Feature learning under anisotropy: Scaling and sample complexity of quadratic networks

Murat A. Erdogdu[1,2], **Xin T. Tong**[3], **Mert Vural**[4,*], **Qingyuan Yu**[5,*]

[1]University of Toronto, [2]Vector Institute, [3]National University of Singapore
[4]UC Berkeley, [5]University of Michigan

erdogdu@cs.toronto.edu, xin.t.tong@nus.edu.sg
vural@berkeley.edu, qyyu@umich.edu

[*]Equal contribution; corresponding authors.

Abstract

Recent theoretical work has established scaling laws for feature learning in tractable nonlinear models, but has largely focused on isotropic inputs. Understanding how these scaling laws extend to anisotropic data remains a central open problem. We address this question by studying two-layer networks with quadratic activations trained on anisotropic Gaussian data. We derive sharp characterizations of the excess risk for both population gradient flow and finite-sample online stochastic gradient descent (SGD), quantifying how the covariance spectrum affects the scaling exponents and sample complexity. We also analyze generalization in the corresponding lazy neural tangent kernel (NTK) regime and establish two distinct advantages of feature learning. With suitable initialization, SGD improves the dependence of sample complexity on the effective dimension when this dimension diverges, and achieves a strictly faster power-law decay of the excess risk when it remains bounded. Finally, experiments with more general activation functions closely match our theoretical predictions, providing evidence that the scaling mechanisms identified by our analysis extend beyond quadratic networks.

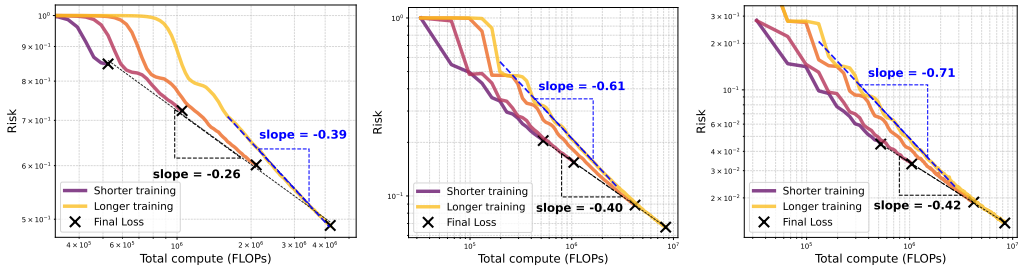
1 Introduction

Scaling laws, which characterize the predictable scaling of generalization error with respect to computational resources, dataset size, and model width, are now foundational to modern machine learning [HNA+17, HBM+22]. Given that empirical loss consistently exhibits robust power-law decay, practitioners can reliably project large-scale performance from a limited set of preliminary measurements [KMH+20, BDK+24]. This phenomenon has sparked significant theoretical interest in deriving scaling laws from first principles [MLGT24, NFLL24]. While much of the foundational work focused on linear models [PPXP24, LWK+24], recent efforts have shifted toward nonlinear solvable models, which are necessary to capture phenomena such as feature learning [NFLL24, DXG+25].

Despite this progress, existing scaling analyses for nonlinear networks largely rely on the restrictive assumption of isotropic inputs [BEVW25, RNWL25]. In this work, we extend this line of research to *anisotropic* settings. Specifically, we study the learning dynamics of a two-layer neural network with quadratic activations under anisotropic Gaussian data. The target function and the input distribution are given by

$$y = \sum_{j=1}^{r} \lambda_j \big(\langle \boldsymbol{\theta}_j, \boldsymbol{x} \rangle^2 - \langle \boldsymbol{\theta}_j, \boldsymbol{\Sigma} \boldsymbol{\theta}_j \rangle \big) + \varepsilon \quad \text{and} \quad \boldsymbol{x} \sim \mathcal{N}(0, \boldsymbol{\Sigma}). \tag{1.1}$$

Here, ε is zero-mean independent sub-Gaussian noise, $\{\boldsymbol{\theta}_j\}_{j=1}^r \subset \mathbb{R}^d$ are unknown signal directions, $\lambda_1 > \lambda_2 > \cdots > \lambda_r > 0$ denote their associated signal strengths, and $\boldsymbol{\Sigma}$ is the data covariance



(a) $(\alpha, \beta) = (0.6, 0.0)$ (b) $(\alpha, \beta) = (0.6, 0.3)$ (c) $(\alpha, \beta) = (0.6, 0.6)$

Figure 1: Population risk vs. compute (width $\times$ steps $\times$ batch size) for two-layer quadratic neural networks for $T \in \{2048, 4096, 8192, 16384\}$ trained with one-pass SGD using a batch size of 512 and squared loss. We tuned the best learning rate for each run and plot the corresponding risk curves. The blue dots correspond to the optimization exponent, which is consistent with our theoretical prediction of $\frac{2\alpha+2\beta-1}{\alpha+2\beta}$. The black lines represent the best risk attained at the end of training, which follows a different slope due to the decay in the fitted learning rates. We set $d = 8192$, $r = r_s = 512$, and $\sigma = 0.5$.

matrix. Our goal is to learn this teacher network using a two-layer student network with width r_s, trained via gradient-based optimization, and to characterize the corresponding scaling behavior.

Several classical high-dimensional inference problems, including phase retrieval ($r = 1$) [Fie82, CC15, TV23] and multi-spike PCA ($r = \Theta(1)$) [OK85, JJK$^+$16, AZL17, BGP25], can be formulated as learning quadratic networks. More broadly, these belong to the family of multi-index models, where the target function depends on a small number of latent directions [BBPV23, DKL$^+$23, CWPPS23, MHWE24]. Such models offer a tractable framework for studying feature learning, specifically how algorithms adapt to hidden low-dimensional structures despite high-dimensional data. While prior work extensively analyzes the isotropic regime ($\boldsymbol{\Sigma} = \boldsymbol{I}_d$)—where risk dynamics admit dimension-independent characterizations [BAGJ21, BAGJ22]—these properties break down under anisotropy. Consequently, the anisotropic regime remains largely underexplored, as it requires solving the dynamics directly in the high-dimensional ambient space [BLMI25, BMI25].

In practice, however, real-world data is inherently anisotropic, and this underlying covariance structure can fundamentally alter the learning dynamics. While recent works have studied these covariance effects in linear regression [PPXP24] and phase retrieval [BLMI25], they focus exclusively on the single-index setting ($r = 1$). This limits the applicability of these works, as scaling laws for feature learning inherently require a large number of latent directions to emerge [BEVW25, RNWL25]. As a result, the impact of data covariance on the scaling behavior of feature learning remains poorly understood. Motivated by this, we ask:

Q1: Can we characterize how data anisotropy affects the scaling behavior of gradient-based training in the feature learning regime?

A natural complementary question is how the scaling behavior of the feature learning regime compares to that of lazy learning. Earlier work demonstrated that in an isotropic single-index setting, SGD in the feature learning regime can significantly improve sample complexity [BAGJ21, LOSW24, DTA$^+$24]. However, these previous results cannot capture the effect of data anisotropy, nor do they explain how scaling exponents change within the feature learning regime itself. Under general anisotropy, existing work establishes that kernel methods require a sample complexity that scales polynomially with the effective dimension—specifically, $\mathcal{O}((r_{\text{eff}}(\boldsymbol{\Sigma}))^k)$ to learn k-degree polynomials, where $r_{\text{eff}}(\boldsymbol{\Sigma}) \coloneqq \operatorname{tr}(\boldsymbol{\Sigma})/\|\boldsymbol{\Sigma}\|_2$ [WL25, GMMM19]. Yet, a direct comparison between these bounds and the feature learning regime under general covariance structures remains elusive. Therefore, we ask:

Q2: How do the sample complexity and scaling behavior of the feature learning regime compare to the lazy regime?

1.1 Our Contributions

To address these questions, we study the learning dynamics of quadratic neural networks, which provide a tractable yet theoretically rich setting for analyzing multi-index learning. Specifically,

Algorithm	Risk scaling	Sample complexity	Result
Online SGD	$(\eta N)^{-\frac{2(\alpha+\beta)-1}{\alpha+2\beta}} + r_s^{-2(\alpha+\beta)+1}$	$\tilde{\mathcal{O}}(d)$	Corollary 1
Online SGD (rescaled initialization)		$\tilde{\mathcal{O}}(r_{\text{eff}}(\mathbf{\Sigma}))$	Corollary 1
NTK	$\begin{cases} (r_{\text{eff}}(\mathbf{\Sigma})^2/N)^{\frac{2\alpha+2\beta-1}{2\beta}} & \beta<1, \alpha<\beta+0.5 \\ (r_{\text{eff}}(\mathbf{\Sigma})^2/N)^2 & \beta<1, \alpha>\beta+0.5 \\ N^{-\frac{2\alpha+2\beta-1}{2\alpha+2\beta+1}} & \beta>1, \alpha<\beta+0.5 \\ N^{-\frac{2\beta}{2\beta+1}} & \beta>1, \alpha>\beta+0.5 \end{cases}$	$\mathcal{O}(r_{\text{eff}}(\mathbf{\Sigma})^2)$	Theorem 2

Table 1: Scaling laws and sample complexities for learning a quadratic network (1.1) with the student model (2.1) using online SGD in the feature learning regime. We consider two initialization schemes for online SGD: the standard $1/\sqrt{d}$ variance scaling, and a covariance-dependent initialization that yields improved sample complexity. For comparison, we provide the generalization bound of the identical model evaluated at the lazy (NTK) global minimizer. Crucially, the SGD bounds track the excess risk along the optimization trajectory at iteration N, whereas the NTK bounds evaluate the empirical risk minimizer fitted on N samples. The parameters η and r_s denote the learning rate and network width, respectively. Logarithmic factors are omitted for clarity.

we investigate the model defined in (1.1) under structured power-law conditions on both the signal strengths and the data covariance spectrum. Letting σ_i denote the eigenvalues of $\mathbf{\Sigma}$, we assume

$$\lambda_i \asymp i^{-\alpha} \qquad \text{and} \qquad \sigma_i \asymp i^{-\beta}, \tag{1.2}$$

for exponents $\alpha, \beta > 0$. In the statistical learning literature, α and β correspond to the *source* and *capacity* conditions, respectively [CdV07a, WL25]. Our main contributions are summarized below (see also Table 1):

1. In Section 3, we analyze both the population gradient flow and online SGD dynamics, deriving a sharp characterization of the excess risk as a function of training time and network width r_s. In particular, we show that the risk exhibits an explicit power-law decay jointly governed by the source and capacity exponents α and β. We validate these predictions empirically and observe strong agreement between our theoretical curves and simulations (Figure 1).
2. We study the sample complexity of learning with SGD under anisotropy. Our analysis reveals that online SGD under standard initialization yields a suboptimal $\tilde{\mathcal{O}}(d)$ sample complexity. To overcome this limitation, we introduce a covariance-dependent rescaled initialization scheme. We demonstrate that this approach achieves an information-theoretically optimal sample complexity of $\tilde{\mathcal{O}}(r_{\text{eff}}(\mathbf{\Sigma}))$, establishing a sharp dimensional dependency for feature learning.
3. Drawing on these results, we compare the scaling exponents and sample complexity of SGD against the lazy (NTK) regime of the identical learning model. As summarized in Table 1, our analysis reveals a strict gap between the feature learning and kernel regimes: SGD strictly improves either the dimension dependency in settings where the effective dimension grows ($\beta < 1$), or the power-law risk decay exponent where the effective dimension remains finite ($\beta > 1$).
4. Finally, we empirically demonstrate that our predicted scaling exponents extend to general 2-homogeneous link functions (Figure 2). This universality arises as a natural consequence of incremental learning, where the target function is learned progressively and the quadratic components drive the initial phase of feature learning [BEVW25]. This observation confirms that our scaling results hold beyond quadratic neural networks.

1.2 Additional related works

Theory of scaling laws. The study of risk scaling under power-law structures is a classical topic in the kernel literature, typically formalized through the framework of source and capacity conditions [CDV07b, CLKZ21]. More recently, attention has shifted toward solvable nonlinear models. This includes establishing fundamental statistical limits [DKLM26b, DKLM26a], deriving exact asymptotic characterizations of empirical risk minimizers [DXG+25], and analyzing optimization dynamics [RNWL25, BEVW25]. Our work advances this line of research by addressing anisotropic setting, which has received comparatively little attention due to technical challenges discussed above.

Feature learning with multi-index teacher models. Learning low-dimensional structure from high-dimensional data is a fundamental problem in statistics [LD89]. The simplest case, where $r = 1$ (the *single-index model*), has been extensively studied in the feature learning literature [BAGJ21, BES+22, MHPG+22, BES+23, MHWSE23, MLHD23, MZD+23, BMZ24, DNGL24, GWB25]. 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 idealized isotropic input assumptions [AAM22, AAM23, BBPV23, DKL+23, CWPPS23, BAGP24, VE24, MHWE24]. In contrast, theoretical understanding under anisotropic data remains scarce. Existing analyses either focus on spiked covariance structures requiring preconditioned optimization methods [BES+22], or impose restrictive constraints, such as analyzing only population-level dynamics [BLMI25] or well-conditioned covariances [BMI25]. In our work, we consider anisotropic power-law spectra consistent with similar settings from earlier literature [PPXP24]. Furthermore, we study the finite-sample setting and provide a rigorous sample complexity analysis.

Comparison with [BLMI25]. Recently, [BLMI25] studied the single-index ($r = 1$) case of our problem—namely, phase retrieval—under power-law data anisotropy at the population level. Although our problem settings are related, our work differs in two essential ways. First, while they analyze population-level dynamics in continuous time, we establish finite-sample complexity guarantees for a discrete-time online algorithm. Second, we consider the multi-index setting with a large number of latent directions ($r \gg 1$), which fundamentally changes how power-law convergence arises. Specifically, their derived exponent characterizes the local convergence rate near the optimum. In contrast, our scaling exponents capture the global learning dynamics across the entire training trajectory, as they arise from a superposition of sequential learning curves that emerge throughout training [RNWL25, BEVW25].

2 Background and 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, respectively, and $\tilde{\mathcal{O}}(\cdot)$ to suppress polylogarithmic factors in d. We write $f \in \Theta(g)$ to denote $\Omega(g) \leq f \leq \mathcal{O}(g)$. We use $o_d(\cdot)$ to denote terms vanishing as $d \to \infty$, and write $f(d) \ll g(d)$ (respectively, $f(d) \gg g(d)$) if $f \in o_d(g)$ (respectively, $g \in o_d(f)$). Finally, all the statements in the main text are formalized in appendix.

2.1 Student-teacher Setting

Teacher Network. We consider learning a two-layer teacher network with a quadratic (second-order Hermite) activation when the input is anisotropic Gaussian. Specifically, we assume the input $\boldsymbol{x} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$ and the target y is generated by:

$$y = \sum_{j=1}^{r} \lambda_j \left(\langle \boldsymbol{\theta}_j, \boldsymbol{x} \rangle^2 - \langle \boldsymbol{\theta}_j, \boldsymbol{\Sigma} \boldsymbol{\theta}_j \rangle \right) + \varepsilon, \tag{2.1}$$

where r is the width of the teacher network, and ε is independent, zero-mean sub-Gaussian noise. The set $\{\boldsymbol{\theta}_j\}_{j=1}^r \subset \mathbb{R}^d$ consists of unknown, orthonormal signal vectors with strictly decreasing coefficients $\lambda_1 > \lambda_2 > \cdots > \lambda_r > 0$. The constant terms $\langle \boldsymbol{\theta}_j, \boldsymbol{\Sigma} \boldsymbol{\theta}_j \rangle$ are included to ensure the labels remain centered, i.e., $\mathbb{E}[y] = 0$.

The interaction between the covariance $\boldsymbol{\Sigma}$ and the signal directions $\{\boldsymbol{\theta}_j\}_{j=1}^r$ fundamentally determines the complexity of the learning problem. To formalize this, we impose a spectral alignment condition:

Assumption 1 (Spectral alignment)**.** *Each $\boldsymbol{\theta}_j$ is the j-th eigenvector of $\boldsymbol{\Sigma}$, such that $\boldsymbol{\Sigma}\boldsymbol{\theta}_j = \sigma_j \boldsymbol{\theta}_j$. For normalization, we assume $\sum_{i=1}^r \sigma_i^2 \lambda_i^2 = 1$ so that the variance of labels, i.e., $\mathbb{E}[y^2]$, is constant.*

This assumption formalizes a classical statistical learning principle: the directions most predictive of the response y coincide with the principal components of the input covariates $\boldsymbol{x}$ [HTF09]. Combined with the power-law assumptions in (1.2), our setting extends classical three-parameter models to the multi-index setting through a shared spectral decay of the covariance and signal components [MRS22, PPXP24].

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

$$\hat{y}(\boldsymbol{x}; \boldsymbol{W}, b) = \frac{1}{r_s} \sum_{j=1}^{r_s} (\langle \boldsymbol{w}_j, \boldsymbol{x} \rangle^2 - b),$$

where r_s is the width of the student network, b is the bias term, and $\{\boldsymbol{w}_j\}_{j=1}^{r_s} \subset \mathbb{R}^d$ denotes the set of trainable weights. We collect these weights as the columns of the matrix $\boldsymbol{W} \in \mathbb{R}^{d \times r_s}$.

We consider minimizing the squared loss, for which the population risk is defined by

$$R(\boldsymbol{W}, b) := \tfrac{1}{2} \mathbb{E}_{(\boldsymbol{x}, y)} \left[(y - \hat{y}(\boldsymbol{x}, \boldsymbol{W}, b))^2 \right].$$

By substituting the teacher and student models in (2.1) and , the population risk admits the following closed-form expression:

$$R(\boldsymbol{W}, b) = \left\| \frac{1}{r_s} \boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{W} \boldsymbol{W}^\top \boldsymbol{\Sigma}^{\frac{1}{2}} - \boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{\Theta} \boldsymbol{\Lambda} \boldsymbol{\Theta}^\top \boldsymbol{\Sigma}^{\frac{1}{2}} \right\|_{\mathrm{F}}^2 + \tfrac{1}{2} \left(b - \tfrac{1}{r_s} \operatorname{tr}(\boldsymbol{\Sigma} \boldsymbol{W} \boldsymbol{W}^\top) \right)^2 + \tfrac{1}{2} \mathbb{E}[\varepsilon^2]. \quad (2.2)$$

Here, we collect the teacher directions $\{\boldsymbol{\theta}_j\}_{j=1}^{r}$ as the columns of the matrix $\boldsymbol{\Theta} \in \mathbb{R}^{d \times r}$, and $\boldsymbol{\Lambda}$ is a diagonal matrix where the j-th diagonal entry is λ_j.

3 Main Result: Risk characterization under anisotropy

3.1 Warm-up: Two-time-scale population dynamics

We begin by analyzing the continuous-time population dynamics to build intuition for our subsequent finite-sample SGD results. To learn the weights and bias simultaneously, we employ a two-time-scale gradient flow where the bias updates faster than the weights. This separation of timescales yields a mathematically tractable learning trajectory and allows the algorithm to learn the covariance-dependent bias term in (2.2) that keeps the predictions centered. Specifically, we consider:

$$\partial_t \boldsymbol{W}_t = -r_s \nabla_{\boldsymbol{W}} R(\boldsymbol{W}_t, b_t), \qquad \partial_t b_t = -\frac{1}{\delta} \nabla_b R(\boldsymbol{W}_t, b_t) \quad \text{(GF)}$$

where the parameters are initialized with $b_0 \sim \mathcal{N}(0, 1)$ and $\boldsymbol{W}_0 \in \mathbb{R}^{d \times r_s}$ where $\boldsymbol{W}_{0,ij} \sim_{\text{i.i.d.}} \mathcal{N}(0, 1/d)$. Here $R(\cdot, \cdot)$ is the population objective defined in (2.2), and $\delta > 0$ controls the relative learning speed of the bias. For notational convenience, we let $\mathcal{R}(t) := R(\boldsymbol{W}_t, b_t)$.

The following proposition characterizes the asymptotic behavior of the population risk under (GF).

Proposition 1. *Let $\alpha, \beta > 0$ denote the source and capacity exponents defined in (1.2), and assume the teacher width $r \gg r_s$. We consider the regime $2\alpha + 2\beta > 1$ and a mildly overparameterized student network with $r_s \gg \operatorname{polylog} d$. For $\delta = \frac{1}{\log^6 d}$, the dynamics in (GF) satisfy*

$$\mathcal{R}(t \log d) - \tfrac{1}{2} \mathbb{E}[\varepsilon^2] \asymp \Theta \left(t^{-\frac{2(\alpha+\beta)-1}{\alpha+2\beta}} + r_s^{-2(\alpha+\beta)+1} \right).$$

This asymptotic risk exhibits a decomposition similar to the neural scaling laws in [KMH+20, HBM+22], where the error cleanly separates into optimization and approximation terms. Here, the capacity exponent β enters both contributions additively due to the alignment structure. In particular, the limit $\beta \to 0$ recovers the isotropic scaling behavior [BEVW25]; however, the above result achieves better scaling exponents for any level of anisotropy, $\beta > 0$. In the next section, we connect this continuous-time analysis to the sample complexity of online SGD.

3.2 Discrete dynamics and sample complexity

We now study the finite-sample, discrete-time counterpart of (GF) and establish both computational and statistical guarantees. In particular, we consider an online (one-pass) discretization of the two-time-scale dynamics. Given a fresh sample $(\boldsymbol{x}_{t+1}, y_{t+1})$, we define the instantaneous loss:

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

where $\hat{y}(\boldsymbol{x}_{t+1}; \boldsymbol{W}, b)$ denotes the network output. The updates are given by:

$$\boldsymbol{W}_{t+1} = \boldsymbol{W}_t - \eta \nabla_{\boldsymbol{W}} \mathcal{L}_{t+1}(\boldsymbol{W}_t, b_t), \qquad b_{t+1} = b_t - \frac{\eta}{\delta} \nabla_b \mathcal{L}_{t+1}(\boldsymbol{W}_t, b_t) \tag{SGD}$$

where the parameters are initialized as in (GF), and $\delta > 0$ controls the timescale of the bias update.

In the following, we provide a master theorem that establishes the sample complexity of online SGD for different choices of learning rate and initialization schemes.

Theorem 1. *Let the parameters α, β, r, r_s, δ be defined as in Proposition 1. We consider the initialization*

$$W_{0,ij} \sim \mathcal{N}\Big(0, \frac{1}{\rho_d}\Big) \quad \textit{and} \quad \eta = \frac{1}{\tilde{\mathcal{O}}(\rho_d)},$$

for a dimension-dependent scaling parameter ρ_d in the range

$$\Theta(d) \geq \rho_d \geq \tilde{\mathcal{O}}(r_{\mathit{eff}}(\boldsymbol{\Sigma})).$$

Almost surely, for any iteration $t \leq \frac{\log^2 d}{\eta}$ with a well defined continuous-time limit ηt, the excess risk satisfies:

$$\mathcal{R}(t) - \tfrac{1}{2}\mathbb{E}[\varepsilon^2] \asymp \Theta\Big((\eta t)^{-\frac{2(\alpha+\beta)-1}{\alpha+2\beta}} + r_s^{-2(\alpha+\beta)+1} \Big).$$

Remark 1. *We make the following remarks regarding Theorem 1:*

- *Theorem 1 evaluates a scaled initialization scheme where the scaling parameter ρ_d interpolates between the standard initialization regime of $\Theta(d)$ and a covariance-aware regime of $\tilde{\mathcal{O}}(r_{\mathit{eff}}(\boldsymbol{\Sigma}))$. Our current analysis requires choosing the initialization variance proportional to the learning rate. We explain the reason for this in the proof overview and conjecture that this coupling between the learning rate and initialization variance is merely a proof artifact.*
- *Additionally, we incorporate an additional logarithmic factor in the learning rate to ensure that the discrete SGD iterates in (SGD) closely track the continuous population flow trajectory (GF) at the effective time ηt, explaining why the scaling exponents coincide.*

The following corollary establishes the resulting sample complexities at both ends of this spectrum.

Corollary 1. *Theorem 1 implies the following convergence rates parameterized by a normalized time scale t:*

$$\left.\begin{array}{ll} \textit{(Standard init)} & \mathcal{R}(td\,\mathrm{polylog}\, d) \\ \textit{(Rescaled init)} & \mathcal{R}(tr_{\mathit{eff}}(\boldsymbol{\Sigma})\,\mathrm{polylog}\, d) \end{array}\right\} - \tfrac{1}{2}\mathbb{E}[\epsilon^2] \asymp \Theta\Big(t^{-\frac{2(\alpha+\beta)-1}{\alpha+2\beta}} + r_s^{-2(\alpha+\beta)+1} \Big),$$

where the standard initialization uses $\rho_d = \Theta(d)$ and the rescaled initialization uses $\rho_d = r_{\mathit{eff}}(\boldsymbol{\Sigma})\,\mathrm{polylog}\, d$.

Consequently, under rescaled initialization, the sample complexity of SGD is $\tilde{\mathcal{O}}(r_{\mathrm{eff}}(\boldsymbol{\Sigma}))$, which is information-theoretically optimal. This strictly improves the $\tilde{\mathcal{O}}(r_{\mathrm{eff}}(\boldsymbol{\Sigma})^2)$ dependency of kernel methods to $\tilde{\mathcal{O}}(r_{\mathrm{eff}}(\boldsymbol{\Sigma}))$, extending the gap established in the isotropic setting to anisotropy [GMMM19].

4 Lazy Regime

In this section, we analyze the training of a two-layer neural network in the lazy regime. Under the standard Neural Tangent Kernel (NTK) parameterization [JGH18], the network output is given by

$$\hat{y}(\boldsymbol{x}; \boldsymbol{W}, b) = \frac{1}{\sqrt{r_s}} \sum_{i=1}^{r_s} \langle \boldsymbol{w}_i, \boldsymbol{x} \rangle^2 - b,$$

where the weights are initialized as $\boldsymbol{w}_i \sim \mathcal{N}(0, \boldsymbol{I}_d)$ and the bias is initialized to zero.

In this lazy regime, the weights remain trapped in a local neighborhood of their initialization throughout training. Consequently, the evolution of the network predictions under gradient flow is

deterministically governed by the empirical tangent kernel, defined over the concatenated parameters $\boldsymbol{\theta} = (\boldsymbol{W}, b)$ as:

$$\hat{\Theta}_{r_s}(\boldsymbol{x}, \boldsymbol{x}') \coloneqq \langle \nabla_{\boldsymbol{\theta}} \hat{y}(\boldsymbol{x}), \nabla_{\boldsymbol{\theta}} \hat{y}(\boldsymbol{x}') \rangle.$$

Standard NTK theory establishes that as $r_s \to \infty$, this empirical kernel concentrates around its expectation over the initialization. The following proposition characterizes this limiting kernel.

Proposition 2. *As $r_s \to \infty$, the induced Neural Tangent Kernel converges almost surely to the deterministic kernel:*

$$\Theta(\boldsymbol{x}, \boldsymbol{x}') = 4\langle \boldsymbol{x}, \boldsymbol{x}' \rangle^2 + 1,$$

which induces the explicit feature map mapping $\boldsymbol{x}$ to $\mathbb{R}^{d^2+1}$ given by:

$$\boldsymbol{x} \mapsto \big[\underbrace{\mathrm{vec}(2\boldsymbol{x}\boldsymbol{x}^\top)}_{\phi(\boldsymbol{x})}, 1\big]^\top,$$

where $\mathrm{vec}(\cdot)$ denotes the matrix vectorization operator.

Because the NTK remains constant throughout training in the infinite-width limit, optimizing the neural network is mathematically equivalent to kernel ridge regression using the fixed kernel Θ. To study the generalization error of the lazy network, we evaluate the excess risk of this ridge estimator over a fresh test sample $(\boldsymbol{x}, y)$:

$$\mathcal{R}_\phi(\lambda) \coloneqq \mathbb{E}_{\boldsymbol{x},y}\left[\frac{1}{2}\big(y - \hat{\boldsymbol{\beta}}(\lambda)^\top \phi(\boldsymbol{x}) - \hat{b}\big)^2\right] - \frac{1}{2}\mathbb{E}[\epsilon^2],$$

where the learned weights and bias are given by the regularized empirical risk minimizer:

$$\big(\hat{\boldsymbol{\beta}}(\lambda), \hat{b}\big) \coloneqq \arg\min_{\boldsymbol{\beta}, b} \frac{1}{N}\sum_{i=1}^{N}\big(y_i - \boldsymbol{\beta}^\top \phi(\boldsymbol{x}_i) - b\big)^2 + \lambda\|\boldsymbol{\beta}\|^2. \tag{4.1}$$

The following theorem characterizes the asymptotic limit of $\mathcal{R}_\phi(\lambda)$.

Theorem 2. *Assume $\alpha + \beta > 0.5$, $\lambda \geq 0$ and that the sample size scales as:*

- *For $\beta < 1$, we take $d \to \infty$ while keeping $r_{\mathit{eff}}(\boldsymbol{\Sigma})^2/N$ fixed.*
- *For $\beta > 1$, we first take $d \to \infty$ and then $N \to \infty$.*

Under the optimal choice of regularization parameter $\lambda^* \geq 0$, the excess risk scales as $d \to \infty$:

$$\mathcal{R}_\phi(\lambda^*) \asymp \begin{cases} (r_{\mathit{eff}}(\boldsymbol{\Sigma})^2/N)^{\frac{2\alpha+2\beta-1}{2\beta}} & \beta < 1, \alpha < \beta + 0.5 \\ (r_{\mathit{eff}}(\boldsymbol{\Sigma})^2/N)^2 & \beta < 1, \alpha > \beta + 0.5 \\ (\log N/N)^{\frac{2\alpha+2\beta-1}{2\alpha+2\beta+1}} & \beta > 1, \alpha < \beta + 0.5 \\ (\log N/N)^{\frac{2\beta}{2\beta+1}} & \beta > 1, \alpha > \beta + 0.5. \end{cases}$$

Remark 2. *We make the following remarks regarding Theorem 2:*

- *Theorem 2 provides an exact characterization of the asymptotic error under the optimally tuned ridge parameter λ^*. We note that although the NTK is often studied in the context of ridgeless kernel regression [HMRT22], evaluating the optimally tuned ridge estimator provides a rigorous lower bound for the generalization error of the quadratic model trained in the lazy regime.*
- *In the $\beta < 1$ regime, the sample complexity of the kernel risk has a worse dependence on the effective dimension compared to SGD, which fundamentally drives the generalization gap between the two methods. In the $\beta > 1$ regime, the effective dimension is finite. In this case, our exponent is strictly larger than that of the kernel method.*
- *Finally, a more direct comparison between SGD and kernel methods would be to compare the tuned SGD's risk, which has a different scaling than the optimization scaling, as illustrated in Figure 1. This direction is out of the scope of this paper, but we leave it as interesting future work.*

5 Overview of Proof Techniques

To provide intuition for the proof, we focus on the population gradient flow dynamics in (GF). Without loss of generality, we assume the teacher directions coincide with the standard basis vectors, so that the teacher matrix $\boldsymbol{\Theta}\boldsymbol{\Lambda}\boldsymbol{\Theta}^\top$ and the covariance matrix $\boldsymbol{\Sigma}$ are diagonal. For notational convenience, we define:

$$\boldsymbol{T}^\star := \boldsymbol{\Sigma}^{1/2}\boldsymbol{\Theta}\boldsymbol{\Lambda}\boldsymbol{\Theta}^\top\boldsymbol{\Sigma}^{1/2}, \quad \boldsymbol{M}(\boldsymbol{W}) := \tfrac{1}{r_s}\boldsymbol{\Sigma}^{1/2}\boldsymbol{W}\boldsymbol{W}^\top\boldsymbol{\Sigma}^{1/2}, \quad q(\boldsymbol{W}, b) := b - \operatorname{tr}(\boldsymbol{M}(\boldsymbol{W})),$$

where $q(\boldsymbol{W}, b)$ denotes the bias mismatch. With this notation, the population risk in (2.2) can be compactly written as:

$$\mathcal{R}(\boldsymbol{W}_t, b_t) = \|\boldsymbol{T}^\star - \boldsymbol{M}(\boldsymbol{W}_t)\|_{\mathrm{F}}^2 + \tfrac{1}{2}q(\boldsymbol{W}, b)^2 + \tfrac{1}{2}\mathbb{E}[\epsilon^2].$$

Because $\boldsymbol{M}(\boldsymbol{W}_t)$ fully characterizes the risk trajectory, our analysis centers on understanding its evolution. *Our primary proof device is to compare the true dynamics $\boldsymbol{M}(\boldsymbol{W}_t)$ with an idealized aligned flow, where the initial weights $\boldsymbol{W}_0$ are perfectly aligned with the teacher directions $\{\boldsymbol{\theta}_j\}_{j=1}^r$.* We execute this comparison in two distinct steps: first, utilizing the fast time-scale parameter $\delta \asymp o_d(1/\log d)$ to enforce continuous centering, and second, leveraging mild overparameterization ($r_s \gg \operatorname{polylog} d$) to tightly couple the true and idealized trajectories.

Two-time-scale argument. The fast bias time scale ($\delta \ll 1$) ensures that the network predictions are rapidly centered; that is, the bias mismatch $q(\boldsymbol{W}_t, b_t)$ becomes negligible early in the effective training horizon. Because the objective is strongly convex with respect to b and satisfies $\nabla_b\mathcal{R}(\boldsymbol{W}_t, b_t) = q(\boldsymbol{W}_t, b_t)$, the accelerated optimization induced by the large $1/\delta$ factor guarantees that $e(\boldsymbol{W}_t, b_t)$ is driven to zero and remains uniformly small throughout training (see Remarks 3 and 4).

Comparison with the idealized dynamics. Assuming the bias mismatch satisfies $q(\boldsymbol{W}, b) = 0$, the trajectory $\boldsymbol{M}_t := \boldsymbol{M}(\boldsymbol{W}_t)$ induced by the gradient flow (GF) satisfies:

$$\partial_t\boldsymbol{M}_t = \boldsymbol{\Sigma}\boldsymbol{T}^\star\boldsymbol{M}_t + \boldsymbol{M}_t\boldsymbol{T}^\star\boldsymbol{\Sigma} - \boldsymbol{\Sigma}\boldsymbol{M}_t^2 - \boldsymbol{M}_t^2\boldsymbol{\Sigma}, \quad \boldsymbol{M}_0 = \frac{1}{\rho_d}\boldsymbol{\Sigma}^{1/2}\Big(\frac{1}{r_s}\boldsymbol{Z}\boldsymbol{Z}^\top\Big)\boldsymbol{\Sigma}^{1/2}, \quad (5.1)$$

where $\boldsymbol{Z} \in \mathbb{R}^{d\times r_s}$ is a random matrix with i.i.d. $\mathcal{N}(0,1)$ entries, and we explicitly isolate the initialization scale ρ_d.

Crucially, by Assumption 1, we can assume that $\boldsymbol{\Sigma}$ and $\boldsymbol{T}^\star$ are diagonal matrices. From this, we make two key observations:

- First, although the evolution in (5.1) is governed by a second-order matrix polynomial, it does not reduce to a classical Riccati matrix ODE of the type studied in prior work [BEVW25]. Consequently, proof techniques developed for isotropic settings cannot be applied directly.
- Second, the dynamics would remain strictly diagonal—and reduce exactly to a Riccati ODE—if the initial state $\boldsymbol{M}_0$ were perfectly diagonal. Under anisotropic covariance, concentration at initialization implies that $\boldsymbol{M}_0$ is indeed asymptotically close to a diagonal matrix.

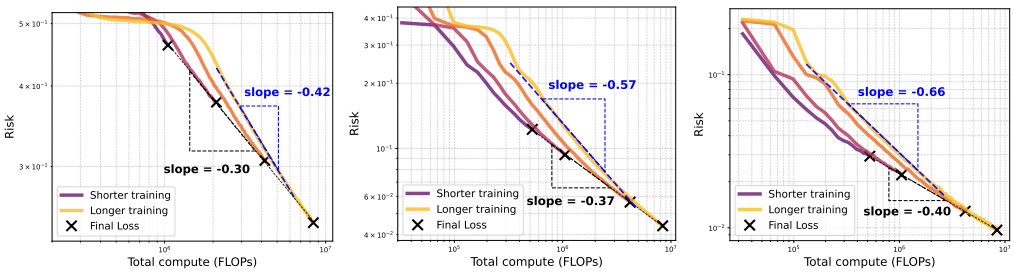
Motivated by these observations, we introduce the idealized diagonal dynamics:

$$\partial_t\boldsymbol{L}_t = \boldsymbol{\Sigma}\boldsymbol{T}^\star\boldsymbol{L}_t + \boldsymbol{L}_t\boldsymbol{T}^\star\boldsymbol{\Sigma} - 2\boldsymbol{L}_t\boldsymbol{\Sigma}\boldsymbol{L}_t, \qquad \boldsymbol{L}_0 = \mathbb{E}[\boldsymbol{M}_0] = \tfrac{1}{\rho_d}\boldsymbol{\Sigma}.$$

We show that for sufficiently large width r_s, overparameterization forces the empirical initialization $\boldsymbol{M}_0$ to concentrate sharply around the anisotropic limit $\boldsymbol{L}_0$. Using this concentration, we prove that the relative discrepancy between the true and idealized trajectories (i.e., $\boldsymbol{R}_t := \boldsymbol{L}_t^{-1/2}\boldsymbol{M}_t\boldsymbol{L}_t^{-1/2}$) remains tightly controlled throughout training (see Theorem 5 and Corollary 3). As a result, the full dynamics of $\boldsymbol{M}_t$ can be accurately approximated by the diagonal flow $\boldsymbol{L}_t$, which admits an explicit analytic characterization. This decoupling ultimately yields the scaling laws stated in Proposition 1.

Extension to online SGD. The analysis for the discrete online SGD dynamics (SGD) follows a similar trajectory. The discrete iteration can be decomposed as:

$$\boldsymbol{M}_{t+1} = \boldsymbol{M}_t + \eta(\boldsymbol{\Sigma}\boldsymbol{T}^\star\boldsymbol{M}_t + \boldsymbol{M}_t\boldsymbol{T}^\star\boldsymbol{\Sigma} - \boldsymbol{\Sigma}\boldsymbol{M}_t^2 - \boldsymbol{M}_t^2\boldsymbol{\Sigma}) + \eta\,\mathrm{Noise}_t + \text{Second-order terms}.$$



(a) $(\alpha, \beta) = (0.6, 0.0)$ (b) $(\alpha, \beta) = (0.6, 0.3)$ (c) $(\alpha, \beta) = (0.6, 0.6)$

Figure 2: Population loss versus compute for a two-layer squared ReLU network. We used the same experimental setting as in Figure 1. The resulting risk trajectories follow a power-law decay, where the empirical exponents closely align with those in Figure 1, as well as with our theoretical result of $\frac{1-2(\alpha+\beta)}{\alpha+2\beta}$ for the quadratic setting.

Similar to the continuous-time case, we study the relative discrepancy $\boldsymbol{R}_t := \boldsymbol{L}_t^{-1/2} \boldsymbol{M}_t \boldsymbol{L}_t^{-1/2}$ against a discrete diagonal sequence $\boldsymbol{L}_t$. We show that with a sufficiently small learning rate, as specified in Theorem 1, the discrete risk closely tracks its continuous-time counterpart.

Under standard initialization with variance $\Theta(1/d)$, the primary technical bottleneck lies in controlling the noise term at initialization. Because the idealized initial state $\boldsymbol{L}_0$ has an operator norm of $\Theta(1/d)$, the learning rate must be severely restricted to an $\tilde{\mathcal{O}}(1/d)$ scale to prevent the stochastic noise terms from overwhelming the initial signal. To overcome this limitation, we consider a rescaled initialization scheme with $\rho_d = \tilde{\mathcal{O}}(r_{\text{eff}}(\boldsymbol{\Sigma}))$. We demonstrate that this calibrated initialization permits a much larger learning rate of $\eta = \tilde{\mathcal{O}}(1/r_{\text{eff}}(\boldsymbol{\Sigma}))$ while preserving theoretical stability, which ultimately yields the near-optimal sample complexity bounds.

6 Conclusion

Existing scaling law analyses for nonlinear networks predominantly rely on the isotropic input assumption, which is overly restrictive in practice. In this work, we bridge this gap by analyzing the anisotropic setting under a specific spectral alignment condition. We derive scaling laws along the dynamics of online SGD and characterize its sample complexity. We show that SGD achieves the near-optimal bound $N \simeq r_{\text{eff}}(\boldsymbol{\Sigma}) \operatorname{polylog} d$, where $r_{\text{eff}}(\boldsymbol{\Sigma}) = \operatorname{tr}(\boldsymbol{\Sigma})/\|\boldsymbol{\Sigma}\|_2$ is the effective ambient dimension. We also derive the asymptotic risk scaling of the NTK for the quadratic network and show how SGD improves the sample complexity.

Universality of exponents: Extending our analysis to general activation functions remains a natural direction for future work. It is conjectured that SGD dynamics exhibit multi-phase risk curves (analogous to the incremental learning phenomenon in [AAM23, BBPV23]), where higher-order Hermite modes affect the learning dynamics only after the low-order terms are learned. Because standard nonlinearities (e.g., GeLU, tanh) contain a nonzero He_2 component, their SGD optimization is expected to exhibit an intermediate phase governed by the same scaling exponents studied in this work. We provide preliminary empirical support for this hypothesis under anisotropic data in Figure 2. By tracking the risk trajectories for networks with squared ReLU[1] activations, we observe a distinct intermediate phase dominated by the quadratic He_2 component. Crucially, the empirical scaling exponents in this regime closely match our theoretical predictions, suggesting a fundamental degree of universality to the scaling laws derived in this work.

Limitations and future work. Here, we list a few limitations of our work and discuss future directions. (i) Spectral alignment condition assumes that teacher weights and the input covariance commute. Despite being in line with traditional methodology, this condition may still be violated in practice; thus, its relaxation is interesting future work. (ii) In this paper, our focus was on two-layer networks with quadratic activations for mathematical convenience. Studying scaling laws under different architectures and more practical activations is left for future studies.

[1] Following [RNWL25, BEVW25], we choose a 2-homogeneous activation function.

References

[AAM22] 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*, pages 4782–4887. PMLR, 2022.

[AAM23] 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*, pages 2552–2623. PMLR, 2023.

[AZL17] 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)*, pages 487–492. IEEE, 2017.

[BAGJ21] 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.

[BAGJ22] 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.

[BAGP24] Gérard Ben Arous, Cédric Gerbelot, and Vanessa Piccolo. High-dimensional optimization for multi-spiked tensor pca. *arXiv preprint arXiv:2408.06401*, 2024.

[BBPV23] Alberto Bietti, Joan Bruna, and Loucas Pillaud-Vivien. On learning Gaussian multi-index models with gradient flow. *arXiv preprint arXiv:2310.19793*, 2023.

[BDK+24] 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.

[BES+22] 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.

[BES+23] 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.

[BEVW25] 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.

[BGP25] Gérard Ben Arous, Cédric Gerbelot, and Vanessa Piccolo. Stochastic gradient descent in high dimensions for multi-spiked tensor pca, 2025.

[BLMI25] Guillaume Braun, Bruno Loureiro, Ha Quang Minh, and Masaaki Imaizumi. Fast escape, slow convergence: Learning dynamics of phase retrieval under power-law data, 2025.

[BMI25] Guillaume Braun, Ha Quang Minh, and Masaaki Imaizumi. Learning a single index model from anisotropic data with vanilla stochastic gradient descent. In *International Conference on Artificial Intelligence and Statistics*, 2025.

[BMZ24] Raphaël Berthier, Andrea Montanari, and Kangjie Zhou. Learning time-scales in two-layers neural networks. *Foundations of Computational Mathematics*, pages 1–84, 2024.

[CC15] 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.

[CdV07a] Andrea Caponnetto and Ernesto de Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7:331–368, 2007.

[CDV07b] Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

[CLKZ21] Hugo Cui, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborov'a. Generalization error rates in kernel regression: the crossover from the noiseless to noisy regime. *Journal of Statistical Mechanics: Theory and Experiment*, 2022, 2021.

[CM22] Chen Cheng and Andrea Montanari. Dimension free ridge regression. *The Annals of Statistics*, 2022.

[CWPPS23] 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.

[DKL+23] Yatin Dandi, Florent Krzakala, Bruno Loureiro, Luca Pesce, and Ludovic Stephan. How two-layer neural networks learn, one (giant) step at a time. *arXiv preprint arXiv:2305.18270*, 2023.

[DKLM26a] 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.

[DKLM26b] Leonardo Defilippis, Florent Krzakala, Bruno Loureiro, and Antoine Maillard. Optimal scaling laws in learning hierarchical multi-index models, 2026.

[DNGL24] 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.

[DTA+24] Yatin Dandi, Emanuele Troiani, Luca Arnaboldi, Luca Pesce, Lenka Zdeborová, and Florent Krzakala. The benefits of reusing batches for gradient descent in two-layer networks: Breaking the curse of information and leap exponents. *arXiv preprint arXiv:2402.03220*, 2024.

[DXG+25] 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.

[Fie82] James R Fienup. Phase retrieval algorithms: a comparison. *Applied optics*, 21(15):2758–2769, 1982.

[GMMM19] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Limitations of lazy training of two-layers neural network. *Advances in Neural Information Processing Systems*, 32, 2019.

[GWB25] 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.

[HBM+22] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

[HKZ11] Daniel Hsu, Sham M. Kakade, and Tong Zhang. A tail inequality for quadratic forms of subgaussian random vectors, 2011.

[HMRT22] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986, 2022.

[HNA+17] Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Patwary, Mostofa Ali, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

[HTF09] Trevor Hastie, Robert Tibshirani, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[JGH18] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580, 2018.

[JJK+16] Prateek Jain, Chi Jin, Sham M Kakade, Praneeth Netrapalli, and Aaron Sidford. Streaming pca: Matching matrix bernstein and near-optimal finite sample guarantees for oja's algorithm. In *Conference on learning theory*, pages 1147–1164. PMLR, 2016.

[KMH+20] 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.

[LD89] Ker-Chau Li and Naihua Duan. Regression analysis under link violation. *Annals of Statistics*, 17:1009–1052, 1989.

[LOSW24] Jason D. Lee, Kazusato Oko, Taiji Suzuki, and Denny Wu. Neural network learns low-dimensional polynomials with sgd near the information-theoretic limit. *ArXiv*, abs/2406.01581, 2024.

[LWK+24] 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.

[LZL23] Yi-Cheng Li, Hao Zhang, and Qian Lin. On the asymptotic learning curves of kernel ridge regression under power-law decay. *ArXiv*, abs/2309.13337, 2023.

[MHPG+22] 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.

[MHWE24] Alireza Mousavi-Hosseini, Denny Wu, and Murat A Erdogdu. Learning multi-index models with neural networks via mean-field langevin dynamics. *arXiv preprint arXiv:2408.07254*, 2024.

[MHWSE23] 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.

[MLGT24] Eric Michaud, Ziming Liu, Uzay Girit, and Max Tegmark. The quantization model of neural scaling. *Advances in Neural Information Processing Systems*, 36, 2024.

[MLHD23] 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.

[MRS22] Alexander Maloney, Daniel A Roberts, and James Sully. A solvable model of neural scaling laws. *arXiv preprint arXiv:2210.16859*, 2022.

[MZD+23] 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.

[NFLL24] 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.

[OK85] 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.

[PPXP24] Elliot Paquette, Courtney Paquette, Lechao Xiao, and Jeffrey Pennington. 4+ 3 phases of compute-optimal neural scaling laws. *arXiv preprint arXiv:2405.15074*, 2024.

[RNWL25] 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.

[Tro15] Joel A. Tropp. An introduction to matrix concentration inequalities, 2015.

[TV23] 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.

[VE24] Nuri Mert Vural and Murat A Erdogdu. Pruning is optimal for learning sparse features in high-dimensions. *arXiv preprint arXiv:2406.08658*, 2024.

[WL25] Arie Wortsman and Bruno Loureiro. Kernel ridge regression under power-law data: spectrum and generalization. *ArXiv*, abs/2510.04780, 2025.

Contents

1 Introduction **1**

1.1 Our Contributions 2

1.2 Additional related works 3

2 Background and Problem Setting **4**

2.1 Student-teacher Setting 4

3 Main Result: Risk characterization under anisotropy **5**

3.1 Warm-up: Two-time-scale population dynamics 5

3.2 Discrete dynamics and sample complexity 5

4 Lazy Regime **6**

5 Overview of Proof Techniques **8**

6 Conclusion **9**

A Preliminaries for Proofs **15**

B Proofs for Main Results **15**

B.1 Proof of statements in Section 3.1 15

B.2 Proof of Theorem 2 16

C Continous-time analysis **16**

C.1 Block matrix analysis 17

D SGD Analysis **21**

D.1 Bounding lower block. 22

D.2 Bounding upper block 23

E Lazy Regime Analysis **27**

E.1 Deterministic equivalent 27

E.1.1 The regime $\beta < 1$ 28

F Auxiliary Results **29**

F.1 Concentration and related results 32

F.2 Noise bounds 35

F.2.1 Upper block 35

F.2.2 Lower block 38

G Miscellaneous **39**

G.1 Proof of Proposition 7 39

A Preliminaries for Proofs

B Proofs for Main Results

B.1 Proof of statements in Section 3.1

We consider the following setting:

- We have orthonormal teacher directions $\{\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_r\} \subseteq \mathbb{R}^d$ and their eigenvalues $\lambda_i = i^{-\alpha}$. Let
$$\boldsymbol{\Theta} \coloneqq [\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_r] \in \mathbb{R}^{d\times r}, \quad \boldsymbol{\Lambda} \coloneqq \operatorname{diag}(\lambda_i) \in \mathbb{R}^{r\times r}, \text{ and } \boldsymbol{T} \coloneqq \boldsymbol{\Theta}\boldsymbol{\Lambda}\boldsymbol{\Theta}^\top \in \mathbb{R}^{d\times d}.$$
We consider a positive definite covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$. We will assume $\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{T}\boldsymbol{\Sigma}^{1/2}\|_F = 1$.
- For $\boldsymbol{x} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$ and $\varepsilon \perp\!\!\!\perp \boldsymbol{x}$ with $\mathbb{E}[\varepsilon] = 0$, the teacher network is given as:
$$y = \tfrac{1}{\sqrt{2}}(\boldsymbol{x}^\top \boldsymbol{T}\boldsymbol{x} - \operatorname{tr}(\boldsymbol{T}\boldsymbol{\Sigma})) + \varepsilon = \left\langle \boldsymbol{T}, \tfrac{1}{\sqrt{2}}(\boldsymbol{x}\boldsymbol{x}^\top - \boldsymbol{\Sigma})\right\rangle + \varepsilon.$$
The student network is given as:
$$\begin{aligned}\hat{y}(\boldsymbol{W}, b) &= \frac{1/\sqrt{2}}{r_s}\sum_{j=1}^{r_s} \langle \boldsymbol{w}_j, \boldsymbol{x}\rangle^2 - b \\ &= \left\langle \frac{1}{r_s}\boldsymbol{W}\boldsymbol{W}^\top, \tfrac{1}{\sqrt{2}}(\boldsymbol{x}\boldsymbol{x}^\top - \boldsymbol{\Sigma})\right\rangle - \tfrac{1}{\sqrt{2}}\Big(b - \operatorname{tr}(\tfrac{1}{r_s}\boldsymbol{W}\boldsymbol{W}^\top\boldsymbol{\Sigma})\Big).\end{aligned}$$
We consider the risk defined by:
$$\begin{aligned}R(\boldsymbol{W}, b) &\coloneqq \mathbb{E}[(y - \hat{y}(\boldsymbol{W}, b))^2] \\ &= \left\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{T}\boldsymbol{\Sigma}^{1/2} - \frac{1}{r_s}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{W}\boldsymbol{W}^\top\boldsymbol{\Sigma}^{\frac{1}{2}}\right\|_F^2 + \frac{1}{2}\Big(b - \operatorname{tr}(\tfrac{1}{r_s}\boldsymbol{W}\boldsymbol{W}^\top\boldsymbol{\Sigma})\Big)^2 + \frac{1}{2}\mathbb{E}[\varepsilon^2].\end{aligned}$$
- We define
$$\begin{aligned}&\boldsymbol{T}^* \coloneqq \boldsymbol{\Sigma}^{1/2}\boldsymbol{T}\boldsymbol{\Sigma}^{1/2}, \quad \boldsymbol{A} \coloneqq \boldsymbol{\Sigma}\boldsymbol{T}^*, \quad \boldsymbol{G}_t \coloneqq \tfrac{1}{r_s}\boldsymbol{W}_t\boldsymbol{W}_t^\top, \\ &\boldsymbol{M}_t \coloneqq \boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{G}_t\boldsymbol{\Sigma}^{\frac{1}{2}}, \quad q_t \coloneqq \operatorname{tr}(\boldsymbol{M}_t) - b_t.\end{aligned}$$

We will make the following assumptions:

Assumption 2. *We assume the following:*

1. *We assume that* $[\boldsymbol{T}, \boldsymbol{\Sigma}] = 0$.
2. *We assume that* $\lambda_i \propto i^{-\alpha}$ *and* $\sigma_i \coloneqq \boldsymbol{\Sigma}_{ii} = i^{-\beta}$ *for* $\alpha, \beta \geq 0$ *and* $\alpha + \beta > 0.5$.

Under these assumptions, we have that $\boldsymbol{T}$, $\boldsymbol{T}^*$ and $\boldsymbol{A}$ are diagonal matrices.

Theorem 3. *We consider* (GF) *with*
$$r \gg r_s \gg \log^{18} d, \qquad \epsilon \ll \frac{1}{\log^5 d}.$$
Under Assumption 2, we have for $t \leq \log d$,
$$\mathcal{R}(t\log d) - \tfrac{1}{2}\mathbb{E}[\varepsilon^2] = \Theta\left(t^{-\frac{2(\alpha+\beta)-1}{\alpha+2\beta}} + r_s^{-2(\alpha+\beta)+1}\right) + o_d(1).$$
with probability $1 - d^{-K}$ *for any* $K \in \mathbb{N}$ *and* $d \geq \Omega_K(1)$.

Proof. We have
$$\mathcal{R}(t\log d) - \tfrac{1}{2}\mathbb{E}[\varepsilon^2] = \|\boldsymbol{T}^* - \boldsymbol{M}_t\|_F^2$$
where by Proposition 11,
$$\|\boldsymbol{T}_{11}^* - \boldsymbol{M}_{t,11}\|_F^2 \leq \|\boldsymbol{T}^* - \boldsymbol{M}_t\|_F^2 \leq \|\boldsymbol{T}_{11}^* - \boldsymbol{M}_{t,11}\|_F^2 + 4\|\boldsymbol{M}_{t,11}\|_2 \operatorname{tr}(\boldsymbol{M}_{t,22})$$

$$+ 2(\|\boldsymbol{T}_{22}^*\|_F^2 + 2\|\boldsymbol{T}_{12}^*\|_F^2 + \|\boldsymbol{M}_{t,22}\|_F^2)$$
$$\overset{(a)}{\leq} \|\boldsymbol{T}_{11}^* - \boldsymbol{M}_{t,11}\|_F^2 + o_d(1)$$

where we consider the block matrix decomposition in (C.1) and used the lower block bound in Proposition 3 for (a). We will use idealized dynamics with Theorem 5 with $\rho_d = d$. By Theorem 5, we have for $t \leq \log^2 d$,

$$\|\boldsymbol{T}_{11}^* - \boldsymbol{M}_t\|_F^2 = \|\boldsymbol{T}_{11}^* - \boldsymbol{L}_t\|_F^2 + o_d(1).$$

By using the Riccati ODE's closed-form (see [BEVW25, Lemma 2]), the statement follows. □

Theorem 4. *We define*

$$r_u = \log^4 d, \qquad J_d = 1 + \sum_{i=1}^{r_u} \sigma_i + \sum_{i=1}^{r_u} \lambda_i \sigma_i,$$

and consider (SGD) *with*

$$r \gg r_s \gg \log^{18} d, \qquad \epsilon = \frac{1}{J_d \log^6 d}, \qquad \eta = \frac{1}{J_d p_d \log^9 d},$$

where

$$d \geq \rho_d \gg \max\{\mathrm{tr}(\boldsymbol{\Sigma}) \log^5 d, \lambda_{r_u}^{-1}\}.$$

Under Assumption 2, we have for $\eta t \leq \log^2 d$

$$\mathcal{R}(t \log d) = \Theta\left((\eta t)^{-\frac{2(\alpha+\beta)-1}{\alpha+2\beta}} + r_s^{-2(\alpha+\beta)+1}\right) + o_d(1).$$

with probability $1 - 3d^{-K}$ for any $K \in \mathbb{N}$ and $d \geq \Omega_K(1)$.

Proof. The proof is very similar to the proof of Theorem 3. We consider the block matrix decomposition in (C.1) with idealized dynamics defined in L-DYN. We used the lower block bound in Proposition 6. We use idealized dynamics with Corollary 3, and the remainig proof is identical to above. □

B.2 Proof of Theorem 2

The $\beta > 1$ regime follows from the usual source/capacity characterization of asymptotic limit [LZL23, Theorem 3.2]. For $\beta < 1$, we will assume Gaussian equivalence [WL25] and use the asymptotic risk formula [CM22] in (E.2). Proposition 10 and Lemma 3 characterizes the stated risk in the theorem.

C Continous-time analysis

In this section, we consider gradient flow:

$$\frac{d\boldsymbol{W}_t}{dt} = -r_s \nabla_{\boldsymbol{W}} R(\boldsymbol{W}_t, b_t) = \boldsymbol{\Sigma}\Big(\boldsymbol{T} - \frac{1}{r_s}\boldsymbol{W}_t\boldsymbol{W}_t^\top\Big)\boldsymbol{\Sigma}\boldsymbol{W}_t - \frac{1}{2}\Big(\mathrm{tr}(\tfrac{1}{r_s}\boldsymbol{W}_t\boldsymbol{W}_t^\top\boldsymbol{\Sigma}) - b_t\Big)\boldsymbol{\Sigma}\boldsymbol{W}_t$$

$$\frac{db_t}{dt} = -\frac{1}{\epsilon}\nabla_b R(\boldsymbol{W}_t, b_t) = \frac{-1}{4\epsilon}\Big(\mathrm{tr}(\tfrac{1}{r_s}\boldsymbol{W}_t\boldsymbol{W}_t^\top\boldsymbol{\Sigma}) - b_t\Big)$$

The effective dynamics reads: For $\boldsymbol{G}_0 = \frac{1}{r_s}\boldsymbol{W}_0\boldsymbol{W}_0^\top$,

$$\frac{d\boldsymbol{G}_t}{dt} = (\boldsymbol{\Sigma}^{1/2}\boldsymbol{T}^*\boldsymbol{\Sigma}^{1/2} - 0.5q_t\boldsymbol{\Sigma})\boldsymbol{G}_t + \boldsymbol{G}_t(\boldsymbol{\Sigma}^{1/2}\boldsymbol{T}^*\boldsymbol{\Sigma}^{1/2} - 0.5q_t\boldsymbol{\Sigma}) - \boldsymbol{\Sigma}\boldsymbol{G}_t\boldsymbol{\Sigma}\boldsymbol{G}_t - \boldsymbol{G}_t\boldsymbol{\Sigma}\boldsymbol{G}_t\boldsymbol{\Sigma},$$

We will consider $\boldsymbol{M}_t := \boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{G}_t\boldsymbol{\Sigma}^{\frac{1}{2}}$ with the dynamics:

$$\frac{d\boldsymbol{M}_t}{dt} = \boldsymbol{\Sigma}(\boldsymbol{T}^* - 0.5q_t\boldsymbol{I}_{r_u})\boldsymbol{M}_t + \boldsymbol{M}_t(\boldsymbol{T}^* - 0.5q_t\boldsymbol{I}_{r_u})\boldsymbol{\Sigma} - \boldsymbol{\Sigma}\boldsymbol{M}_t^2 - \boldsymbol{M}_t^2\boldsymbol{\Sigma}.$$

We write

$$\boldsymbol{W}_t =: \begin{bmatrix} \boldsymbol{W}_{t,1} \\ \boldsymbol{W}_{t,2} \end{bmatrix}$$

where $\boldsymbol{W}_{t,1} \in \mathbb{R}^{r_u \times r_s}$. Moreover,

$$\boldsymbol{\Sigma} =: \begin{bmatrix} \boldsymbol{\Sigma}_{11} & 0 \\ 0 & \boldsymbol{\Sigma}_{22} \end{bmatrix} \text{ and } \boldsymbol{T}^* =: \begin{bmatrix} \boldsymbol{T}^*_{11} & \boldsymbol{T}^*_{12} \\ \boldsymbol{T}^*_{21} & \boldsymbol{T}^*_{22} \end{bmatrix} \text{ and } \boldsymbol{A} =: \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{bmatrix} \text{ and } \boldsymbol{M}_t =: \begin{bmatrix} \boldsymbol{M}_{t,11} & \boldsymbol{M}_{t,12} \\ \boldsymbol{M}_{t,12}^\top & \boldsymbol{M}_{t,22} \end{bmatrix} \tag{C.1}$$

where $\boldsymbol{T}^*_{11} \in \mathbb{R}^{r_u \times r_u}$ and $\boldsymbol{M}_{t,11} \in \mathbb{R}^{r_u \times r_u}$ and $\boldsymbol{T}^*_{22}, \boldsymbol{M}_{t,22} \in \mathbb{R}^{(d-r_u)\times(d-r_u)}$. We have

$$\boldsymbol{M}_t^2 = \begin{bmatrix} \boldsymbol{M}_{t,11}^2 + \boldsymbol{M}_{t,12}\boldsymbol{M}_{t,12}^\top & \boldsymbol{M}_{t,11}\boldsymbol{M}_{t,12} + \boldsymbol{M}_{t,12}\boldsymbol{M}_{t,22} \\ \boldsymbol{M}_{t,12}^\top\boldsymbol{M}_{t,11}^\top + \boldsymbol{M}_{t,22}^\top\boldsymbol{M}_{t,12}^\top & \boldsymbol{M}_{t,12}^\top\boldsymbol{M}_{t,12} + \boldsymbol{M}_{t,22}^2 \end{bmatrix}$$

We start with the following statement:

Remark 3. *Under our initialization assumptions, one can verify that choosing $\epsilon = o_d(\log^{-5} d)$ ensures*

$$\sup_{0\le t\le \log^2 d} \int_0^t |q_s|\, ds = o_d(\log^{-3} d).$$

The change of variables

$$\widetilde{\boldsymbol{M}}_t = \exp\left(\frac{1}{2}\boldsymbol{\Sigma}\int_0^t q_s\, ds\right)\boldsymbol{M}_t \exp\left(\frac{1}{2}\boldsymbol{\Sigma}\int_0^t q_s\, ds\right)$$

eliminates the q_t-term from the dynamics and perturbs the coefficients of the quadratic terms by $o_d(\log^{-3} d)$, uniformly on this time interval. Using the integral bound above, the centering correction can be absorbed into the error estimates below without changing their asymptotic rates. We omit the details of this reduction and, for brevity, present the proof for the centered case.

C.1 Block matrix analysis

We have

$$\frac{d\boldsymbol{M}_{t,11}}{dt} = \boldsymbol{A}_{11}\boldsymbol{M}_{t,11} + \boldsymbol{M}_{t,11}\boldsymbol{A}_{11} - \boldsymbol{\Sigma}_{11}(\boldsymbol{M}_{t,11}^2 + \boldsymbol{M}_{t,12}\boldsymbol{M}_{t,12}^\top) - (\boldsymbol{M}_{t,11}^2 + \boldsymbol{M}_{t,12}\boldsymbol{M}_{t,12}^\top)\boldsymbol{\Sigma}_{11}$$

and

$$\frac{d\boldsymbol{M}_{t,22}}{dt} = \boldsymbol{A}_{22}\boldsymbol{M}_{t,22} + \boldsymbol{M}_{t,22}\boldsymbol{A}_{22}^\top - \boldsymbol{\Sigma}_{22}(\boldsymbol{M}_{t,22}^2 + \boldsymbol{M}_{t,12}^\top\boldsymbol{M}_{t,12}) - (\boldsymbol{M}_{t,22}^2 + \boldsymbol{M}_{t,12}^\top\boldsymbol{M}_{t,12})\boldsymbol{\Sigma}_{22}$$

We define

$$\frac{d\boldsymbol{L}_t}{dt} = \boldsymbol{A}_{11}\boldsymbol{L}_t + \boldsymbol{L}_t\boldsymbol{A}_{11} - 2\boldsymbol{L}_t\boldsymbol{\Sigma}_{11}\boldsymbol{L}_t,$$

with a diagonal initialization. Whenever we want to make the initalization explicit we will write $\boldsymbol{L}_t(\boldsymbol{L}_0)$.

Bounding lower block. We have the following proposition.

Proposition 3 (Lower block)**.** *We have*

- $\|\boldsymbol{M}_{t,22}\|_2 \le C(1 + \log\frac{\lambda_{max}(\boldsymbol{\Sigma}_{22})}{\lambda_{min}(\boldsymbol{\Sigma}_{22})})\|\boldsymbol{M}_{0,22}\|_2 \exp(2\|\boldsymbol{A}_{22}\|_2 t)$.
- $\mathrm{tr}(\boldsymbol{M}_{t,22}) \le \mathrm{tr}(\boldsymbol{M}_{0,22})\exp(2\|\boldsymbol{A}_{22}\|_2 t)$.

Proof. For the second item, we have

$$\frac{d\,\mathrm{tr}(\boldsymbol{M}_{t,22})}{dt} \le 2\,\mathrm{tr}(\boldsymbol{M}_{t,22})\|\boldsymbol{A}_{22}\|_2 \Rightarrow \mathrm{tr}(\boldsymbol{M}_{t,22}) \le \mathrm{tr}(\boldsymbol{M}_{0,22})\exp(2\|\boldsymbol{A}_{22}\|_2 t).$$

For the first item, we define

$$\boldsymbol{X}_t \coloneqq e^{-\boldsymbol{A}_{22}t}\boldsymbol{M}_t e^{-\boldsymbol{A}_{22}t} \text{ and } \boldsymbol{J}_t \coloneqq \int_0^t e^{-\boldsymbol{A}_{22}s}(\boldsymbol{M}_{s,22}^2 + \boldsymbol{M}_{s,12}^\top \boldsymbol{M}_{s,12})e^{-\boldsymbol{A}_{22}s}\,ds \succeq 0.$$

We have

$$\boldsymbol{X}_t = \boldsymbol{M}_{0,22} - \boldsymbol{\Sigma}_{22}\boldsymbol{J}_t - \boldsymbol{J}_t\boldsymbol{\Sigma}_{22}$$

We define

$$\Phi_{\boldsymbol{\Sigma}_{22}}(\boldsymbol{Y}) \coloneqq \int_0^\infty e^{-u\boldsymbol{\Sigma}_{22}}\boldsymbol{Y}e^{-u\boldsymbol{\Sigma}_{22}}\,du, \text{ where } (\Phi(Y))_{ij} = \frac{\boldsymbol{Y}_{ij}}{\sigma_i + \sigma_j}$$

Since $\Phi(\boldsymbol{Y})$ is a linear operator, we have

$$\Phi_{\boldsymbol{\Sigma}_{22}}(\boldsymbol{X}_t) = \Phi_{\boldsymbol{\Sigma}_{22}}(\boldsymbol{M}_{0,22}) - \Phi_{\boldsymbol{\Sigma}_{22}}(\boldsymbol{\Sigma}_{22}\boldsymbol{J}_t + J_t\boldsymbol{\Sigma}_{22}) = \Phi_{\boldsymbol{\Sigma}_{22}}(\boldsymbol{M}_{0,22}) - \boldsymbol{J}_t \tag{C.2}$$

We observe that for general psd matrix Y,

$$\|\Phi_{\boldsymbol{\Sigma}_{22}}(\boldsymbol{Y})\|_2 \le \|\boldsymbol{Y}\|_2 \Big\| \int_0^\infty e^{-2u\boldsymbol{\Sigma}_{22}}\,du \Big\|_2 = \frac{\|\boldsymbol{Y}\|_2}{2\lambda_{min}(\boldsymbol{\Sigma}_{22})}. \tag{C.3}$$

To avoid suffering $1/\lambda_{min}(\boldsymbol{\Sigma}_{22})$, we partition the lower-block coordinates into spectral blocks of $\boldsymbol{\Sigma}_{22}$, with each part containing eigenvalues that differ by at most a factor of 2. Let P_j denote the corresponding diagonal projections. For one block, write

$$\boldsymbol{X}_{t,j} = \boldsymbol{P}_j\boldsymbol{X}_t\boldsymbol{P}_j, \qquad \boldsymbol{M}_{t,j} = \boldsymbol{P}_j\boldsymbol{M}_{t,22}\boldsymbol{P}_j,$$

and let $\boldsymbol{\Sigma}_j$ be the restriction of $\boldsymbol{\Sigma}_{22}$ to that block. By (C.2), we have

$$0 \preceq \Phi_{\boldsymbol{\Sigma}_j}(X_{t,j}) \preceq \Phi_{\boldsymbol{\Sigma}_j}(\boldsymbol{M}_{0,j}).$$

Then, by writing

$$X_{t,j} = \boldsymbol{\Sigma}_j\Phi_{\boldsymbol{\Sigma}_j}(X_{t,j}) + \Phi_{\boldsymbol{\Sigma}_j}(X_{t,j})\boldsymbol{\Sigma}_j$$

we have

$$\|\boldsymbol{M}_{t,j}\|_2 \le 2\|\boldsymbol{\Sigma}_j\|_2\|\Phi_{\boldsymbol{\Sigma}_j}(\boldsymbol{M}_{0,j})\|e^{2\|\boldsymbol{A}_{22}\|_2 t} \le 2\|\boldsymbol{M}_{0,j}\|_2 e^{2\|\boldsymbol{A}_{22}\|_2 t}.$$

where we used (C.3) in the last inequality. Then,

$$\begin{aligned}
\|\boldsymbol{M}_{t,22}\|_2 \le \sum_j \|\boldsymbol{M}_{t,j}\|_2 &\le 2e^{2\|\boldsymbol{A}_{22}\|_2 t}\sum_j \|\boldsymbol{M}_{0,j}\|_2 \\
&\le 2\Big(1 + \lceil \log_2 \frac{\lambda_{max}(\boldsymbol{\Sigma}_{22})}{\lambda_{min}(\boldsymbol{\Sigma}_{22})} \rceil\Big)\|\boldsymbol{M}_{0,22}\|_2 \exp(2\|\boldsymbol{A}_{22}\|_2 t) \\
&\le C\Big(1 + \log \frac{\lambda_{max}(\boldsymbol{\Sigma}_{22})}{\lambda_{min}(\boldsymbol{\Sigma}_{22})}\Big)\|\boldsymbol{M}_{0,22}\|_2 \exp(2\|\boldsymbol{A}_{22}\|_2 t).
\end{aligned}$$

□

Bounding upper block. We start with the following proposition.

Proposition 4. *We define* $\boldsymbol{R}_t \coloneqq \boldsymbol{L}_t^{\frac{-1}{2}}\boldsymbol{M}_{t,11}\boldsymbol{L}_t^{\frac{-1}{2}}$. *We have*

$$\begin{aligned}
\frac{d\boldsymbol{R}_t}{dt} &= \boldsymbol{\Sigma}_{11}\boldsymbol{L}_t\boldsymbol{R}_t + \boldsymbol{R}_t\boldsymbol{\Sigma}_{11}\boldsymbol{L}_t - 2\boldsymbol{R}_t\boldsymbol{\Sigma}_{11}\boldsymbol{L}_t\boldsymbol{R}_t + [\boldsymbol{R}_t, \boldsymbol{\Sigma}_{11}]\boldsymbol{L}_t\boldsymbol{R}_t + \boldsymbol{R}_t\boldsymbol{L}_t[\boldsymbol{R}_t, \boldsymbol{\Sigma}_{11}]^\top \\
&\quad - \frac{1}{r_s^2}\boldsymbol{\Sigma}_{11}\boldsymbol{L}_t^{\frac{-1}{2}}(\boldsymbol{\Sigma}_{11}^{\frac{1}{2}}\boldsymbol{W}_{t,1}\boldsymbol{W}_{t,2}^\top\boldsymbol{\Sigma}_{22}\boldsymbol{W}_{t,2}\boldsymbol{W}_{t,1}^\top\boldsymbol{\Sigma}_{11}^{\frac{1}{2}})\boldsymbol{L}_t^{\frac{-1}{2}} \\
&\quad - \frac{1}{r_s^2}\boldsymbol{L}_t^{\frac{-1}{2}}(\boldsymbol{\Sigma}_{11}^{\frac{1}{2}}\boldsymbol{W}_{t,1}\boldsymbol{W}_{t,2}^\top\boldsymbol{\Sigma}_{22}\boldsymbol{W}_{t,2}\boldsymbol{W}_{t,1}^\top\boldsymbol{\Sigma}_{11}^{\frac{1}{2}})\boldsymbol{L}_t^{\frac{-1}{2}}\boldsymbol{\Sigma}_{11}
\end{aligned}$$

Proof. We define $\boldsymbol{\Omega}_t \coloneqq [\boldsymbol{M}_{t,11}, \boldsymbol{\Sigma}_{11}]$. We observe that

$$\begin{aligned}
\frac{d\boldsymbol{M}_{t,11}}{dt} &= \boldsymbol{A}_{11}\boldsymbol{M}_{t,11} + \boldsymbol{M}_{t,11}\boldsymbol{A}_{11} - 2\boldsymbol{M}_{t,11}\boldsymbol{\Sigma}_{11}\boldsymbol{M}_{t,11}^\top + [[\boldsymbol{M}_{t,11}, \boldsymbol{\Sigma}_{11}], \boldsymbol{M}_{t,11}] \\
&\quad - \frac{1}{r_s^2}\boldsymbol{\Sigma}_{11}(\boldsymbol{\Sigma}_{11}^{\frac{1}{2}}\boldsymbol{W}_{t,1}\boldsymbol{W}_{t,2}^\top\boldsymbol{\Sigma}_{22}\boldsymbol{W}_{t,2}\boldsymbol{W}_{t,1}^\top\boldsymbol{\Sigma}_{11}^{\frac{1}{2}}) \\
&\quad - \frac{1}{r_s^2}(\boldsymbol{\Sigma}_{11}^{\frac{1}{2}}\boldsymbol{W}_{t,1}\boldsymbol{W}_{t,2}^\top\boldsymbol{\Sigma}_{22}\boldsymbol{W}_{t,2}\boldsymbol{W}_{t,1}^\top\boldsymbol{\Sigma}_{11}^{\frac{1}{2}})\boldsymbol{\Sigma}_{11} \\
&= (\boldsymbol{A}_{11} + \boldsymbol{\Omega}_t)\boldsymbol{M}_{t,11} + \boldsymbol{M}_{t,11}(\boldsymbol{A}_{11} + \boldsymbol{\Omega}_t^\top) - 2\boldsymbol{M}_{t,11}\boldsymbol{\Sigma}_{11}\boldsymbol{M}_{t,11}^\top \\
&\quad - \frac{1}{r_s^2}\boldsymbol{\Sigma}_{11}(\boldsymbol{\Sigma}_{11}^{\frac{1}{2}}\boldsymbol{W}_{t,1}\boldsymbol{W}_{t,2}^\top\boldsymbol{\Sigma}_{22}\boldsymbol{W}_{t,2}\boldsymbol{W}_{t,1}^\top\boldsymbol{\Sigma}_{11}^{\frac{1}{2}}) \\
&\quad - \frac{1}{r_s^2}(\boldsymbol{\Sigma}_{11}^{\frac{1}{2}}\boldsymbol{W}_{t,1}\boldsymbol{W}_{t,2}^\top\boldsymbol{\Sigma}_{22}\boldsymbol{W}_{t,2}\boldsymbol{W}_{t,1}^\top\boldsymbol{\Sigma}_{11}^{\frac{1}{2}})\boldsymbol{\Sigma}_{11}
\end{aligned}$$

We have

$$\begin{aligned}
\frac{d\boldsymbol{R}_t}{dt} &= -\frac{1}{2}\frac{d\boldsymbol{L}_t}{dt}\boldsymbol{L}_t^{-1}\boldsymbol{R}_t - \frac{1}{2}\boldsymbol{R}_t\boldsymbol{L}_t^{-1}\frac{d\boldsymbol{L}_t}{dt} + \boldsymbol{L}_t^{\frac{-1}{2}}\frac{d\boldsymbol{M}_{t,11}}{dt}\boldsymbol{L}_t^{\frac{-1}{2}} \\
&= -\big(\boldsymbol{A}_{11} - \boldsymbol{L}_t\boldsymbol{\Sigma}_{11}\big)\boldsymbol{R}_t - \boldsymbol{R}_t\big(\boldsymbol{A}_{11} - \boldsymbol{L}_t\boldsymbol{\Sigma}_{11}\big) \\
&\quad + (\boldsymbol{A}_{11} + \boldsymbol{L}_t^{\frac{-1}{2}}\boldsymbol{\Omega}_t\boldsymbol{L}_t^{\frac{1}{2}})\boldsymbol{R}_t + \boldsymbol{R}_t(\boldsymbol{A}_{11} + \boldsymbol{L}_t^{\frac{1}{2}}\boldsymbol{\Omega}_t^\top\boldsymbol{L}_t^{\frac{-1}{2}}) - 2\boldsymbol{R}_t\boldsymbol{L}_t\boldsymbol{\Sigma}_{11}\boldsymbol{R}_t \\
&\quad - \frac{1}{r_s^2}\boldsymbol{\Sigma}_{11}\boldsymbol{L}_t^{\frac{-1}{2}}(\boldsymbol{\Sigma}_{11}^{\frac{1}{2}}\boldsymbol{W}_{t,1}\boldsymbol{W}_{t,2}^\top\boldsymbol{\Sigma}_{22}\boldsymbol{W}_{t,2}\boldsymbol{W}_{t,1}^\top\boldsymbol{\Sigma}_{11}^{\frac{1}{2}})\boldsymbol{L}_t^{\frac{-1}{2}} \\
&\quad - \frac{1}{r_s^2}\boldsymbol{L}_t^{\frac{-1}{2}}(\boldsymbol{\Sigma}_{11}^{\frac{1}{2}}\boldsymbol{W}_{t,1}\boldsymbol{W}_{t,2}^\top\boldsymbol{\Sigma}_{22}\boldsymbol{W}_{t,2}\boldsymbol{W}_{t,1}^\top\boldsymbol{\Sigma}_{11}^{\frac{1}{2}})\boldsymbol{L}_t^{\frac{-1}{2}}\boldsymbol{\Sigma}_{11} \\
&= \boldsymbol{L}_t\boldsymbol{\Sigma}_{11}\boldsymbol{R}_t + \boldsymbol{R}_t\boldsymbol{L}_t\boldsymbol{\Sigma}_{11} + \boldsymbol{L}_t^{\frac{-1}{2}}\boldsymbol{\Omega}_t\boldsymbol{L}_t^{\frac{1}{2}}\boldsymbol{R}_t + \boldsymbol{R}_t\boldsymbol{L}_t^{\frac{1}{2}}\boldsymbol{\Omega}_t^\top\boldsymbol{L}_t^{\frac{-1}{2}} - 2\boldsymbol{R}_t\boldsymbol{L}_t\boldsymbol{\Sigma}_{11}\boldsymbol{R}_t \\
&\quad - \frac{1}{r_s^2}\boldsymbol{\Sigma}_{11}\boldsymbol{L}_t^{\frac{-1}{2}}(\boldsymbol{\Sigma}_{11}^{\frac{1}{2}}\boldsymbol{W}_{t,1}\boldsymbol{W}_{t,2}^\top\boldsymbol{\Sigma}_{22}\boldsymbol{W}_{t,2}\boldsymbol{W}_{t,1}^\top\boldsymbol{\Sigma}_{11}^{\frac{1}{2}})\boldsymbol{L}_t^{\frac{-1}{2}} \\
&\quad - \frac{1}{r_s^2}\boldsymbol{L}_t^{\frac{-1}{2}}(\boldsymbol{\Sigma}_{11}^{\frac{1}{2}}\boldsymbol{W}_{t,1}\boldsymbol{W}_{t,2}^\top\boldsymbol{\Sigma}_{22}\boldsymbol{W}_{t,2}\boldsymbol{W}_{t,1}^\top\boldsymbol{\Sigma}_{11}^{\frac{1}{2}})\boldsymbol{L}_t^{\frac{-1}{2}}\boldsymbol{\Sigma}_{11}.
\end{aligned}$$

By observing that $\boldsymbol{L}_t^{\frac{-1}{2}}\boldsymbol{\Omega}_t\boldsymbol{L}_t^{\frac{-1}{2}} = [\boldsymbol{R}_t, \boldsymbol{\Sigma}_{11}]$, we have the result. □

Corollary 2. *We have*

$$\begin{aligned}
\frac{d\boldsymbol{R}_t}{dt} &= -\boldsymbol{\Sigma}_{11}(\boldsymbol{R}_t - \boldsymbol{I}_{r_u})\boldsymbol{L}_t - \boldsymbol{L}_t(\boldsymbol{R}_t - \boldsymbol{I}_{r_u})\boldsymbol{\Sigma}_{11} \\
&\quad - \boldsymbol{\Sigma}_{11}(\boldsymbol{R}_t - \boldsymbol{I}_{r_u})\boldsymbol{L}_t(\boldsymbol{R}_t - \boldsymbol{I}_{r_u}) - (\boldsymbol{R}_t - \boldsymbol{I}_{r_u})\boldsymbol{L}_t(\boldsymbol{R}_t - \boldsymbol{I}_{r_u})\boldsymbol{\Sigma}_{11} \\
&\quad - \frac{1}{r_s^2}\boldsymbol{\Sigma}_{11}\boldsymbol{L}_t^{\frac{-1}{2}}(\boldsymbol{\Sigma}_{11}^{\frac{1}{2}}\boldsymbol{W}_{t,1}\boldsymbol{W}_{t,2}^\top\boldsymbol{\Sigma}_{22}\boldsymbol{W}_{t,2}\boldsymbol{W}_{t,1}^\top\boldsymbol{\Sigma}_{11}^{\frac{1}{2}})\boldsymbol{L}_t^{\frac{-1}{2}} \\
&\quad - \frac{1}{r_s^2}\boldsymbol{L}_t^{\frac{-1}{2}}(\boldsymbol{\Sigma}_{11}^{\frac{1}{2}}\boldsymbol{W}_{t,1}\boldsymbol{W}_{t,2}^\top\boldsymbol{\Sigma}_{22}\boldsymbol{W}_{t,2}\boldsymbol{W}_{t,1}^\top\boldsymbol{\Sigma}_{11}^{\frac{1}{2}})\boldsymbol{L}_t^{\frac{-1}{2}}\boldsymbol{\Sigma}_{11}.
\end{aligned}$$

Proof. We have

$$\begin{aligned}
&\boldsymbol{\Sigma}_{11}\boldsymbol{L}_t\boldsymbol{R}_t + \boldsymbol{R}_t\boldsymbol{\Sigma}_{11}\boldsymbol{L}_t - 2\boldsymbol{R}_t\boldsymbol{\Sigma}_{11}\boldsymbol{L}_t\boldsymbol{R}_t \\
&+ (\boldsymbol{R}_t\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{11}\boldsymbol{R}_t)\boldsymbol{L}_t\boldsymbol{R}_t + \boldsymbol{R}_t\boldsymbol{L}_t(\boldsymbol{\Sigma}_{11}\boldsymbol{R}_t - \boldsymbol{R}_t\boldsymbol{\Sigma}_{11}) \\
&= -\boldsymbol{\Sigma}_{11}(\boldsymbol{R}_t - \boldsymbol{I}_{r_u})\boldsymbol{L}_t\boldsymbol{R}_t - \boldsymbol{R}_t\boldsymbol{L}_t(\boldsymbol{R}_t - \boldsymbol{I}_{r_u})\boldsymbol{\Sigma}_{11} \\
&= -\boldsymbol{\Sigma}_{11}(\boldsymbol{R}_t - \boldsymbol{I}_{r_u})\boldsymbol{L}_t - \boldsymbol{L}_t(\boldsymbol{R}_t - \boldsymbol{I}_{r_u})\boldsymbol{\Sigma}_{11} \\
&\quad - \boldsymbol{\Sigma}_{11}(\boldsymbol{R}_t - \boldsymbol{I}_{r_u})\boldsymbol{L}_t(\boldsymbol{R}_t - \boldsymbol{I}_{r_u}) - (\boldsymbol{R}_t - \boldsymbol{I}_{r_u})\boldsymbol{L}_t(\boldsymbol{R}_t - \boldsymbol{I}_{r_u})\boldsymbol{\Sigma}_{11}.
\end{aligned}$$

□

Theorem 5. *Let $e_t \coloneqq \|\boldsymbol{R}_t - \boldsymbol{I}_{r_u}\|_2$ and we consider $\alpha + \beta > 0.5$,*

$$c_d = \frac{1}{\log^4 d}, \qquad r_u = \log^{10} d, \qquad r_s \gg \frac{r_u}{c_d^2}, \qquad \rho_d \gg \mathrm{tr}(\boldsymbol{\Sigma}) \log^5 d, \qquad \boldsymbol{L}_0 = \frac{\boldsymbol{\Sigma}_{11}}{\rho_d}.$$

$\mathcal{G}_{init}$ implies that for $d \geq \Omega_K(1)$,

$$\max_{0 \leq t \leq \log^2 d} e_t \leq \frac{2C(1 + \beta \log\log d)^2}{\log^3 d}.$$

Proof. We start by observing that by (E.2) and Proposition 3 for $t \leq \log^2 d$,

$$\sup_{s \leq t} \|\boldsymbol{M}_{s,22}\|_2 \leq C_K\big(\frac{\log d}{r_s} + \frac{\|\boldsymbol{\Sigma}_{22}\|_2}{\rho_d}\big) \leq \frac{C_K}{\log^5 d}$$

We define

$$\mathcal{T}_{hit} = \inf\{t \geq 0 : e_t > 2\} \wedge \log^2 d.$$

We observe that for $t \leq \log^2 d$,

$$K_U^2(e_0 + 6 \sup_{s \leq t} \|\boldsymbol{M}_{s,22}\|_2 t) \leq \underbrace{K_U^2(e_0 + 6 \sup_{s \leq \log^2 d} \|\boldsymbol{M}_{s,22}\|_2 \log^2 d)}_{\coloneqq e_*} < \frac{C_K(1 + \beta \log\log d)}{\log^3 d}$$

where we used $e_0 < c_d$ and K_U is defined in Proposition 12. Let

$$\mathcal{A}_t(\boldsymbol{X}) \coloneqq -\boldsymbol{\Sigma}_{11}\boldsymbol{X}\boldsymbol{L}_t - \boldsymbol{L}_t\boldsymbol{X}\boldsymbol{\Sigma}_{11}.$$

We have

$$\mathcal{A}_t(\boldsymbol{X})_{ij} = (-\sigma_i \ell_{t,j} - \sigma_j \ell_{t,i}) X_{ij}$$

We define

$$(U_{t,s}[\boldsymbol{X}])_{ij} \coloneqq \exp\Big(- \sigma_i \int_s^t \ell_{u,j}\, du - \sigma_j \int_s^t \ell_{u,i}\, du\Big) X_{ij}$$

and

$$\begin{aligned}\boldsymbol{Q}_t = & -\frac{1}{r_s^2}\boldsymbol{\Sigma}_{11}\boldsymbol{L}_t^{\frac{-1}{2}}(\boldsymbol{\Sigma}_{11}^{\frac{1}{2}}\boldsymbol{W}_{t,1}\boldsymbol{W}_{t,2}^\top\boldsymbol{\Sigma}_{22}\boldsymbol{W}_{t,2}\boldsymbol{W}_{t,1}^\top\boldsymbol{\Sigma}_{11}^{\frac{1}{2}})\boldsymbol{L}_t^{\frac{-1}{2}} \\ & -\frac{1}{r_s^2}\boldsymbol{L}_t^{\frac{-1}{2}}(\boldsymbol{\Sigma}_{11}^{\frac{1}{2}}\boldsymbol{W}_{t,1}\boldsymbol{W}_{t,2}^\top\boldsymbol{\Sigma}_{22}\boldsymbol{W}_{t,2}\boldsymbol{W}_{t,1}^\top\boldsymbol{\Sigma}_{11}^{\frac{1}{2}})\boldsymbol{L}_t^{\frac{-1}{2}}\boldsymbol{\Sigma}_{11}.\end{aligned}$$

Let $\hat{\boldsymbol{R}}_t \coloneqq \boldsymbol{R}_t - \boldsymbol{I}_{r_u}$. We have

$$\hat{\boldsymbol{R}}_t = U_{t,0}(\hat{\boldsymbol{R}}_0) - \int_0^t U_{t,s}(\boldsymbol{\Sigma}_{11}\hat{\boldsymbol{R}}_s\boldsymbol{L}_s\hat{\boldsymbol{R}}_s + \hat{\boldsymbol{R}}_s\boldsymbol{L}_s\hat{\boldsymbol{R}}_s\boldsymbol{\Sigma}_{11})ds + \int_0^t U_{t,s}(\boldsymbol{Q}_s)ds.$$

Therefore, by using Proposition 12 for $t \leq \mathcal{T}_{hit}$

$$e_t \leq K_U^2\Big(e_0 + 2\sup_{s \leq t}\|\boldsymbol{M}_{s,22}\|_2 \int_0^t (1 + e_s)ds + 2\int_0^t e_s^2 ds\Big) \leq K_U^2\Big(e_* + 2\int_0^t e_s^2 ds\Big).$$

By using Proposition 14, for large enough d, we have

$$e_t \leq \frac{K_U^2 e_*}{1 - 2K_U^4 e_* t} \leq \frac{2C_K(1 + \beta\log\log d)}{\log^3 d}.$$

Hence, $\mathcal{T}_{hit} = \log^2 d$ and we have the result. □

D SGD Analysis

We define $\mathrm{He}_2(\boldsymbol{z}_t) \coloneqq \frac{1}{\sqrt{2}}(\boldsymbol{z}_t\boldsymbol{z}_t^\top - \boldsymbol{I}_d)$. We recall that

$$y_t = \langle \boldsymbol{T}^*, \mathrm{He}_2(\boldsymbol{z}_t)\rangle + \varepsilon_t, \qquad \hat{y}_{t+1} = \langle \boldsymbol{M}_t, \mathrm{He}_2(\boldsymbol{z}_{t+1})\rangle - \tfrac{1}{\sqrt{2}}q_t.$$

For notational convenience we introduce $\bar{\boldsymbol{W}}_t = \boldsymbol{\Sigma}^{1/2}\boldsymbol{W}_t$. The online SGD updates with instantaneous loss $L(\boldsymbol{W}_t, b_t) \coloneqq 0.25(y_{t+1} - \hat{y}_{t+1})^2$ is

$$\begin{aligned}\boldsymbol{W}_{t+1} &= \boldsymbol{W}_t + \frac{\eta}{\rho_d}(y_{t+1} - \hat{y}_{t+1})\tfrac{1}{\sqrt{2}}\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}^\top\boldsymbol{W}_t \\ &= \boldsymbol{W}_t + \frac{\eta}{\rho_d}\boldsymbol{\Sigma}^{\frac{1}{2}}(y_{t+1} - \hat{y}_{t+1})\tfrac{1}{\sqrt{2}}\boldsymbol{z}_{t+1}\boldsymbol{z}_{t+1}^\top\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{W}_t\end{aligned}$$

and for $b_0 = 0$,

$$b_{t+1} = b_t + \frac{1}{2\sqrt{2}\epsilon}\frac{\eta}{\rho_d}(y_{t+1} - \hat{y}_{t+1}) \tag{D.1}$$

We define

- The symmetric noise term:

$$\begin{aligned}\boldsymbol{N}_{t+1} \coloneqq &\Big(\big(\langle \boldsymbol{T}^* - \boldsymbol{M}_t, \mathrm{He}_2(\boldsymbol{z}_{t+1})\rangle + \tfrac{1}{\sqrt{2}}q_t\big)\tfrac{1}{\sqrt{2}}\boldsymbol{z}_{t+1}\boldsymbol{z}_{t+1}^\top - (\boldsymbol{T}^* - 0.5q_t\boldsymbol{I}_d - \boldsymbol{M}_t)\Big) \\ &+ \tfrac{1}{\sqrt{2}}\varepsilon_{t+1}\boldsymbol{z}_{t+1}\boldsymbol{z}_{t+1}^\top.\end{aligned}$$

- The effective learning rate $\eta \coloneqq \frac{\eta}{\rho_d}$

We have

$$\begin{aligned}\bar{\boldsymbol{W}}_{t+1} &= \bar{\boldsymbol{W}}_t + \eta\boldsymbol{\Sigma}(y_{t+1} - \hat{y}_{t+1})\tfrac{1}{\sqrt{2}}\boldsymbol{z}_{t+1}\boldsymbol{z}_{t+1}^\top\bar{\boldsymbol{W}}_t \\ &= \bar{\boldsymbol{W}}_t + \eta\boldsymbol{\Sigma}(\boldsymbol{T}^* - 0.5q_t\boldsymbol{I}_d - \boldsymbol{M}_t)\bar{\boldsymbol{W}}_t + \eta\boldsymbol{\Sigma}\boldsymbol{N}_{t+1}\bar{\boldsymbol{W}}_t.\end{aligned}$$

Then, we write

$$\begin{aligned}\boldsymbol{M}_{t+1} = \boldsymbol{M}_t &+ \eta\big(\boldsymbol{\Sigma}(\boldsymbol{T}^* - 0.5q_t\boldsymbol{I}_d - \boldsymbol{M}_t)\boldsymbol{M}_t + \boldsymbol{M}_t(\boldsymbol{T}^* - 0.5q_t\boldsymbol{I}_d - \boldsymbol{M}_t)\boldsymbol{\Sigma}\big) \\ &+ \eta(\boldsymbol{\Sigma}\boldsymbol{N}_{t+1}\boldsymbol{M}_t + \boldsymbol{M}_t\boldsymbol{N}_{t+1}\boldsymbol{\Sigma}) \\ &+ \eta^2\boldsymbol{\Sigma}(\boldsymbol{T}^* - 0.5q_t\boldsymbol{I}_d - \boldsymbol{M}_t)\boldsymbol{M}_t(\boldsymbol{T}^* - 0.5q_t\boldsymbol{I}_d - \boldsymbol{M}_t)\boldsymbol{\Sigma} \\ &+ \eta^2\boldsymbol{\Sigma}(\boldsymbol{T}^* - 0.5q_t\boldsymbol{I}_d - \boldsymbol{M}_t)\boldsymbol{M}_t\boldsymbol{N}_{t+1}\boldsymbol{\Sigma} + \eta^2\boldsymbol{\Sigma}\boldsymbol{N}_{t+1}\boldsymbol{M}_t(\boldsymbol{T}^* - 0.5q_t\boldsymbol{I}_d - \boldsymbol{M}_t)\boldsymbol{\Sigma} \\ &+ \eta^2\boldsymbol{\Sigma}\boldsymbol{N}_{t+1}\boldsymbol{M}_t\boldsymbol{N}_{t+1}\boldsymbol{\Sigma}.\end{aligned} \tag{D.2}$$

We define

- We define $\boldsymbol{H}_t \coloneqq \mathbb{E}_t[\boldsymbol{\Sigma}\boldsymbol{N}_{t+1}\boldsymbol{M}_t\boldsymbol{N}_{t+1}\boldsymbol{\Sigma}] + \boldsymbol{\Sigma}(\boldsymbol{T}^* - 0.5q_t\boldsymbol{I}_d - \boldsymbol{M}_t)\boldsymbol{M}_t(\boldsymbol{T}^* - 0.5q_t\boldsymbol{I}_d - \boldsymbol{M}_t)\boldsymbol{\Sigma}$
- The noise term

$$\begin{aligned}\Xi_{t+1} \coloneqq &(\boldsymbol{\Sigma}\boldsymbol{N}_{t+1}\boldsymbol{M}_t + \boldsymbol{M}_t\boldsymbol{N}_{t+1}\boldsymbol{\Sigma}) \\ &+ \eta\Big(\boldsymbol{\Sigma}(\boldsymbol{T}^* - 0.5q_t\boldsymbol{I}_d - \boldsymbol{M}_t)\boldsymbol{M}_t\boldsymbol{N}_{t+1}\boldsymbol{\Sigma} + \boldsymbol{\Sigma}\boldsymbol{N}_{t+1}\boldsymbol{M}_t(\boldsymbol{T}^* - 0.5q_t\boldsymbol{I}_d - \boldsymbol{M}_t)\boldsymbol{\Sigma}\Big) \\ &+ \eta\Big(\boldsymbol{\Sigma}\boldsymbol{N}_{t+1}\boldsymbol{M}_t\boldsymbol{N}_{t+1}\boldsymbol{\Sigma} - \mathbb{E}_t[\boldsymbol{\Sigma}\boldsymbol{N}_{t+1}\boldsymbol{M}_t\boldsymbol{N}_{t+1}\boldsymbol{\Sigma}]\Big)\end{aligned}$$

We have the following.

Proposition 5. *We have*

$$\boldsymbol{H}_t \preceq C(\|\boldsymbol{T}^* - \boldsymbol{M}_t\|_F^2 + \tfrac{q_t^2}{2} + \mathbb{E}_t[\varepsilon_{t+1}^2])\,\mathrm{tr}(\boldsymbol{M}_t)\boldsymbol{\Sigma}^2.$$

Remark 4. *Similar to the continuous-time case, under our initialization and noise moment assumptions, the two-time scale SGD admits the same asymptotic bounds as the centered SGD. By using* $\epsilon = (J_d \log^6 d)^{-1}$*, we can show that* (D.1) *concentrates around its mean for large* d *in the time horizon* $t \le N = \lfloor \log^2 d/\eta \rfloor$*. Then, by applying the change of variables*

$$\mathsf{E}_t = \prod_{s=0}^{t-1} \Big(\boldsymbol{I}_d - \frac{\eta}{2} q_s \boldsymbol{\Sigma} \Big)^{-1}, \qquad \widetilde{\boldsymbol{M}}_t = \mathsf{E}_t \boldsymbol{M}_t \mathsf{E}_t,$$

we can remove the q_t *factor from the weight update* (D.2)*. The resulting drift perturbations contribute* $o_d(\log^{-3} d)$ *over the time horizon* $\eta N \le \log^2 d$ *by the same reasoning as in the continuous-time case. The remaining stochastic and discretization corrections can be controlled by the same concentration arguments used below, without changing the learning rate or asymptotic bounds. We omit the details of this reduction for brevity and assume* $q_t = 0$ *below.*

Therefore, we have

$$\boldsymbol{M}_{t+1} = \boldsymbol{M}_t + \eta\big(\boldsymbol{\Sigma}(\boldsymbol{T}^* - \boldsymbol{M}_t)\boldsymbol{M}_t + \boldsymbol{M}_t(\boldsymbol{T}^* - \boldsymbol{M}_t)\boldsymbol{\Sigma}\big) + \eta^2 \boldsymbol{H}_t + \eta \Xi_{t+1}$$

By using the decomposition structure in (C.1), we write

$$\begin{aligned}\boldsymbol{M}_{t+1,11} = \boldsymbol{M}_{t,11} &+ \eta(\boldsymbol{A}_{11}\boldsymbol{M}_{t,11} + \boldsymbol{M}_{t,11}\boldsymbol{A}_{11}) \\ &- \eta\big(\boldsymbol{\Sigma}_{11}(\boldsymbol{M}_{t,11}^2 + \boldsymbol{M}_{t,12}\boldsymbol{M}_{t,12}^\top) + (\boldsymbol{M}_{t,11}^2 + \boldsymbol{M}_{t,12}\boldsymbol{M}_{t,12}^\top)\boldsymbol{\Sigma}_{11}\big) \\ &+ \eta^2 \boldsymbol{H}_{t,11} + \eta \Xi_{t+1,11}\end{aligned}$$

and

$$\begin{aligned}\boldsymbol{M}_{t+1,22} = \boldsymbol{M}_{t,22} &+ \eta(\boldsymbol{A}_{22}\boldsymbol{M}_{t,22} + \boldsymbol{M}_{t,22}\boldsymbol{A}_{22}) \\ &- \big(\boldsymbol{\Sigma}_{22}(\boldsymbol{M}_{t,22}^2 + \boldsymbol{M}_{t,12}^\top\boldsymbol{M}_{t,12}) - (\boldsymbol{M}_{t,22}^2 + \boldsymbol{M}_{t,12}^\top\boldsymbol{M}_{t,12})\boldsymbol{\Sigma}_{22}\big) \\ &+ \eta^2 \boldsymbol{H}_{t,22} + \eta \Xi_{t+1,22}\end{aligned}$$

We define

$$\boldsymbol{L}_{t+1} = \big[\boldsymbol{I}_{r_u} - \eta\boldsymbol{\Sigma}_{11}(\boldsymbol{T}_{11}^* - \boldsymbol{L}_t) + \eta^2\boldsymbol{\Sigma}_{11}^2(\boldsymbol{T}_{11}^* - \boldsymbol{L}_t)^2\big]^{-2}\boldsymbol{L}_t \tag{L-DYN}$$

with a diagonal initialization. We will use $\boldsymbol{L}_t(\boldsymbol{L}_0)$ whenever we want to make the initial condition explicit.

We define the following events:

- $\mathcal{E}_{t,1} \equiv \{\|\boldsymbol{L}_t^{-1/2}\boldsymbol{M}_{t,11}\boldsymbol{L}_t^{-1/2}\|_2 \le 2 \text{ and } \operatorname{tr}(\boldsymbol{M}_{t,22}) \le b_{22} \text{ and } \|\boldsymbol{M}_{t,22}\|_2 \le m_{22}\}$, where
$$b_{22} = m_{22} \coloneqq \frac{16\operatorname{tr}(\boldsymbol{\Sigma}_{22})}{\rho_d}.$$
- $\mathcal{E}_{t,2} \equiv \{\operatorname{tr}(\boldsymbol{M}_{t,11}) \le b_{11} \text{ and } \|\boldsymbol{L}_t^{-1/2}\boldsymbol{M}_{t,12}\boldsymbol{M}_{t,12}^\top\boldsymbol{L}_t^{-1/2}\|_2 \le 2m_{22} \text{ and } \|\Delta_t\|_F^2 \le 5\}$ where
$$b_{11} = 16\operatorname{tr}(\boldsymbol{T}_{11}^*).$$

For $N \coloneqq \lfloor \frac{\log^2 d}{\eta} \rfloor$, we define the hitting time

$$\mathcal{T}_{disc} \coloneqq \inf\{t \ge 0 : \mathcal{E}_{t,1} \cap \mathcal{E}_{t,2} \text{ fails}\} \wedge N.$$

D.1 Bounding lower block.

We have the following proposition.

Proposition 6 (Lower block)**.** *We define*

$$Z_{t,22} \coloneqq \eta \sum_{s=1}^{t} (\boldsymbol{I}_{d-r_u} + \eta\boldsymbol{A}_{22})^{-s}\Xi_{s,22}(\boldsymbol{I}_{d-r_u} + \eta\boldsymbol{A}_{22})^{-s}. \tag{D.7}$$

and for deterministic $\zeta_{bot}, \zeta_{tr} > 0$*, we define*

$$\mathcal{G}_{bot} \coloneqq \Big\{ \max_{0\le t\le N\wedge\mathcal{T}_{disc}} \|Z_{t,22}\|_2 \le \zeta_{bot}, \quad \max_{0\le t\le N\wedge\mathcal{T}_{disc}} |\operatorname{tr}(Z_{t,22})| \le \zeta_{tr} \Big\}$$

Under $\mathcal{G}_{bot}$*, for* $n = t \wedge \mathcal{T}_{disc}$*, we have*

- $\|\boldsymbol{M}_{n,22}\|_2 \le C_\varepsilon e^{2\|\boldsymbol{A}_{22}\|_2\eta n}\big((1+\log\frac{\lambda_{max}(\boldsymbol{\Sigma}_{22})}{\lambda_{min}(\boldsymbol{\Sigma}_{22})})(\|\boldsymbol{M}_{0,22}\|_2+\zeta_{bot})+\eta^2 n(b_{11}+b_{22})\|\boldsymbol{\Sigma}_{22}\|_2^2\big)$
- $\operatorname{tr}(\boldsymbol{M}_{n,22}) \le e^{2\|\boldsymbol{A}_{22}\|_2\eta n}\big(\operatorname{tr}(\boldsymbol{M}_{0,22})+\zeta_{tr}+C_\varepsilon\eta^2 n(b_{11}+b_{22})\operatorname{tr}(\boldsymbol{\Sigma}_{22}^2)\big)$.

Proof. For notational convenience, we introduce $\boldsymbol{D} \coloneqq (\boldsymbol{I}_{d-r_u}+\eta\boldsymbol{A}_{22})$ and $\boldsymbol{C}_t \coloneqq \boldsymbol{M}_{t,22}^2 + \boldsymbol{M}_{t,12}^\top\boldsymbol{M}_{t,12}$. We can write

$$\boldsymbol{M}_{t+1,22} = \boldsymbol{D}\boldsymbol{M}_{t,22}\boldsymbol{D} - \eta(\boldsymbol{\Sigma}_{22}\boldsymbol{C}_t + \boldsymbol{C}_t\boldsymbol{\Sigma}_{22}) - \eta^2\boldsymbol{A}_{22}\boldsymbol{M}_{t,22}\boldsymbol{A}_{22} + \eta^2\boldsymbol{H}_{t,22} + \eta\boldsymbol{\Xi}_{t+1,22}$$

We define $\boldsymbol{X}_t \coloneqq \boldsymbol{D}^{-t}\boldsymbol{M}_{t,22}\boldsymbol{D}^{-t}$. We have

$$\begin{aligned}\boldsymbol{X}_{t+1} &= \boldsymbol{X}_t - \eta(\boldsymbol{\Sigma}_{22}\boldsymbol{D}^{-(t+1)}\boldsymbol{C}_t\boldsymbol{D}^{-(t+1)} + \boldsymbol{D}^{-(t+1)}\boldsymbol{C}_t\boldsymbol{D}^{-(t+1)}\boldsymbol{\Sigma}_{22}) \\ &\quad - \eta^2\boldsymbol{D}^{-1}\boldsymbol{A}_{22}X_t\boldsymbol{A}_{22}\boldsymbol{D}^{-1} + \eta^2\boldsymbol{D}^{-(t+1)}\boldsymbol{H}_{t,22}\boldsymbol{D}^{-(t+1)} + \eta\boldsymbol{D}^{-(t+1)}\boldsymbol{\Xi}_{t+1,22}\boldsymbol{D}^{-(t+1)} \\ &= \boldsymbol{X}_0 - \eta\sum_{s=0}^{t}(\boldsymbol{\Sigma}_{22}\boldsymbol{D}^{-(s+1)}\boldsymbol{C}_s\boldsymbol{D}^{-(s+1)} + \boldsymbol{D}^{-(s+1)}\boldsymbol{C}_s\boldsymbol{D}^{-(s+1)}\boldsymbol{\Sigma}_{22}) \\ &\quad - \eta^2\sum_{s=0}^{t}\boldsymbol{D}^{-1}\boldsymbol{A}_{22}X_s\boldsymbol{A}_{22}\boldsymbol{D}^{-1} + \eta^2\sum_{s=0}^{t}\boldsymbol{D}^{-(s+1)}\boldsymbol{H}_{s,22}\boldsymbol{D}^{-(s+1)} + \boldsymbol{Z}_{t+1,22} \qquad \text{(D.8)}\end{aligned}$$

We use $\Phi_{\boldsymbol{\Sigma}_{22}}$ defined in the proof of Proposition 3. We have

$$\begin{aligned}0 \preceq \Phi_{\boldsymbol{\Sigma}_{22}}(\boldsymbol{X}_n) &\preceq \Phi_{\boldsymbol{\Sigma}_{22}}(\boldsymbol{X}_0) + \eta^2\sum_{s=0}^{n-1}\Phi_{\boldsymbol{\Sigma}_{22}}(\boldsymbol{D}^{-(s+1)}\boldsymbol{H}_{s,22}\boldsymbol{D}^{-(s+1)}) + \Phi_{\boldsymbol{\Sigma}_{22}}(\boldsymbol{Z}_{n,22}) \\ &\preceq \Phi_{\boldsymbol{\Sigma}_{22}}(\boldsymbol{X}_0) + C_\varepsilon\eta^2 n(b_{11}+b_{22})\boldsymbol{\Sigma}_{22} + \Phi_{\boldsymbol{\Sigma}_{22}}(\boldsymbol{Z}_{n,22}).\end{aligned}$$

Trace bound. By using (D.8), we have

$$\operatorname{tr}(\boldsymbol{X}_n) \le \operatorname{tr}(\boldsymbol{X}_0) + C_\varepsilon\eta^2 n(b_{11}+b_{22})\operatorname{tr}(\boldsymbol{\Sigma}_{22}^2) + \operatorname{tr}(\boldsymbol{Z}_{n,22}).$$

Since $\operatorname{tr}(\boldsymbol{M}_{n,22}) \le e^{2\|\boldsymbol{A}_{22}\|_2\eta n}\operatorname{tr}(\boldsymbol{X}_n)$, we have the trace bound.

Operator norm bound. We will use in block matrix decomposition for $\boldsymbol{\Sigma}_{22}$ in Propoisition 3 where we define $\boldsymbol{Z}_{n,j} \coloneqq \boldsymbol{P}_j\boldsymbol{Z}_{n,22}\boldsymbol{P}_j$. We have

$$\boldsymbol{X}_{n,j} = \boldsymbol{\Sigma}_j\Phi_{\boldsymbol{\Sigma}_j}(\boldsymbol{X}_{n,j}) + \Phi_{\boldsymbol{\Sigma}_j}(\boldsymbol{X}_{n,j})\boldsymbol{\Sigma}_j$$

Therefore, we have

$$\|\boldsymbol{X}_{n,j}\|_2 \le 2\|\boldsymbol{X}_{0,j}+\boldsymbol{Z}_{n,j}\|_2 + C\eta^2 n(b_{11}+b_{22})\|\boldsymbol{\Sigma}_j\|_2^2.$$

Then, we get

$$\|\boldsymbol{X}_n\|_2 \le C\Big(1+\log\frac{\lambda_{max}(\boldsymbol{\Sigma}_{22})}{\lambda_{min}(\boldsymbol{\Sigma}_{22})}\Big)(\|\boldsymbol{M}_{0,22}\|_2+\zeta_{bot}) + C\eta^2 n(b_{11}+b_{22})\|\boldsymbol{\Sigma}_{22}\|_2^2.$$

Since $\|\boldsymbol{M}_{n,22}\|_2 \le e^{2\|\boldsymbol{A}_{22}\|_2\eta n}\|X_n\|_2$, we have the operator norm bound. □

D.2 Bounding upper block

We start with the following proposition.

Proposition 7. *We define*

$$\boldsymbol{R}_t \coloneqq \boldsymbol{L}_t^{-1/2}\boldsymbol{M}_{t,11}\boldsymbol{L}_t^{-1/2}, \qquad \hat{\boldsymbol{R}}_t \coloneqq \boldsymbol{R}_t - \boldsymbol{I}_{r_u}, \qquad \tilde{\boldsymbol{\Xi}}_{t+1} \coloneqq \boldsymbol{L}_t^{-1/2}\boldsymbol{\Xi}_{t+1,11}\boldsymbol{L}_t^{-1/2}$$

and

$$\tilde{\boldsymbol{H}}_t \coloneqq \boldsymbol{L}_t^{-1/2}\Big(\boldsymbol{H}_{t,11} - \boldsymbol{\Sigma}_{11}(\boldsymbol{T}_{11}^* - \boldsymbol{M}_{t,11})\boldsymbol{M}_{t,11}(\boldsymbol{T}_{11}^* - \boldsymbol{M}_{t,11})\boldsymbol{\Sigma}_{11}\Big)\boldsymbol{L}_t^{-1/2}.$$

We have

$$\hat{\boldsymbol{R}}_{t+1} = \hat{\boldsymbol{R}}_t - \eta(\boldsymbol{\Sigma}_{11}\hat{\boldsymbol{R}}_t\boldsymbol{L}_t + \boldsymbol{L}_t\hat{\boldsymbol{R}}_t\boldsymbol{\Sigma}_{11} - \eta\boldsymbol{\Sigma}_{11}\boldsymbol{L}_t\hat{\boldsymbol{R}}_t\boldsymbol{L}_t\boldsymbol{\Sigma}_{11}) + \eta\mathcal{D}_t + \eta\boldsymbol{P}_t\tilde{\boldsymbol{\Xi}}_{t+1}\boldsymbol{P}_t,$$

where

$$\begin{aligned}\mathcal{D}_t \coloneqq & -\Big[\boldsymbol{\Sigma}_{11}\hat{\boldsymbol{R}}_t\boldsymbol{L}_t\hat{\boldsymbol{R}}_t + \hat{\boldsymbol{R}}_t\boldsymbol{L}_t\hat{\boldsymbol{R}}_t\boldsymbol{\Sigma}_{11} + \boldsymbol{\Sigma}_{11}\boldsymbol{K}_t + \boldsymbol{K}_t\boldsymbol{\Sigma}_{11}\Big] \\ & + \eta\Big[\boldsymbol{\Sigma}_{11}^2(\boldsymbol{T}_{11}^* - \boldsymbol{L}_t)\hat{\boldsymbol{R}}_t\boldsymbol{L}_t\boldsymbol{R}_t + \boldsymbol{R}_t\boldsymbol{L}_t\hat{\boldsymbol{R}}_t(\boldsymbol{T}_{11}^* - \boldsymbol{L}_t)\boldsymbol{\Sigma}_{11}^2 \\ & \qquad + \boldsymbol{\Sigma}_{11}\hat{\boldsymbol{R}}_t\boldsymbol{L}_t\boldsymbol{R}_t\boldsymbol{L}_t\hat{\boldsymbol{R}}_t\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{11}\boldsymbol{L}_t\hat{\boldsymbol{R}}_t\boldsymbol{L}_t\boldsymbol{\Sigma}_{11} \\ & \qquad + \boldsymbol{\Sigma}_{11}(\boldsymbol{T}_{11}^* - \boldsymbol{L}_t)(\boldsymbol{\Sigma}_{11}\boldsymbol{K}_t + \boldsymbol{K}_t\boldsymbol{\Sigma}_{11}) + (\boldsymbol{\Sigma}_{11}\boldsymbol{K}_t + \boldsymbol{K}_t\boldsymbol{\Sigma}_{11})\boldsymbol{\Sigma}_{11}(\boldsymbol{T}_{11}^* - \boldsymbol{L}_t)\Big] \\ & + \eta\boldsymbol{P}_t\widetilde{\boldsymbol{H}}_t\boldsymbol{P}_t + \eta^2\boldsymbol{F}_t\end{aligned}$$

and

$$\boldsymbol{K}_t \coloneqq \boldsymbol{L}_t^{-1/2}\boldsymbol{M}_{t,12}\boldsymbol{M}_{t,12}^\top\boldsymbol{L}_t^{-1/2}, \qquad \boldsymbol{P}_t \coloneqq \boldsymbol{I}_{r_u} - \eta\boldsymbol{\Sigma}_{11}(\boldsymbol{T}_{11}^* - \boldsymbol{L}_t) + \eta^2\boldsymbol{\Sigma}_{11}^2(\boldsymbol{T}_{11}^* - \boldsymbol{L}_t)^2.$$

Moreover, under

$$0 \prec \boldsymbol{L}_t \preceq I, \qquad 0 \preceq \boldsymbol{T}_{11}^* \preceq I, \qquad 0 \preceq \boldsymbol{\Sigma}_{11} \preceq I, \qquad 0 < \eta \leq 0.1. \tag{D.9}$$

we have

$$\|\boldsymbol{F}_t\|_2 \leq 3(1 + \|\boldsymbol{R}_t\|_2)^3 + 7\|\boldsymbol{K}_t\|_2. \tag{D.10}$$

Proof. The proof is postponed to Section G.1. □

Bounding $\mathcal{D}_t$. As notational convenience, we define

$$e_t \coloneqq \|\boldsymbol{R}_t - \boldsymbol{I}_{r_u}\|_2 \text{ and } n = t \wedge \mathcal{T}_{disc}.$$

In the next proposition, we bound $\mathcal{D}_t$.

Proposition 8. *For $s < \mathcal{T}_{disc}$, we have*

$$\|\mathcal{D}_s\|_2 \leq C_\varepsilon(e_s^2 + m_{22} + \eta\operatorname{tr}(\boldsymbol{M}_s)\|\mathsf{A}_s\|_2 + \eta^2).$$

where $\mathsf{A}_s \coloneqq \boldsymbol{P}_s\boldsymbol{\Sigma}_{11}\boldsymbol{L}_s^{-1}\boldsymbol{\Sigma}_{11}\boldsymbol{P}_s$

Proof. We start with observing that

$$\begin{aligned}-C(\|\Delta_n\|_F^2 + \mathbb{E}_t[\varepsilon_{t+1}^2])\operatorname{tr}(\boldsymbol{M}_n)\boldsymbol{\Sigma}_{11}\boldsymbol{L}_t^{-1}\boldsymbol{\Sigma}_{11} &\preceq \tilde{\boldsymbol{H}}_n \\ &\preceq C(\|\Delta_n\|_F^2 + \mathbb{E}_t[\varepsilon_{t+1}^2])\operatorname{tr}(\boldsymbol{M}_n)\boldsymbol{\Sigma}_{11}\boldsymbol{L}_t^{-1}\boldsymbol{\Sigma}_{11}.\end{aligned}$$

Therefore,

$$\|\boldsymbol{P}_n\tilde{\boldsymbol{H}}_n\boldsymbol{P}_n\|_2 \leq C_\varepsilon\operatorname{tr}(\boldsymbol{M}_n)\|\mathsf{A}_n\|_2.$$

By using triangle inequality, we can write

$$\begin{aligned}\|\mathcal{D}_n\|_2 &\leq C_\varepsilon\Big((e_n^2 + \|\boldsymbol{K}_n\|_2) + \eta(e_n + e_n^2 + \|\boldsymbol{K}_n\|_2 + \operatorname{tr}(\boldsymbol{M}_n)\|\mathsf{A}_n\|_2) + \eta^2(1 + e_n^3 + \|\boldsymbol{K}_n\|_2)\Big) \\ &\leq C_\varepsilon\Big(e_n^2 + m_{22} + \eta\operatorname{tr}(\boldsymbol{M}_n)\|\mathsf{A}_n\|_2 + \eta^2\Big).\end{aligned}$$

where we used the definition of $\mathcal{T}_{disc}$ in the last step. □

Main result. We define $\boldsymbol{Y}_{s+1} \coloneqq \boldsymbol{P}_s\widetilde{\boldsymbol{\Xi}}_{s+1}\boldsymbol{P}_s$, and

$$\mathcal{A}_t(\boldsymbol{R}) \coloneqq \boldsymbol{R} - \eta(\boldsymbol{\Sigma}_{11}\boldsymbol{R}\boldsymbol{L}_t + \boldsymbol{L}_t\boldsymbol{R}\boldsymbol{\Sigma}_{11} - \eta\boldsymbol{\Sigma}_{11}\boldsymbol{L}_t\boldsymbol{R}\boldsymbol{L}_t\boldsymbol{\Sigma}_{11}), \qquad \mathcal{U}_{t,s} \coloneqq \mathcal{A}_{t-1} \circ \cdots \circ \mathcal{A}_s$$

where $t > s$ and $\mathcal{U}_{s,s} = \mathrm{Id}$. We have the following:

Theorem 6. *We consider $\alpha+\beta>0.5$, and $\boldsymbol{L}_0=\frac{\boldsymbol{\Sigma}_{11}}{\rho_d}$, and define*

$$\boldsymbol{Z}_{t,11}\coloneqq \eta\sum_{s=0}^{t-1}\mathcal{U}_{t,s+1}(\boldsymbol{Y}_{s+1}),\qquad \mathcal{G}_{top}\coloneqq\big\{\max_{0\le t\le N\wedge\mathcal{T}_{disc}}\|\boldsymbol{Z}_{t,11}\|_2\le K_U^2\zeta_{top}\big\},\tag{D.11}$$

$$e_*\coloneqq e_0+\zeta_{top}+C_\varepsilon(\eta N m_{22}+\eta^2 N\rho_d(b_{11}+b_{22})+\eta^3 N)$$

where $\zeta_{top}>0$ is a deterministic number and K_U is defined in Proposition 12. Under $\mathcal{G}_{top}$ if $\boldsymbol{L}_0\preceq\boldsymbol{T}_{11}^*$ and $C_\varepsilon K_U^4\eta N e_*<0.5$, we have

$$\max_{1\le t\le N}e_n\le 2K_U^2 e_*.$$

Proof. By using Proposition 13, we can write

$$\hat{\boldsymbol{R}}_t=\mathcal{U}_{t,0}(\hat{\boldsymbol{R}}_0)+\eta\sum_{s=0}^{t-1}\mathcal{U}_{t,s+1}(\mathcal{D}_s)+Z_{t,11}$$

Then, by Proposition 12, we have

$$e_n\le K_U^2\Big(e_0+\zeta_{top}+\eta\sum_{s=0}^{n-1}\|\mathcal{D}_s\|_2\Big)$$

Since by Proposition 15, we have $\boldsymbol{L}_0\preceq\boldsymbol{L}_t\preceq\boldsymbol{T}_{11}^*$, we have $\|\mathsf{A}_t\|_2\le\rho_d$. Then by Proposition 8,

$$\begin{aligned}
e_n&\le K_U^2\Big((e_0+\zeta_{top})+C_\varepsilon\eta\sum_{s=0}^{n-1}(e_s^2+m_{22}+\eta\rho_d\operatorname{tr}(\boldsymbol{M}_s)+\eta^2)\Big)\\
&\le K_U^2\Big(\big[e_0+\zeta_{top}+C_\varepsilon(\eta N m_{22}+\eta^2 N\rho_d(b_{11}+b_{22})+\eta^3 N)\big]+C_\varepsilon\eta\sum_{s=0}^{n-1}e_s^2\Big)\\
&\le K_U^2\Big(e_*+C_\varepsilon\eta\sum_{s=0}^{n-1}e_s^2\Big)
\end{aligned}$$

By using Proposition 14, we have

$$e_n\le\frac{K_U^2 e_*}{1-C_\varepsilon K_U^4\eta n e_*}\le\frac{K_U^2 e_*}{1-C_\varepsilon K_U^4\eta N e_*}\le 2K_U^2 e_*.$$

□

Corollary 3. *We define $J_d\coloneqq 1+\operatorname{tr}(\boldsymbol{\Sigma}_{11})+\operatorname{tr}(\boldsymbol{T}_{11}^*)$ and let*

$$c_d\le\frac{1}{\log^3 d},\qquad r_u=\log^4 d,\qquad r_s\gg C_K\frac{r_u}{c_d^2},\qquad \eta=\frac{1}{J_d\log^9 d},$$

and

$$d\ge\rho_d\gg\max\{\operatorname{tr}(\boldsymbol{\Sigma})\log^5 d,\lambda_{r_u}^{-1}\}.$$

We have

$$\mathbb{P}\Big(\mathcal{T}_{disc}=N,\qquad\max_{0\le t\le N}e_t\le\frac{C_{\varepsilon,\alpha,\beta,K}(1+\log\log d)}{\log^3 d}\Big)\ge 1-Cd^{-K}.$$

Proof. We choose

$$\zeta_{bot}=\zeta_{tr}=\frac{\operatorname{tr}(\boldsymbol{\Sigma}_{22})}{\rho_d},\qquad\zeta_{top}=\frac{C_{\varepsilon,K}}{\log^3 d}$$

We first show that under this choice $\mathcal{G}_{init} \cap \mathcal{G}_{top} \cap \mathcal{G}_{bot}$ implies the event in the theorem. Fist, we observe that $\mathcal{E}_{t,2} \supseteq \mathcal{E}_{t,1}$ in the definition of $\mathcal{T}_{disc}$. Therefore, it is enough to show that $\mathcal{E}_{t,1}$ holds. For the remainig, we observe that

$$\|\boldsymbol{A}_{22}\|_2 \ll \frac{1}{\log^2 d}, \qquad \underbrace{b_{11}+b_{22}}_{:=b} = 16\Big(\operatorname{tr}(\boldsymbol{T}_{11}^*) + \frac{\operatorname{tr}(\boldsymbol{\Sigma}_{22})}{\rho_d}\Big) \le O(\log^2 d), \qquad e_* \le \frac{C_{\varepsilon,K}}{\log^3 d}.$$

By $\mathcal{G}_{init}$, we have

$$e_0 \le c_d, \quad 0 \preceq \boldsymbol{M}_{0,11} \preceq (1+c_d)\boldsymbol{T}_{11}^*, \quad \|\boldsymbol{M}_{0,22}\|_2 \le \operatorname{tr}(\boldsymbol{M}_{0,22}) \le \frac{1.01\operatorname{tr}(\boldsymbol{\Sigma}_{22})}{\rho_d}.$$

Therefore, we have $\mathcal{T}_{disc} > 0$. In the following, we will assume $\mathcal{G}_{init}$ holds.

By $\mathcal{G}_{bot}$, we observe that

$$\begin{aligned}\operatorname{tr}(\boldsymbol{M}_{n,22}) &\le 1.1\Big(\frac{2.01\operatorname{tr}(\boldsymbol{\Sigma}_{22})}{\rho_d} + C_\varepsilon \eta \log^2 d (b_{11}+b_{22})\operatorname{tr}(\boldsymbol{\Sigma}_{22}^2)\Big) \\ &\le 1.1\big(\frac{2.01\operatorname{tr}(\boldsymbol{\Sigma}_{22})}{\rho_d} + \frac{C_\varepsilon}{\log^9 d}\frac{\operatorname{tr}(\boldsymbol{\Sigma}_{22})}{\rho_d}\big) < 16\frac{\operatorname{tr}(\boldsymbol{\Sigma}_{22})}{\rho_d}.\end{aligned}$$

By $\mathcal{G}_{top}$, we observe that $C_\varepsilon K_U^4 \eta N e_* < \frac{C_{\varepsilon,\beta}(1+\log\log d)}{\log d} < 0.5$ and thus

$$e_n \le \frac{C_{\varepsilon,\beta}(1+\log\log d)^2}{\log^3 d} \tag{D.12}$$

Since $\|\boldsymbol{R}_t\|_2 \le 1+e_t$, by induction, we can show that $\mathcal{T}_{disc} = N$ and (D.12) holds for all $0 \le t \le N$.

Next, we will prove a concentration bound. Specifically, we will bound

$$\mathbb{P}[\mathcal{G}_{top}^c \cup \mathcal{G}_{bot}^c \cup \mathcal{G}_{init}^c] \le \mathbb{P}[\mathcal{G}_{top}^c] + \mathbb{P}[\mathcal{G}_{bot}^c] + d^{-K}.$$

For the top part, we will rely on Propositions 16 and 17. We observe that

$$\|\mathsf{A}_t\|_2 \le \rho_d, \qquad \operatorname{tr}(\mathsf{A}_t) \le \rho_d \operatorname{tr}(\boldsymbol{\Sigma}_{11}), \qquad \|\mathsf{B}_t\|_2 \le 6, \qquad \operatorname{tr}(\mathsf{B}_t) \le 2b$$

By choosing

$$a_* = C\sqrt{\rho_d \operatorname{tr}(\boldsymbol{\Sigma}_{11})b}, \qquad b_* = C\rho_d \operatorname{tr}(\boldsymbol{\Sigma}_{11})b, \qquad v_* = C_\varepsilon \rho_d J_d,$$

we can show that $\mathcal{T}_{disc} \le \mathcal{T}(v_*, a_*, b_*)$ and for $\kappa_\delta = \log\frac{CNr_u}{\delta}$,

$$\sqrt{\eta^2 N v_* \kappa_\delta} + \eta a_* \kappa_\delta^3 + \eta^2 b_* \kappa_\delta^5 \le \frac{C_{\varepsilon,K}}{\log^3 d},$$

Hence, with the given ζ_{top}, we have $\mathbb{P}[\mathcal{G}_{top}^c] < d^{-K}$.

For the bottom part, we will rely on Propositions 18 and 19. We observe that

$$\|\mathsf{C}_t\|_2 \le \|\boldsymbol{\Sigma}_{22}\|_2^2, \qquad \operatorname{tr}(\mathsf{C}_t) \le \operatorname{tr}(\boldsymbol{\Sigma}_{22}^2), \qquad \|\mathsf{F}_t\|_2 \le \frac{C\operatorname{tr}(\boldsymbol{\Sigma}_{22})}{\rho_d}, \qquad \operatorname{tr}(\mathsf{F}_t) \le \frac{C\operatorname{tr}(\boldsymbol{\Sigma}_{22})}{\rho_d}$$

By choosing

$$c_* = C\sqrt{\frac{\operatorname{tr}(\boldsymbol{\Sigma}_{22}^2)\operatorname{tr}(\boldsymbol{\Sigma}_{22})}{\rho_d}}, \qquad f_* = C\operatorname{tr}(\boldsymbol{\Sigma}_{22})b, \qquad v_* = C_\varepsilon \operatorname{tr}(\boldsymbol{\Sigma}_{22}^2)\frac{\operatorname{tr}(\boldsymbol{\Sigma}_{22})}{\rho_d}$$

we can show that $\mathcal{T}_{disc} \le \mathcal{T}(v_*, c_*, f_*)$ and for $\kappa_\delta = \log\frac{CNd}{\delta}$,

$$\sqrt{\eta^2 N v_* \kappa_\delta} + \eta c_* \kappa_\delta^3 + \eta^2 f_* \kappa_\delta^5 \le \frac{C_{\varepsilon,K}}{\log d}\frac{\operatorname{tr}(\boldsymbol{\Sigma}_{22})}{\rho_d}.$$

Hence, with the given ζ_{bot} and ζ_{tr}, we have $\mathbb{P}[\mathcal{G}_{bot}^c] < d^{-K}$, which gives the result. □

E Lazy Regime Analysis

Proposition 9. *As the network width $r_s \to \infty$, the induced Neural Tangent Kernel converges to*

$$\Theta(\boldsymbol{x}, \boldsymbol{x}') = 4\langle \boldsymbol{x}, \boldsymbol{x}' \rangle^2 + 1.$$

Proof. Write $\boldsymbol{\theta} = (\boldsymbol{W}, b)$ with $\boldsymbol{W} = [\boldsymbol{w}_1, \dots, \boldsymbol{w}_{r_s}]$, and evaluate all gradients at initialization, $\boldsymbol{w} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \boldsymbol{I}_d)$, $b = 0$. Since $\hat{y}(\boldsymbol{x}; \boldsymbol{W}, b) = \frac{1}{\sqrt{r_s}} \sum_{i=1}^{r_s} \langle \boldsymbol{w}, \boldsymbol{x} \rangle^2 - b$,

$$\nabla_{\boldsymbol{w}} \hat{y}(\boldsymbol{x}) = \frac{2}{\sqrt{r_s}} \langle \boldsymbol{w}, \boldsymbol{x} \rangle \, \boldsymbol{x}, \qquad \partial_b \hat{y}(\boldsymbol{x}) = -1.$$

Hence, for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d$, and $\widehat{\boldsymbol{G}}_{r_s} \coloneqq \frac{1}{r_s} \sum_{i=1}^{r_s} \boldsymbol{w}\boldsymbol{w}^\top$.

$$\begin{aligned} \hat{\Theta}_{r_s}(\boldsymbol{x}, \boldsymbol{x}') &= \sum_{i=1}^{r_s} \left\langle \nabla_{\boldsymbol{w}_i} \hat{y}(\boldsymbol{x}), \nabla_{\boldsymbol{w}_i} \hat{y}(\boldsymbol{x}') \right\rangle + \partial_b \hat{y}(\boldsymbol{x}) \, \partial_b \hat{y}(\boldsymbol{x}') \\ &= 4\langle \boldsymbol{x}, \boldsymbol{x}' \rangle \, \boldsymbol{x}^\top \widehat{\boldsymbol{G}}_{r_s} \boldsymbol{x}' + 1, \end{aligned}$$

Since $\hat{\boldsymbol{G}}_{r_s} \to \boldsymbol{I}_d$, we have the result. □

We can show that using (4.1) is asymptotically equivalent to learning with centered feature map $\boldsymbol{x} \to 2(\boldsymbol{x}\boldsymbol{x}^\top - \boldsymbol{\Sigma})$. Therefore, w.l.o.g, we will assume $\Phi(\boldsymbol{x}) = 2(\boldsymbol{x}\boldsymbol{x}^\top - \boldsymbol{\Sigma})$ in the following.

We observe that the covariance of $\Phi(\boldsymbol{x})$ can be diagonalized with orthonormal basis

$$\boldsymbol{E}_{ij} = \begin{cases} \boldsymbol{e}_i \boldsymbol{e}_i^\top, & i = j \\ \frac{1}{\sqrt{2}} (\boldsymbol{e}_i \boldsymbol{e}_j + \boldsymbol{e}_j \boldsymbol{e}_i^\top), & i \neq j \end{cases}$$

We define $\mathcal{I} \coloneqq \{(i, j) : i \leq j\}$ and

$$\psi_{ij}(\boldsymbol{x}) = \operatorname{tr}(\boldsymbol{E}\Phi(\boldsymbol{x})) = \begin{cases} 2(x_i^2 - \sigma_i), & i = j \\ 2\sqrt{2} x_i x_j, & i < j. \end{cases}$$

Lemma 1. *We have the following:*

- $\mathbb{E}[\psi_{ij}(\boldsymbol{x}) \psi_{kl}(\boldsymbol{x})] = \mu_{ij} \delta_{ik} \delta_{kl}$ *with* $\mu_{ij} \coloneqq 8\sigma_i \sigma_j$ *for* $(i, j), (k, l) \in \mathcal{I}$.
- *The operator* $\boldsymbol{\Gamma} \coloneqq \mathbb{E}[\Phi(\boldsymbol{x}) \otimes \Phi(\boldsymbol{x})]$ *satisfies* $\operatorname{tr}(\boldsymbol{\Gamma}) = 4(\operatorname{tr} \boldsymbol{\Sigma})^2 + 4\|\boldsymbol{\Sigma}\|_F^2$, $\operatorname{tr}(\boldsymbol{\Gamma}^2) = 64 \sum_{i \leq j} \sigma_i^2 \sigma_j^2 \leq 64 \|\boldsymbol{\Sigma}\|_F^4$.

Proof. By Isserlis' theorem, $\mathbb{E}[(x_i x_j - \Sigma_{ij})(x_k x_l - \Sigma_{kl})] = \Sigma_{ik}\Sigma_{jl} + \Sigma_{il}\Sigma_{jk} = \sigma_i \sigma_j (\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk})$. For $i < j$ and $k < l$ the second product vanishes, giving $\mathbb{E}[\psi_{ij}\psi_{kl}] = 8\sigma_i\sigma_j\delta_{ik}\delta_{jl}$. For $i = j$, $k = l$ both products equal $\sigma_i^2 \delta_{ik}$, giving $4 \cdot 2\sigma_i^2$. For $i = j$ and $k < l$ both vanish since $k \neq l$. The traces follow from $\sum_{i \leq j} \sigma_i \sigma_j = \frac{1}{2}[(\operatorname{tr} \boldsymbol{\Sigma})^2 + \|\boldsymbol{\Sigma}\|_F^2]$ and $\sum_{i \leq j} \sigma_i^2 \sigma_j^2 \leq (\sum_i \sigma_i^2)^2$. □

E.1 Deterministic equivalent

For $\kappa > 0$ define the degrees of freedom and the bias functional

$$\nu_1(\kappa) \coloneqq \sum_{(i,j) \in \mathcal{I}} \frac{\mu_{ij}}{\mu_{ij} + \kappa}, \quad \nu_2(\kappa) \coloneqq \sum_{(i,j) \in \mathcal{I}} \frac{\mu_{ij}^2}{(\mu_{ij} + \kappa)^2}, \quad \mathcal{B}(\kappa) \coloneqq \sum_{i \leq r} 2\sigma_i^2 \lambda_i^2 \frac{\kappa^2}{(8\sigma_i^2 + \kappa)^2}.$$

Lemma 2. *Let* $h(\kappa) =: 1 - \frac{1 + \nu_1(\kappa)}{N}$. *If* $\lambda > 0$, *or if* $\lambda = 0$ *and* $p + 1 > N$, *the equation*

$$\lambda = \kappa \, h(\kappa) \tag{E.1}$$

has a unique solution $\kappa = \kappa_N(\lambda) \in (0, \infty)$. *Moreover* $\lambda \mapsto \kappa_N(\lambda)$ *is continuous and nondecreasing,* $\kappa_N(\lambda) \geq \lambda$, *the map* $\kappa \mapsto \kappa h(\kappa)$ *is an increasing bijection from* $(\kappa_N(0), \infty)$ *onto* $(0, \infty)$, *and* $1 + \nu_2(\kappa_N(\lambda)) < N$.

Proof. h is continuous and strictly increasing, with $h(0^+) = 1 - \frac{1+p}{N}$ and $h(\infty) = 1 - \frac{1}{N}$. (E.1) is equivalent to $h(\kappa) = \lambda/\kappa$. Its right side is nonincreasing in κ, so there is exactly one crossing, and $h(\kappa) > \lambda/\kappa$ iff $\kappa > \kappa_N(\lambda)$. On $\{h > 0\} = (\kappa_N(0), \infty)$, $\kappa h(\kappa)$ is a product of positive increasing functions, so it is an increasing bijection onto $(0, \infty)$. This gives continuity and monotonicity of its inverse. $\kappa_N(\lambda) \geq \lambda$ since $h \leq 1$. Lastly, $\nu_2 < \nu_1$ and $1 + \nu_1(\kappa_N(\lambda)) = N(1 - \lambda/\kappa_N(\lambda)) \leq N$. □

With $\kappa = \kappa_N(\lambda)$, we will consider

$$\mathcal{R}_N(\lambda) \coloneqq \frac{1}{2} \cdot \frac{\mathcal{B}(\kappa) + \frac{1+\nu_2(\kappa)}{N}}{1 - \frac{1+\nu_2(\kappa)}{N}}. \tag{E.2}$$

E.1.1 The regime $\beta < 1$

Proposition 10. *Let $\beta < 1$, $d \to \infty$, and $r_{\mathit{eff}}(\mathbf{\Sigma})^2/N \to \bar{\kappa} \in (0, \infty)$. Set $\tau \coloneqq 4\sigma_1^2\bar{\kappa}$. For every $\gamma > 0$,*

$$\sup_{\lambda \in [0,\gamma]} \left|\kappa_{d,N}(\lambda) - (\lambda + \tau)\right| \to 0, \qquad \sup_{\lambda \in [0,\gamma]} \left|\mathcal{R}_{d,N}(\lambda) - \tfrac{1}{2}\mathcal{B}_\infty(\lambda + \tau)\right| \to 0$$

$$\inf_{\lambda \geq 0} \mathcal{R}_{d,N}(\lambda) \to \tfrac{1}{2}\mathcal{B}_\infty(\tau).$$

Proof. Since $\beta < 1$, we have

$$\frac{\|\mathbf{\Sigma}\|_F^2}{(\operatorname{tr}\mathbf{\Sigma})^2} \to 0 \qquad \text{and} \qquad \frac{\|\mathbf{\Sigma}\|_F^4}{(\operatorname{tr}\mathbf{\Sigma})^2} \to 0.$$

By Lemma 1, $r_{\text{eff}} = \operatorname{tr}\mathbf{\Sigma}/\sigma_1$, and $N \sim r_{\text{eff}}^2/\bar{\kappa}$, we get

$$\tau_d \coloneqq \frac{\operatorname{tr}\mathbf{\Gamma}}{N} = \frac{4\sigma_1^2 r_{\text{eff}}^2}{N}\left(1 + \frac{\|\mathbf{\Sigma}\|_F^2}{(\operatorname{tr}\mathbf{\Sigma})^2}\right) \to \tau, \qquad \epsilon_d \coloneqq \frac{\operatorname{tr}(\mathbf{\Gamma}^2)}{N} \leq \frac{64\|\mathbf{\Sigma}\|_F^4}{N} \to 0.$$

Also $p \asymp d^2 \gg N \asymp d^{2-2\beta}$, so Lemma 2 applies at $\lambda = 0$.

By using $\frac{\mu\kappa}{\mu+\kappa} = \mu - \frac{\mu^2}{\mu+\kappa}$,

$$g_d(\kappa) \coloneqq \kappa h(\kappa) = \kappa - \tau_d - \frac{\kappa}{N} + E_d(\kappa), \qquad 0 \leq E_d(\kappa) \coloneqq \frac{1}{N}\sum_{\mathcal{I}} \frac{\mu_{ij}^2}{\mu_{ij} + \kappa} \leq \frac{\epsilon_d}{\kappa}.$$

Fix $\epsilon \in (0, \tau/4)$ and let d be large enough that $|\tau_d - \tau| < \epsilon/4$. For all $\lambda \in [0, \gamma]$, $g_d(\lambda + \tau_d + \epsilon) \geq \lambda + \epsilon - \frac{\gamma+2\tau}{N} > \lambda$. Similarly, $g_d(\lambda + \tau_d - \epsilon) \leq \lambda - \epsilon + \frac{4\epsilon_d}{\tau} < \lambda$, using $\lambda + \tau_d - \epsilon \geq \tau/4$. Since $g_d(\kappa) > \lambda$ iff $\kappa > \kappa_{d,N}(\lambda)$ (Lemma 2), we get $|\kappa_{d,N}(\lambda) - \lambda - \tau_d| \leq \epsilon$ uniformly on $[0, \gamma]$. This proves the first claim, and also $\kappa_{d,N}(\lambda) \geq \tau/4$ for d large.

Since $\nu_2(\kappa) \leq \operatorname{tr}(\Gamma^2)/\kappa^2$,

$$\frac{1 + \nu_2(\kappa_{d,N}(\lambda))}{N} \leq \frac{1}{N} + \frac{16\,\epsilon_d}{\tau^2} \to 0$$

uniformly in $\lambda \geq 0$. So the variance term and the denominator in (E.2) tend to 0 and 1. Next,

$$\sup_{\kappa > 0} |\mathcal{B}_d(\kappa) - \mathcal{B}_\infty(\kappa)| \to 0$$

since $\alpha + \beta > 0.5$. We observe that $\mathcal{B}_\infty$ is uniformly continuous. Combined with the earlier result, this gives $\mathcal{B}_d(\kappa_{d,N}(\lambda)) \to \mathcal{B}_\infty(\lambda + \tau)$ uniformly on $[0, \gamma]$.

Finally, for every $\lambda \geq 0$, using that the denominator in (E.2) lies in $(0, 1]$, that $\mathcal{B}_d$ is nondecreasing, and that $\kappa_{d,N}$ is nondecreasing (Lemma 2),

$$\tfrac{1}{2}\mathcal{B}_d(\kappa_{d,N}(0)) \leq \tfrac{1}{2}\mathcal{B}_d(\kappa_{d,N}(\lambda)) \leq \mathcal{R}_{d,N}(\lambda).$$

Hence $\frac{1}{2}\mathcal{B}_d(\kappa_{d,N}(0)) \leq \inf_{\lambda \geq 0} \mathcal{R}_{d,N}(\lambda) \leq \mathcal{R}_{d,N}(0)$, and both bounds converge to $\frac{1}{2}\mathcal{B}_\infty(\tau)$. □

Lemma 3. *Let $K > 0$. Uniformly in $\kappa \in (0, K]$,*

$$\mathcal{B}_\infty(\kappa) \asymp \begin{cases} \kappa^{\frac{2\alpha+2\beta-1}{2\beta}}, & \alpha < \beta + \frac{1}{2} \\ \kappa^2, & \alpha > \beta + \frac{1}{2}, \end{cases} \qquad \textit{and} \qquad \mathcal{B}_d(\kappa) \leq C\,\mathcal{B}_\infty(\kappa).$$

Proof. The first result follows elementary power sum inequalities. The comparison $\mathcal{B}_d \leq C\mathcal{B}_\infty$ holds termwise since $c_d^2 \leq C$. □

F Auxiliary Results

Proposition 11. *We have*

$$\|\boldsymbol{T}_{11}^* - \boldsymbol{M}_{t,11}\|_F^2 \le \|\boldsymbol{T}^* - \boldsymbol{M}_t\|_F^2 \le \|\boldsymbol{T}_{11}^* - \boldsymbol{M}_{t,11}\|_F^2 + 4\|\boldsymbol{M}_{t,11}\|_2 \operatorname{tr}(\boldsymbol{M}_{t,22}) + 2(\|\boldsymbol{T}_{22}^*\|_F^2 + 2\|\boldsymbol{T}_{12}^*\|_F^2 + \|\boldsymbol{M}_{t,22}\|_F^2)$$

Proof. We can write

$$\|\boldsymbol{T}^* - \boldsymbol{M}_t\|_F^2 = \|\boldsymbol{T}_{11}^* - \boldsymbol{M}_{t,11}\|_F^2 + 2\|\boldsymbol{T}_{12}^* - \boldsymbol{M}_{t,12}\|_F^2 + \|\boldsymbol{T}_{22}^* - \boldsymbol{M}_{t,22}\|_F^2$$

The lower bound follows. For the upper bound, we have

$$\|\boldsymbol{T}^* - \boldsymbol{M}_t\|_F^2 \le \big\|\boldsymbol{T}_{11}^* - \boldsymbol{M}_{t,11}\|_F^2 + 4(\|\boldsymbol{T}_{12}^*\|_F^2 + \|\boldsymbol{M}_{t,12}\|_F^2) + 2(\|\boldsymbol{T}_{22}^*\| + \|\boldsymbol{M}_{t,22}\|_F^2)$$

Since $\boldsymbol{M}_t = \frac{1}{r_s}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{W}_t\boldsymbol{W}_t^\top\boldsymbol{\Sigma}^{\frac{1}{2}}$, we have

$$\boldsymbol{M}_{t,12} = \frac{1}{r_s}\boldsymbol{\Sigma}_{11}^{\frac{1}{2}}\boldsymbol{W}_{t,1}\boldsymbol{W}_{t,2}^\top\boldsymbol{\Sigma}_{22}^{\frac{1}{2}}$$

Therefore,

$$\|\boldsymbol{M}_{t,12}\|_F^2 = \frac{1}{r_s}\operatorname{tr}\Big(\boldsymbol{\Sigma}_{22}^{\frac{1}{2}}\boldsymbol{W}_{t,2}\frac{1}{r_s}(\boldsymbol{W}_{t,1}^\top\boldsymbol{\Sigma}_{11}\boldsymbol{W}_{t,1})\boldsymbol{W}_{t,2}^\top\boldsymbol{\Sigma}_{22}^{\frac{1}{2}}\Big) \le \|\boldsymbol{M}_{t,11}\|_2\operatorname{tr}(\boldsymbol{M}_{t,22}),$$

which gives the upper bound. □

Proposition 12. *We define linear operators*

$$(U_{t,s}[\boldsymbol{X}])_{ij} = \boldsymbol{X}_{ij}\begin{cases}\exp\Big(-\sigma_i\int_s^t \ell_{u,j}\,du - \sigma_j\int_s^t \ell_{u,i}\,du\Big), & \textit{continuous}\\ \prod_{u=s}^{t-1}(1-\eta\sigma_i\ell_{u,j})(1-\eta\sigma_j\ell_{u,i}), & \textit{discrete}.\end{cases} \tag{F.1}$$

If $0 \le \ell_{s,i} \le 1$*, and* $0 < \eta\max_i\sigma_i < 0.01$*, for both cases we have*

$$\|U_{t,s}[\boldsymbol{X}]\|_2 \le K_U^2\|\boldsymbol{X}\|_2,$$

for $K_U \coloneqq 1 + \frac{1}{2}\sqrt{1+\log\left(\frac{\max_i\sigma_i}{\min_i\sigma_i}\right)}$.

Proof. W.l.o.g. we assume $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_d$. For both cases, we have

$$U_{t,s}(\boldsymbol{X}) = \boldsymbol{H}^\top \odot (\boldsymbol{H}\odot\boldsymbol{X}),$$

where

$$\boldsymbol{H}_{ij} \coloneqq \begin{cases}\exp(-\sigma_i\int_s^t \ell_{u,j}du), & \text{continuous}\\ \prod_{u=s}^{t-1}(1-\eta\sigma_i\ell_{u,j}), & \text{discrete.}\end{cases} \tag{F.2}$$

It is enough to show that

$$\|\boldsymbol{H}\odot\boldsymbol{X}\|_2 \le K_U\|\boldsymbol{X}\|_2.$$

We introduce

$$b_j \coloneqq \begin{cases}\int_s^t \ell_{u,j}, & \text{continuous}\\ \eta\sum_{u=s}^{t-1}\ell_{u,j}, & \text{discrete}\end{cases}, \qquad f_j(z) \coloneqq \begin{cases}e^{-zb_j}, & \text{continuous}\\ \prod_{u=s}^{t-1}(1-\eta z\ell_{u,j}), & \text{discrete.}\end{cases}$$

In both cases, $\boldsymbol{H}_{ij} = f_j(\sigma_i)$. We define

$$q \coloneqq \begin{cases}0, & \text{continuous,}\\ \eta, & \text{discrete.}\end{cases}$$

We will first show that for $\sigma_d \le z \le \sigma_1$,

$$0 \le f_j(z) \le 1, \qquad 0 \le -f_j'(z) \le \frac{b_j e^{-zb_j}}{1-qz}$$

The first inequality is straightforward. For the second inequality, the continuous case is straightforward. For the discrete case,

$$-f_j'(z) = f_j(z) \sum_{u=s}^{t-1} \frac{\eta \ell_{u,j}}{1-\eta z \ell_{u,j}} \overset{(a)}{\le} f_j(z) \frac{\eta \sum_{u=s}^{t-1} \ell_{u,j}}{1-\eta z} \le \frac{b_j e^{-zb_j}}{1-\eta z}.$$

where we use $0 \le \ell_{u,j} \le 1$ in (a). We define

$$\Pi_z \coloneqq \text{diag}(\mathbb{1}_{z \le \sigma_i}), \qquad Q_z \coloneqq \text{diag}(-f_j'(z)). \tag{F.3}$$

We have

$$\boldsymbol{H} \odot \boldsymbol{X} \coloneqq \boldsymbol{X}\text{diag}(f_j(\sigma_d)) - \int_{\sigma_d}^{\sigma_1} \boldsymbol{\Pi}_u \boldsymbol{X} \boldsymbol{Q}_u du. \tag{F.4}$$

We observe that $\|\boldsymbol{X}\text{diag}(f_j(\sigma_d))\|_2 \le \|\boldsymbol{X}\|_2$ and for arbitrary unit vectors $\boldsymbol{w}, \boldsymbol{y}$

$$\begin{aligned}
\boldsymbol{w}^\top \Big(\int_{\sigma_d}^{\sigma_1} \boldsymbol{\Pi}_z \boldsymbol{X} \boldsymbol{Q}_z dz \Big) \boldsymbol{y} &\le \|\boldsymbol{X}\|_2 \int_{\sigma_d}^{\sigma_1} \|\boldsymbol{\Pi}_z \boldsymbol{w}\|_2 \|\boldsymbol{Q}_z \boldsymbol{y}\|_2 dz \\
&\le \|\boldsymbol{X}\|_2 \Big(\int_{\sigma_d}^{\sigma_1} \frac{\|\boldsymbol{\Pi}_z \boldsymbol{w}\|_2^2}{z(1-qz)^2} dz \Big)^{\frac{1}{2}} \Big(\int_{\sigma_d}^{\sigma_1} z(1-qz)^2 \|\boldsymbol{Q}_z \boldsymbol{y}\|_2^2 dz \Big)^{\frac{1}{2}} \\
&\le \|\boldsymbol{X}\|_2 \Big(\int_{\sigma_d}^{\sigma_1} \frac{1}{z(1-qz)^2} dz \Big)^{\frac{1}{2}} \Big(\sum_{j=1}^{d} y_j^2 b_j^2 \int_0^{\infty} z e^{-2zb_j} dz \Big)^{\frac{1}{2}} \\
&\le 0.5 \|\boldsymbol{X}\|_2 \Big(\int_{\sigma_d}^{\sigma_1} \frac{1}{z(1-qz)^2} dz \Big)^{\frac{1}{2}}
\end{aligned} \tag{F.5}$$

In the following, we will bound (F.5). For continuous case, we can show

$$\int_{\sigma_d}^{\sigma_1} \frac{1}{z} dz \le \log \frac{\sigma_1}{\sigma_d}$$

For the discrete case,

$$\begin{aligned}
\int_{\sigma_d}^{\sigma_1} \frac{1}{z(1-qz)^2} dz &= \int_{\sigma_d}^{\sigma_1} \Big(\frac{1}{z} + \eta \frac{2-\eta z}{(1-\eta z)^2} \Big) dz \\
&\overset{(b)}{\le} \int_{\sigma_d}^{\sigma_1} \Big(\frac{1}{z} + 3\eta \Big) dz \\
&\le \log \frac{\sigma_1}{\sigma_d} + 3\eta\sigma_1 \\
&\overset{(c)}{\le} \log \frac{\sigma_1}{\sigma_d} + 1
\end{aligned}$$

where we use $\eta\sigma_1 \le 0.01$ in (b) and (c).

Therefore, we have

$$\|\boldsymbol{H} \odot \boldsymbol{X}\|_2 \le (1 + 0.5\sqrt{1 + \log \frac{\sigma_1}{\sigma_d}}) \|\boldsymbol{X}\|_2$$

which gives the result. □

Proposition 13. *We have the following:*

- *Let $t \to \mathcal{A}_t$ be a linear operator (from $\mathbb{R}^{d \times d} \to \mathbb{R}^{d \times d}$) valued continuous flow.*

- *For $t \geq s$, let $\mathcal{U}_{t,s}$ be the solution of*

$$\begin{cases} \frac{d\boldsymbol{Z}_t}{dt} = \mathcal{A}_t(\boldsymbol{Z}_t), & \textit{continous} \\ \boldsymbol{Z}_{t+1} = \mathcal{A}_t(\boldsymbol{Z}_t), & \textit{discrete}. \end{cases}$$

- *Let $t \to \mathcal{D}_t$ be a $d \times d$-dimensional matrix valued continuous flow.*

Consider

$$\begin{cases} \frac{d\boldsymbol{E}_t}{dt} = \mathcal{A}_t(\boldsymbol{E}_t) + \mathcal{D}_t, & \textit{continous} \\ \boldsymbol{E}_{t+1} = \mathcal{A}_t(\boldsymbol{Z}_t) + \mathcal{D}_t, & \textit{discrete}. \end{cases}$$

We have

$$E_t = \begin{cases} \mathcal{U}_{t,0}(\boldsymbol{E}_0) + \int_0^t \mathcal{U}_{t,s}(\mathcal{D}_s)\, ds, & \textit{continous} \\ \mathcal{U}_{t,0}(\boldsymbol{E}_0) + \sum_{s=0}^{t-1} \mathcal{U}_{t,s+1}(\mathcal{D}_s), & \textit{discrete}. \end{cases}$$

Proof. We start with the continous case. We define

$$\boldsymbol{Y}_t = \mathcal{U}_{t,0}(\boldsymbol{E}_0) + \int_0^t \mathcal{U}_{t,s}(\mathcal{D}_s)\, ds.$$

We have $\boldsymbol{Y}_0 = \boldsymbol{E}_0$. Moreover,

$$\begin{aligned} \frac{d}{dt}\boldsymbol{Y}_t = \mathcal{U}_{t,0}(\mathcal{A}_t(\boldsymbol{E}_0)) + \mathcal{D}_t + \int_0^t \mathcal{U}_{t,s}(\mathcal{A}_t(\mathcal{D}_s))\, ds &= \mathcal{A}_t\Big(\mathcal{U}_{t,0}(\boldsymbol{E}_0) + \int_0^t \mathcal{U}_{t,s}(\mathcal{D}_s)\, ds\Big) + \mathcal{D}_t \\ &= \mathcal{A}_t(\boldsymbol{Y}_t) + \mathcal{D}_t. \end{aligned}$$

Since initial value problems have unique solutions, we have $Y_t = E_t$.

For the discrete case, we observe that

$$\mathcal{U}_{t,s} = \mathcal{A}_{t-1} \circ \cdots \circ \mathcal{A}_s, \Rightarrow \boldsymbol{Z}_t = \mathcal{U}_{t,s}(\boldsymbol{Z}_s).$$

We define

$$Y_t = \mathcal{U}_{t,0}(\boldsymbol{E}_0) + \sum_{s=0}^{t-1} \mathcal{U}_{t,s+1}(\mathcal{D}_s).$$

we have

$$\begin{aligned} Y_{t+1} &= \mathcal{U}_{t+1,0}(\boldsymbol{E}_0) + \sum_{s=0}^{t-1} \mathcal{U}_{t+1,s+1}(\mathcal{D}_s) + \mathcal{D}_t \\ &= \mathcal{A}_t\Big(\mathcal{U}_{t,0}(\boldsymbol{E}_0) + \sum_{s=0}^{t-1} \mathcal{U}_{t,s+1}(\mathcal{D}_s)\Big) + \mathcal{D}_t = \mathcal{A}_t(\boldsymbol{Y}_t) + \mathcal{D}_t. \end{aligned}$$

Hence, $\boldsymbol{Y}_t = \boldsymbol{E}_t$. □

Proposition 14. *For $e_0, \eta, C > 0$, we have the following:*

- *If $e_t \leq b + C\int_0^t e_s^2 ds$, then $e_t \leq \frac{b}{1-Cbt}$, for $t \in [0, \frac{1}{Cb})$.*
- *If $e_t \leq b + C\eta \sum_{s=0}^{t-1} e_s^2$, then $e_t \leq \frac{b}{1-Cb\eta t}$, for $t \in [0, \frac{1}{Cb\eta})$.*

Proposition 15. *We consider the iteration*

$$\ell_{t+1} = \big(1 - \eta(\lambda - \ell_t) + \eta^2(\lambda - \ell_t)^2\big)^{-2}\ell_t$$

For $0 < \ell_0 \leq \lambda$ and $\eta\lambda \leq 0.1$, we have

$$\frac{\lambda\ell_0}{\ell_0 + (\lambda - \ell_0)(1 - \lambda\eta + \lambda^2\eta^2)^{2t}} \leq \ell_t \leq \frac{\lambda\ell_0}{\ell_0 + (\lambda - \ell_0)(1 - 2\lambda\eta)^t}.$$

Consequently under Assumption 2, if $0 \prec \boldsymbol{L}_0 \preceq \boldsymbol{T}_{11}^*$, (L-DYN) satisfies

- *If* $\eta\lambda_1\sigma_1^2 \le 0.1$, *then* $0 \prec \boldsymbol{L}_t \preceq \boldsymbol{T}_{11}^*$
- *If* $\eta^2 N_d \in o_d(1)$, *we have*

$$\max_{t\le N_d}\max_{i\in[r_u]}\Big|\boldsymbol{L}_{t,ii}\Big(\frac{\lambda_i\sigma_i\boldsymbol{L}_{0,ii}}{\boldsymbol{L}_{0,ii}+(\lambda_i\sigma_i-\boldsymbol{L}_{0,ii})\exp(-2\lambda_i\sigma_i^2\eta t)}\Big)^{-1}-1\Big| \xrightarrow[d\to\infty]{} 0.$$

Proof. We define $a \coloneqq \lambda\eta$ and $g_t \coloneqq \eta(\lambda-\ell_t)$. We define $z_t \coloneqq \lambda/\ell_t - 1$, which implies $g_t = \frac{az_t}{1+z_t}$. We have

$$z_{t+1} = (1+z_t)\big(1-g_t+g_t^2\big)^2 - 1 = (1+z_t)(1-2g_t+3g_t^2-2g_t^3+g_t^4)-1$$

By using $(1+z_t)g_t = az_t$, we have

$$z_{t+1} = z_t(1-2a+a(3g_t-2g_t^2+g_t^3))$$

We first observe that if $g_t \ge 0$, then $(3g_t-2g_t^2+g_t^3)\ge 0$. Therefore, if $\ell_t \le \lambda$, then $\ell_{t+1}\le\lambda$. Hence, given $\ell_0\le\lambda$, we have $\ell_t\le\lambda$ for all $t>0$. Moreover, since $\eta\lambda<0.1$ and $\ell_0>0$, we have $\ell_t>0$. In the following we will use $0<\ell_t\le\lambda$ for all $t\ge 0$.

First, by using $g_t\ge 0$, we can write

$$z_{t+1}\ge z_t(1-2a) \Rightarrow z_t \ge z_0(1-2a)^t,$$

which gives the upper bound for ℓ_t. Second, since $g_t\le a$, we have

$$z_{t+1}\le z_t(1-a+a^2)^2 \Rightarrow z_t \le z_0(1-a+a^2)^{2t},$$

which gives the lower bound for ℓ_t.

First item for the results on $\boldsymbol{L}_t$ is direct application of the result above. For the second item, we define $a_i \coloneqq \eta\lambda_i\sigma_i^2$, $p_i \coloneqq 1-2a_i$, $q_i \coloneqq (1-a_i+a_i^2)^2$ and

$$\boldsymbol{F}_{t,i} = \frac{\lambda_i\sigma_i\boldsymbol{L}_{0,ii}}{\boldsymbol{L}_{0,ii}+(\lambda_i\sigma_i-\boldsymbol{L}_{0,ii})e^{-2a_it}}.$$

We have

$$\log q_i = 2\log(1-a_i+a_i^2)\le -2a_i+2a_i^2 \Rightarrow q_i\le e^{-2a_i}e^{2a_i^2}.$$

Moreover,

$$-\log p_i - 2a_i \le \frac{2a_i}{1-2a_i}-2a_i = \frac{4a_i^2}{1-2a_i}\le 5a_i^2. \Rightarrow p_i \ge e^{-2a_i}e^{-5a_i^2}.$$

Since $a_i^2N_d\in o_d(1)$, for sufficiently large d, we have

$$e^{-2a_i^2t}-1\le\frac{\boldsymbol{L}_{t,ii}}{\boldsymbol{F}_{t,i}}-1\le e^{5a_i^2t}-1 \Rightarrow \Big|\frac{\boldsymbol{L}_{t,ii}}{\boldsymbol{F}_{t,i}}-1\Big|\le 6a_i^2t.$$

Therefore,

$$\max_{t\le N_d}\max_{i\in[r_u]} 6a_i^2t \le O(\eta^2N_d)\xrightarrow[d\to\infty]{}0.$$

□

F.1 Concentration and related results

In this section, $\boldsymbol{z}\sim N(0,\boldsymbol{I}_d)$.

Lemma 4. *For symmetric matrices* $\Delta,\boldsymbol{M}\in\mathbb{R}^{d\times d}$ *with* $\boldsymbol{M}\succeq 0$ *and a general matrix* $\boldsymbol{B}\in\mathbb{R}^{m\times d}$, *we have*

- $\mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\top\boldsymbol{M}\boldsymbol{z}\boldsymbol{z}^\top]\preceq C\operatorname{tr}(\boldsymbol{M})\boldsymbol{I}_d$
- $\mathbb{E}[(\langle\Delta,\mathrm{He}_2(\boldsymbol{z})\rangle+\varepsilon)^2\boldsymbol{z}\boldsymbol{z}^\top\boldsymbol{M}\boldsymbol{z}\boldsymbol{z}^\top]\preceq C_\varepsilon(\|\Delta\|_F^2+\mathbb{E}[\varepsilon^2])\operatorname{tr}(\boldsymbol{M})\boldsymbol{I}_d$.

- $\mathbb{E}\Big[\Big((\langle\Delta, \mathrm{He}_2(\boldsymbol{z})\rangle + \varepsilon)^2 \boldsymbol{B}\boldsymbol{z}\boldsymbol{z}^\top \boldsymbol{M}\boldsymbol{z}\boldsymbol{z}^\top \boldsymbol{B}^\top\Big)^2\Big] \preceq C_\varepsilon(\|\Delta\|_F^2 + \mathbb{E}[\varepsilon^2])^2\|\boldsymbol{B}\|_F^2 \operatorname{tr}(\boldsymbol{M})^2 \boldsymbol{B}\boldsymbol{B}^\top$.

where C_ε is a constant depending on only the noise distribution.

Lemma 5. *For $p \geq 2$, the following holds:*

- $\mathbb{E}[|\boldsymbol{z}^\top \Delta \boldsymbol{z} - \operatorname{tr}(\Delta)|^p]^{1/p} \leq Cp\|\Delta\|_F$
- $\mathbb{E}[\|\Delta \boldsymbol{z}\|_2^p]^{1/p} \leq C\sqrt{p}\|\Delta\|_F$

Lemma 6. *Let $\boldsymbol{w}_i \sim N(0, \boldsymbol{I}_d)$ i.i.d. for $i \in [n]$, and let $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ be a positive definite matrix. We let*

$$\hat{\boldsymbol{\Sigma}}_n \coloneqq \boldsymbol{\Sigma}^{\frac{1}{2}}\Big(\frac{1}{n}\sum_{i=1}^n \boldsymbol{w}_i \boldsymbol{w}_i^\top\Big)\boldsymbol{\Sigma}^{\frac{1}{2}} \;\; \textit{and} \;\; r_{\mathit{eff}}(\boldsymbol{\Sigma}) = \frac{\operatorname{tr}(\boldsymbol{\Sigma})}{\|\boldsymbol{\Sigma}\|_2}.$$

We have

- $\mathbb{P}\Big[\operatorname{tr}(\boldsymbol{\Sigma}_n) - \operatorname{tr}(\boldsymbol{\Sigma}) \geq 2\sqrt{\frac{\|\boldsymbol{\Sigma}\|_F^2 t}{n}} + \frac{2\|\boldsymbol{\Sigma}\|_2 t}{n}\Big] \leq e^{-t}$.
- $\mathbb{P}\Big[\|\hat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}\|_2 \geq C\|\boldsymbol{\Sigma}\|_2(\sqrt{\frac{r_{\mathit{eff}}(\boldsymbol{\Sigma})+t}{n}} + \frac{r_{\mathit{eff}}(\boldsymbol{\Sigma})+t}{n})\Big] \leq e^{-t}$.

Proof. We can write $n(\operatorname{tr}(\hat{\boldsymbol{\Sigma}}_n) - \operatorname{tr}(\boldsymbol{\Sigma}))$ as a quadratic form with a $n \cdot d$ dimensional standard Gaussian vector with $\boldsymbol{\Sigma} \otimes \boldsymbol{I}_n$ matrix. By using [HKZ11], the first item follows.

For the second item, we define

$$\mathcal{K} = \{\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{u} : \|\boldsymbol{u}\|_2 \leq 1\}, \quad \boldsymbol{W} = \{\boldsymbol{w}_i\}_{i=1}^n \in \mathbb{R}^{n\times d}, \quad Z(\boldsymbol{W}) \coloneqq \sup_{u\in\mathcal{K}}\Big|\|\boldsymbol{W}\boldsymbol{u}\|_2 - \sqrt{n}\|\boldsymbol{u}\|_2\Big|.$$

We have

$$\|\hat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}\|_2 = \sup_{u\in\mathcal{K}}\Big|\frac{\|\boldsymbol{W}\boldsymbol{u}\|_2^2}{n} - \|\boldsymbol{u}\|_2^2\Big|$$

For any fixed $\boldsymbol{u} \in \mathcal{K}$, we have

$$\begin{aligned}\Big|\frac{\|\boldsymbol{W}\boldsymbol{u}\|_2^2}{n} - \|\boldsymbol{u}\|_2^2\Big| &= \Big|\frac{\|\boldsymbol{W}\boldsymbol{u}\|_2}{\sqrt{n}} - \|\boldsymbol{u}\|_2\Big|\Big|\frac{\|\boldsymbol{W}\boldsymbol{u}\|_2}{\sqrt{n}} + \|\boldsymbol{u}\|_2\Big| \\ &\leq \frac{Z(\boldsymbol{W})}{\sqrt{n}}\Big(\frac{Z(\boldsymbol{W})}{\sqrt{n}} + 2\|\boldsymbol{u}\|_2\Big) \leq \frac{2\|\boldsymbol{\Sigma}\|_2^{\frac{1}{2}} Z(\boldsymbol{W})}{\sqrt{n}} + \frac{Z(\boldsymbol{W})^2}{n}\end{aligned}$$

Therefore, we have

$$\|\hat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}\|_2 \leq \frac{2\|\boldsymbol{\Sigma}\|_2^{\frac{1}{2}} Z(\boldsymbol{W})}{\sqrt{n}} + \frac{Z(\boldsymbol{W})^2}{n}. \tag{F.6}$$

On the other hand, we observe that

$$\begin{aligned}Z(\boldsymbol{W}+\Delta) - Z(\boldsymbol{W}) &= \sup_{u\in\mathcal{K}}\Big|\|(\boldsymbol{W}+\Delta)\boldsymbol{u}\|_2 - \sqrt{n}\|\boldsymbol{u}\|_2\Big| - \sup_{u\in\mathcal{K}}\Big|\|\boldsymbol{W}\boldsymbol{u}\|_2 - \sqrt{n}\|\boldsymbol{u}\|_2\Big| \\ &\leq \sup_{u\in\mathcal{K}}|\|(\boldsymbol{W}+\Delta)\boldsymbol{u}\|_2 - \|\boldsymbol{W}\boldsymbol{u}\|_2| \leq \|\boldsymbol{\Sigma}\|_2^{\frac{1}{2}}\|\Delta\|_2\end{aligned}$$

Hence, by Lipshitz concentration bound, we have

$$\mathbb{P}[Z(\boldsymbol{W}) \geq \mathbb{E}[Z(\boldsymbol{W})] + \sqrt{2\|\boldsymbol{\Sigma}\|_2 t}] \leq e^{-t}. \tag{F.7}$$

Hence, it is enough to bound $\mathbb{E}[Z(\boldsymbol{W})]$ for the result. We define

$$Z_+(\boldsymbol{W}) \coloneqq \sup_{u\in\mathcal{K}}(\|\boldsymbol{W}\boldsymbol{u}\|_2 - \sqrt{n}\|\boldsymbol{u}\|_2) \;\text{ and }\; Z_-(\boldsymbol{W}) \coloneqq \sup_{u\in\mathcal{K}}(\sqrt{n}\|\boldsymbol{u}\|_2 - \|\boldsymbol{W}\boldsymbol{u}\|_2).$$

We have $Z(\boldsymbol{W}) = Z_+(\boldsymbol{W}) \vee Z_-(\boldsymbol{W})$. Since $0 \in \mathcal{K}$, we have $Z_+(\boldsymbol{W}), Z_-(\boldsymbol{W}) \geq 0$, and thus,

$$\mathbb{E}[Z(\boldsymbol{W})] \leq \mathbb{E}[Z_+(\boldsymbol{W})] + \mathbb{E}[Z_-(\boldsymbol{W})].$$

For $v \in \mathbb{R}^n$ with $\|v\|_2 \leq 1$, and $h \sim N(0, \boldsymbol{I}_d)$ and $\boldsymbol{g} \sim N(0, \boldsymbol{I}_n)$, we define

$$X_{u,v} \coloneqq v^\top \boldsymbol{W} \boldsymbol{u} - \sqrt{n}\|\boldsymbol{u}\|_2 \;\text{ and }\; Y_{u,v} \coloneqq \|\boldsymbol{u}\|_2 v^\top \boldsymbol{g} + \boldsymbol{u}^\top h - \sqrt{n}\|\boldsymbol{u}\|_2.$$

We observe that

$$\mathbb{E}[(Y_{u,v} - Y_{u',v'})^2] - \mathbb{E}[(X_{u,v} - X_{u',v'})^2] = 2(\|\boldsymbol{u}\|_2\|\boldsymbol{u}'\|_2 - \boldsymbol{u}^\top \boldsymbol{u}')(1 - v^\top v') + (\|\boldsymbol{u}\|_2 - \|\boldsymbol{u}'\|_2)^2$$

$$\geq 0.$$

We have

$$\begin{aligned}
\mathbb{E}[Z_+(\boldsymbol{W})] &= \mathbb{E}[\max_{u \in \mathcal{K}, \|v\| \leq 1} X_{u,v}] \\
&\overset{(a)}{\leq} \mathbb{E}[\max_{u \in \mathcal{K}, \|v\| \leq 1} Y_{u,v}] \\
&= \mathbb{E}\big[\sup_{u \in \mathcal{K}} \big(\|\boldsymbol{u}\|_2(\|\boldsymbol{g}\|_2 - \sqrt{n}) + \boldsymbol{u}^\top h\big)\big] \\
&\leq \mathbb{E}[\sup_{u \in \mathcal{K}} \boldsymbol{u}^\top h] + \|\boldsymbol{\Sigma}\|_2^{\frac{1}{2}} \mathbb{E}[(\|\boldsymbol{g}\|_2 - \sqrt{n})_+].
\end{aligned}$$

where we use Sudakov-Fernique inequality in (a). By using Gordon's comparison inequality, we have

$$\mathbb{E}\big[\inf_{u \in \mathcal{K}} \big(\|\boldsymbol{W}\boldsymbol{u}\|_2 - \sqrt{n}\|\boldsymbol{u}\|_2\big)\big] \geq \mathbb{E}\big[\inf_{u \in \mathcal{K}} \big((\|\boldsymbol{g}\|_2 - \sqrt{n})\|\boldsymbol{u}\|_2 + \boldsymbol{u}^\top h\big)\big].$$

Therefore,

$$\begin{aligned}
\mathbb{E}[Z_-(\boldsymbol{W})] = -\mathbb{E}\big[\inf_{u \in \mathcal{K}} \big(\|\boldsymbol{W}\boldsymbol{u}\|_2 - \sqrt{n}\|\boldsymbol{u}\|_2\big)\big] &\leq \mathbb{E}\big[\sup_{u \in \mathcal{K}} \big((\sqrt{n} - \|\boldsymbol{g}\|_2)\|\boldsymbol{u}\|_2 - \boldsymbol{u}^\top h\big)\big] \\
&\leq \mathbb{E}[\sup_{u \in \mathcal{K}} \boldsymbol{u}^\top h] + \|\boldsymbol{\Sigma}\|_2^{\frac{1}{2}} \mathbb{E}[(\sqrt{n} - \|\boldsymbol{g}\|_2)_+].
\end{aligned}$$

Therefore, we get

$$\mathbb{E}[Z_+(\boldsymbol{W})] + \mathbb{E}[Z_-(\boldsymbol{W})] \leq 2\mathbb{E}[\sup_{u \in \mathcal{K}} \boldsymbol{u}^\top h] + \|\boldsymbol{\Sigma}\|_2^{\frac{1}{2}} \mathbb{E}[|\|\boldsymbol{g}\|_2 - \sqrt{n}|]$$

We have

$$\mathbb{E}[\|\boldsymbol{\Sigma}^{\frac{1}{2}} h\|_2] \leq \sqrt{\operatorname{tr}(\boldsymbol{\Sigma})} = \|\boldsymbol{\Sigma}\|_2^{\frac{1}{2}} \sqrt{r_{\text{eff}}(\boldsymbol{\Sigma})}.$$

and

$$\mathbb{E}[|\|\boldsymbol{g}\|_2 - \sqrt{n}|] = \mathbb{E}[\frac{|\|\boldsymbol{g}\|_2^2 - n|}{|\|\boldsymbol{g}\|_2 + \sqrt{n}|}] \leq \frac{1}{\sqrt{n}} \mathbb{E}[(\|\boldsymbol{g}\|_2^2 - n)^2]^{1/2} \leq \sqrt{2}.$$

Therefore, we have

$$\begin{aligned}
\mathbb{E}[Z(\boldsymbol{W})] \leq \mathbb{E}[Z_+(\boldsymbol{W})] + \mathbb{E}[Z_-(\boldsymbol{W})] &\leq \|\boldsymbol{\Sigma}\|_2^{\frac{1}{2}} (2\sqrt{r_{\text{eff}}(\boldsymbol{\Sigma})} + \sqrt{2}) \\
&\leq (2 + \sqrt{2})\|\boldsymbol{\Sigma}\|_2^{\frac{1}{2}} \sqrt{r_{\text{eff}}(\boldsymbol{\Sigma})}.
\end{aligned}$$

By using (F.7) in (F.6), we have the second item. □

Events for initialization. For the events for initalization, let $\boldsymbol{Z} \in \mathbb{R}^{d \times r_s}$ be a random matrix whose entries are i.i.d. and distributed as $N(0, 1/\rho_d)$. We have $\boldsymbol{G}_0 =_d \frac{1}{r_s} \boldsymbol{\Sigma}^{\frac{1}{2}} Z Z^\top \boldsymbol{\Sigma}^{\frac{1}{2}}$. We define the $\boldsymbol{Z}_m \in \mathbb{R}^{m \times r_s}$ as the submatrix formed by the first m rows and $Z_{-m} \in \mathbb{R}^{m \times r_s}$ as the submatrix formed by the last m rows, i.e.,

$$\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{Z}_m \\ \vdots \end{bmatrix} \;\text{ or }\; \boldsymbol{Z} = \begin{bmatrix} \vdots \\ \boldsymbol{Z}_{-m} \end{bmatrix}$$

We start with characterizing "good events" for initial matrices given by the following lemma:

Lemma 7. *For $c_d, K > 0$, and a constant $C_K > 0$, we define the events*

E.1 $\frac{1-c_d}{\rho_d} \leq \lambda_{min}(\frac{1}{r_s}\boldsymbol{Z}_{r_u}\boldsymbol{Z}_{r_u}^\top) \leq \lambda_{max}(\frac{1}{r_s}\boldsymbol{Z}_{r_u}Z_{r_u}^\top) \leq \frac{1+c_d}{\rho_d}$

E.2 $\lambda_{max}(\frac{1}{r_s}\boldsymbol{\Sigma}_{22}^{\frac{1}{2}}\boldsymbol{Z}_{-(d-r_u)}\boldsymbol{Z}_{-(d-r_u)}^\top\boldsymbol{\Sigma}_{22}^{\frac{1}{2}}) \leq \frac{C_K}{\rho_d}\Big(\|\boldsymbol{\Sigma}_{22}\|_2 + \frac{\mathrm{tr}(\boldsymbol{\Sigma}_{22})+\|\boldsymbol{\Sigma}_{22}\|_2\log d}{r_s}\Big)$.

E.3 $\mathrm{tr}(\frac{1}{r_s}\boldsymbol{\Sigma}_{22}^{\frac{1}{2}}\boldsymbol{Z}_{-(d-r_u)}\boldsymbol{Z}_{-(d-r_u)}^\top\boldsymbol{\Sigma}_{22}^{\frac{1}{2}}) \leq \frac{\mathrm{tr}(\boldsymbol{\Sigma}_{22})}{\rho_d} + \frac{C_K}{\rho_d}\Big(\sqrt{\frac{\|\boldsymbol{\Sigma}_{22}\|_F^2\log d}{r_s}} + \frac{\|\boldsymbol{\Sigma}_{22}\|_2\log d}{r_s}\Big)$

and

$$\mathcal{G}_{init} \equiv (\text{E.1}) \cap (\text{E.2}) \cap (\text{E.3}).$$

For $r_u > \log d$ and $\frac{r_u}{c_d^2} \ll r_s$ and $d \geq \Omega_K(1)$, we have

$$\mathbb{P}[\mathcal{G}_{init}] > 1 - d^{-K}.$$

Proof. All items directly follow from Lemma 6. □

F.2 Noise bounds

F.2.1 Upper block

In this part, we will bound $\sum_{s=0}^{t-1}\mathcal{U}_{t,s+1}\big(\boldsymbol{P}_s\widetilde{\boldsymbol{\Xi}}_{s+1}\boldsymbol{P}_s\big)$ in (D.11) .

- For notational convenience, we use $\boldsymbol{Y}_{s+1} = \boldsymbol{P}_s\widetilde{\boldsymbol{\Xi}}_{s+1}\boldsymbol{P}_s$.
- We observe that $\boldsymbol{Y}_{s+1}$ is a $\mathcal{F}_{s+1}$-measurable symmetric random matrix and $\mathbb{E}_s[\boldsymbol{Y}_{s+1}] = 0$.
- By using the definition of (F.1), $\mathcal{U}_{t,s+1}\big(\boldsymbol{Y}_{s+1}\big)$ is also a $\mathcal{F}_{s+1}$-measurable symmetric random matrix and $\mathbb{E}_s[\mathcal{U}_{t,s+1}\big(\boldsymbol{Y}_{s+1}\big)] = 0$.

We start with the following lemma.

Lemma 8. *We have*

$$\|\mathbb{E}_s[U_{t,s+1}(\boldsymbol{Y}_{s+1})^2]\|_2 \leq K_U^4\|\mathbb{E}_s[\boldsymbol{Y}_{s+1}^2]\|_2$$

Proof. We will use the definitions in (F.2) and (F.3). We let $\Phi_{\boldsymbol{H}}(\boldsymbol{X}) \coloneqq \boldsymbol{H}\odot\boldsymbol{X}$. By using (F.4), we write

$$\Phi_{\boldsymbol{H}}(\boldsymbol{X}) \coloneqq \boldsymbol{X}\mathrm{diag}(f_j(\sigma_d)) - \int_{\sigma_d}^{\sigma_1}\boldsymbol{\Pi}_u\boldsymbol{X}\boldsymbol{Q}_u du.$$

Let $\boldsymbol{X}$ be an arbitrary random matrix (not necessarily symmetric), and we define

$$v_s \coloneqq \|\mathbb{E}_s[\boldsymbol{X}\boldsymbol{X}^\top]\|_2$$

For arbitrary unit vectors $\boldsymbol{w}$ and $\boldsymbol{y}$, by using Cauchy–Schwarz, we can write

$$\boldsymbol{y}^\top\Big(\int\boldsymbol{Q}_z\boldsymbol{X}^\top\boldsymbol{\Pi}_z\,dz\Big)\boldsymbol{w} \leq \int_{\sigma_d}^{\sigma_1}\|\boldsymbol{X}^\top\boldsymbol{\Pi}_z\boldsymbol{w}\|_2\|\boldsymbol{Q}_z\boldsymbol{y}\|_2 dz \leq 0.5\Big(\int_{\sigma_d}^{\sigma_1}\frac{\|\boldsymbol{X}^\top\boldsymbol{\Pi}_z\boldsymbol{w}\|_2^2}{z(1-qz)^2}dz\Big)^{\frac{1}{2}}.$$

Therefore,

$$\mathbb{E}_s\Big[\Big\|\int\boldsymbol{Q}_z\boldsymbol{X}^\top\boldsymbol{\Pi}_z\boldsymbol{w}\,dz\Big\|_2^2\Big] \leq 0.25\int_{\sigma_d}^{\sigma_1}\frac{\mathbb{E}_s[\|\boldsymbol{X}^\top\boldsymbol{\Pi}_z\boldsymbol{w}\|_2^2]}{z(1-qz)^2}dz \leq 0.25v_s(1+\log\frac{\sigma_1}{\sigma_d})$$

Therefore, we have

$$\mathbb{E}_s[\|\Phi_{\boldsymbol{H}}(\boldsymbol{X})^\top\boldsymbol{w}\|_2^2]^{\frac{1}{2}} \leq K_U\sqrt{v} \Rightarrow \|\mathbb{E}_s[\Phi_{\boldsymbol{H}}(\boldsymbol{X})\Phi_{\boldsymbol{H}}(\boldsymbol{X})^\top]\|_2 \leq K_U^2 v_s.$$

Since $U_{t,s+1} = \Phi_{\boldsymbol{H}^\top}(\Phi_{\boldsymbol{H}}(\boldsymbol{X}))$, we have the statement. □

Proposition 16. *We define* $\nu_s^2 \coloneqq (\|\boldsymbol{T}^* - \boldsymbol{M}_s\|_F^2 + \mathbb{E}_s[\varepsilon_{s+1}^2])$ *and*

$$\mathsf{A}_s \coloneqq \boldsymbol{P}_s\boldsymbol{\Sigma}_{11}\boldsymbol{L}_s^{-1}\boldsymbol{\Sigma}_{11}\boldsymbol{P}_s, \qquad \mathsf{B}_s \coloneqq \boldsymbol{P}_s(\boldsymbol{R}_s\boldsymbol{L}_s\boldsymbol{R}_s + \boldsymbol{K}_s)\boldsymbol{P}_s,$$
$$a_s \coloneqq \nu_s\sqrt{\operatorname{tr}\mathsf{A}_s \operatorname{tr}\mathsf{B}_s}, \qquad b_s \coloneqq \nu_s^2 \operatorname{tr}\mathsf{A}_s \operatorname{tr}\boldsymbol{M}_s.$$

For a constant $C_\varepsilon > 0$ *depending on only the noise distribution, we have*

$$\mathbb{E}_s[\boldsymbol{Y}_{s+1}^2] \preceq C_\varepsilon\nu_s^2\left[(\operatorname{tr}\mathsf{B}_s)\mathsf{A}_s + (\operatorname{tr}\mathsf{A}_s)\mathsf{B}_s\right] + C_\varepsilon\eta^2\nu_s^4(\operatorname{tr}\mathsf{A}_s)(\operatorname{tr}\boldsymbol{M}_s)^2\mathsf{A}_s$$
$$\mathbb{P}_s[\|\boldsymbol{Y}_{s+1}\|_2 > C_\varepsilon(a_s u^2 + \eta b_s u^4)] \le e^{-u}, \text{ for } u \ge 2. \tag{F.8}$$

Proof. We start with the variance bound. We define

$$\boldsymbol{\Delta}_s = \boldsymbol{T}^* - \boldsymbol{M}_s, \qquad \boldsymbol{J}_s \coloneqq \boldsymbol{P}_s\boldsymbol{L}_s^{\frac{-1}{2}}\left[\boldsymbol{I}_{r_u} \quad 0\right], \qquad \boldsymbol{C}_s \coloneqq \boldsymbol{J}_s\boldsymbol{\Sigma}, \qquad \boldsymbol{D}_s \coloneqq \boldsymbol{M}_s\boldsymbol{J}_s^\top.$$

We have

$$\mathsf{A}_s = \boldsymbol{C}_s\boldsymbol{C}_s^\top \text{ and } \mathsf{B}_s = \boldsymbol{D}_s^\top\boldsymbol{D}_s.$$

We define $\boldsymbol{Y}_{s+1} \coloneqq \boldsymbol{Q}_{s+1} - \mathbb{E}_s[\boldsymbol{Q}_{s+1}]$, where

$$\boldsymbol{Q}_{s+1} = \underbrace{(\langle\boldsymbol{\Delta}_s, \operatorname{He}_2(\boldsymbol{z}_{s+1})\rangle + \varepsilon_{s+1})(\boldsymbol{C}_s\tfrac{1}{\sqrt{2}}\boldsymbol{z}_{s+1}\boldsymbol{z}_{s+1}^\top\boldsymbol{D}_s + \boldsymbol{D}_s^\top\tfrac{1}{\sqrt{2}}\boldsymbol{z}_{s+1}\boldsymbol{z}_{s+1}^\top\boldsymbol{C}_s^\top)}_{\coloneqq\boldsymbol{Q}_{s+1,1}}$$
$$+\eta\underbrace{(\langle\boldsymbol{\Delta}_s, \operatorname{He}_2(\boldsymbol{z}_{s+1})\rangle + \varepsilon_{s+1})^2\boldsymbol{C}_s\tfrac{1}{\sqrt{2}}\boldsymbol{z}_{s+1}\boldsymbol{z}_{s+1}^\top\boldsymbol{M}_s\tfrac{1}{\sqrt{2}}\boldsymbol{z}_{s+1}\boldsymbol{z}_{s+1}^\top\boldsymbol{C}_s^\top}_{\coloneqq\boldsymbol{Q}_{s+1,2}}.$$

We have

$$\begin{aligned}
\mathbb{E}_s[\boldsymbol{Y}_{s+1}^2] &\preceq C\mathbb{E}_s[(\langle\boldsymbol{\Delta}_s, \operatorname{He}_2(\boldsymbol{z}_{s+1})\rangle + \varepsilon_{s+1})^2\boldsymbol{C}_s\tfrac{1}{\sqrt{2}}\boldsymbol{z}_{s+1}\boldsymbol{z}_{s+1}^\top\boldsymbol{D}_s\boldsymbol{D}_s^\top\tfrac{1}{\sqrt{2}}\boldsymbol{z}_{s+1}\boldsymbol{z}_{s+1}^\top\boldsymbol{C}_s^\top]\\
&+ C\mathbb{E}_s[(\langle\boldsymbol{\Delta}_s, \operatorname{He}_2(\boldsymbol{z}_{s+1})\rangle + \varepsilon_{s+1})^2\boldsymbol{D}_s^\top\tfrac{1}{\sqrt{2}}\boldsymbol{z}_{s+1}\boldsymbol{z}_{s+1}^\top\boldsymbol{C}_s^\top\boldsymbol{C}_s\tfrac{1}{\sqrt{2}}\boldsymbol{z}_{s+1}\boldsymbol{z}_{s+1}^\top\boldsymbol{D}_s]\\
&+ C\eta^2\mathbb{E}_s[(\langle\boldsymbol{\Delta}_s, \operatorname{He}_2(\boldsymbol{z}_{s+1})\rangle + \varepsilon_{s+1})^4\big(\boldsymbol{C}_s\tfrac{1}{\sqrt{2}}\boldsymbol{z}_{s+1}\boldsymbol{z}_{s+1}^\top\boldsymbol{M}_s\tfrac{1}{\sqrt{2}}\boldsymbol{z}_{s+1}\boldsymbol{z}_{s+1}^\top\boldsymbol{C}_s^\top\big)^2]\\
&\preceq C(\|\boldsymbol{\Delta}_s\|_F^2 + \mathbb{E}_s[\varepsilon_{s+1}^2])\operatorname{tr}(\boldsymbol{D}_s\boldsymbol{D}_s^\top)\boldsymbol{C}_s\boldsymbol{C}_s^\top\\
&+ C(\|\boldsymbol{\Delta}_s\|_F^2 + \mathbb{E}_s[\varepsilon_{s+1}^2])\operatorname{tr}(\boldsymbol{C}_s\boldsymbol{C}_s^\top)\boldsymbol{D}_s^\top\boldsymbol{D}_s\\
&+ C\eta^2(\|\boldsymbol{\Delta}_s\|_F^2 + \mathbb{E}_s[\varepsilon_{s+1}^2])^2\operatorname{tr}(\boldsymbol{C}_s\boldsymbol{C}_s^\top)\operatorname{tr}(\boldsymbol{M}_s)^2\boldsymbol{C}_s\boldsymbol{C}_s^\top
\end{aligned}$$

where we use Lemma 4 in the second inequality.

For the probability bound,

$$\mathbb{E}_s[\|\boldsymbol{Y}_{s+1}\|_2^p]^{1/p} \le 2\mathbb{E}_s[\|\boldsymbol{Q}_{s+1}\|_2^p]^{1/p} \le 2\mathbb{E}_s[\|\boldsymbol{Q}_{s+1,1}\|_2^p]^{1/p} + 2\eta\mathbb{E}_s[\|\boldsymbol{Q}_{s+1,2}\|_2^p]^{1/p}.$$

We have

$$\begin{aligned}
\mathbb{E}_s[\|\boldsymbol{Q}_{s+1,1}\|_2^p]^{1/p} &\le 2\mathbb{E}_s[|\langle\boldsymbol{\Delta}_s, \operatorname{He}_2(\boldsymbol{z}_{s+1})\rangle + \varepsilon_{s+1}|^{2p}]^{1/2p}\mathbb{E}_s[\|\boldsymbol{C}_s\boldsymbol{z}_{s+1}\boldsymbol{z}_{s+1}^\top\boldsymbol{D}_s\|_2^{2p}]^{1/2p}\\
&\le 2(\mathbb{E}_s[|\langle\boldsymbol{\Delta}_s, \operatorname{He}_2(\boldsymbol{z}_{s+1})\rangle|^{2p}]^{1/2p} + \mathbb{E}_s[|\varepsilon_{s+1}|^{2p}]^{1/2p})\\
&\qquad\times(\mathbb{E}_s[\|\boldsymbol{C}_s\boldsymbol{z}_{s+1}\|_2^{4p}]^{1/4p}\mathbb{E}_s[\|\boldsymbol{D}_s^\top\boldsymbol{z}_{s+1}\|_2^{4p}]^{1/4p})\\
&\le C_\varepsilon p^2(\|\boldsymbol{\Delta}_s\|_F^2 + \mathbb{E}_s[\varepsilon_{s+1}^2])^{1/2}\|\boldsymbol{C}_s\|_F\|\boldsymbol{D}_s\|_F\\
&= C_\varepsilon p^2 a_s.
\end{aligned}$$

Moreover,

$$\begin{aligned}
\mathbb{E}_s[\|\boldsymbol{Q}_{s+1,2}\|_2^p]^{1/p} &\le 2\mathbb{E}_s[|\langle\boldsymbol{\Delta}_s, \operatorname{He}_2(\boldsymbol{z}_{s+1})\rangle + \varepsilon_{s+1}|^{4p}]^{1/2p}\mathbb{E}_s[\|\boldsymbol{C}_s\tfrac{1}{\sqrt{2}}\boldsymbol{z}_{s+1}\boldsymbol{z}_{s+1}^\top\boldsymbol{M}_s^{1/2}\|_2^{4p}]^{1/2p}\\
&\le C_\varepsilon p^4(\|\boldsymbol{\Delta}_s\|_F^2 + \mathbb{E}_s[\varepsilon_{s+1}^2])\|\boldsymbol{C}_s\|_F^2\operatorname{tr}(\boldsymbol{M}_s)\\
&= C_\varepsilon p^4 b_s.
\end{aligned}$$

Therefore,

$$\mathbb{E}_s[\|\boldsymbol{Y}_{s+1}\|_2^p]^{1/p} \le C_\varepsilon(a_s p^2 + \eta b_s p^4)$$

By using $u = p$ with Markov's bound, we have the probability bound. □

For the next statement, we introduce $\kappa_\delta := \log \frac{Cr_u N}{\delta}$ and

$$v_s := C_\varepsilon \nu_s^2 \left[(\operatorname{tr} \mathsf{B}_s)\|\mathsf{A}_s\|_2 + (\operatorname{tr} \mathsf{A}_s)\|\mathsf{B}_s\|_2\right] + C_\varepsilon \eta^2 \nu_s^4 (\operatorname{tr} \mathsf{A}_s)(\operatorname{tr} \boldsymbol{M}_s)^2 \|\mathsf{A}_s\|_2.$$

Proposition 17. *For deterministic $v_*, a_*, b_* > 0$, we define a hitting time*

$$\mathcal{T}(v_*, a_*, b_*) := \inf\{s \geq 0 : v_s > v_* \ \text{or} \ a_s > a_* \ \text{or} \ b_s > b_*\}, \tag{F.9}$$

where we will use $\mathcal{T} = \mathcal{T}(v_*, a_*, b_*)$ whenever the arguments can be understood from the context. We have

$$\mathbb{P}\Big[\max_{1\leq t\leq N\wedge\mathcal{T}} \Big\|\eta \sum_{s=0}^{t-1} U_{t,s+1}(\boldsymbol{Y}_{s+1})\Big\|_2 \leq C_\varepsilon K_U^2 (\sqrt{\eta^2 N v_* \kappa_\delta} + \eta a_* \kappa_\delta^3 + \eta^2 b_* \kappa_\delta^5)\Big] \geq 1 - \delta.$$

Proof. We define

$$\boldsymbol{Z}_t := \sum_{s=0}^{t-1} \mathbb{1}_{s+1\leq\mathcal{T}} U_{t,s+1}(\boldsymbol{Y}_{s+1}) = \sum_{s=1}^{t} \underbrace{\mathbb{1}_{s\leq\mathcal{T}} U_{t,s}(\boldsymbol{Y}_s)}_{:=\boldsymbol{X}_{t,s}}.$$

We observe that

- For a fixed $t \in [N]$, $(\boldsymbol{X}_{t,s})_{s\leq t}$ is a martingale difference sequence.
- Moreover,

$$\max_{1\leq t\leq N\wedge\mathcal{T}} \Big\|\eta \sum_{s=0}^{t-1} U_{t,s+1}(\boldsymbol{Y}_{s+1})\Big\|_2 = \max_{1\leq t\leq N} \|\eta \boldsymbol{Z}_{t\wedge\mathcal{T}}\|_2 \leq \max_{1\leq t\leq N} \|\eta \boldsymbol{Z}_t\|_2.$$

Hence, in the following we will bound $\|\boldsymbol{Z}_t\|_2$ for a fixed $t \in [N]$ and then use the union bound for the result. We will use a clipping argument to bound $\|\boldsymbol{Z}_t\|_2$. We define $B(u) := C_\varepsilon K_U^2 (4a_* u^2 + 16\eta b_* u^4)$ for $u \geq 1$, and

$$\hat{\boldsymbol{Z}}_{t,r} := \sum_{s=1}^{r} \underbrace{\boldsymbol{X}_{t,s} \mathbb{1}_{\|\boldsymbol{X}_{t,s}\|_2\leq B(u)} - \mathbb{E}_{s-1}[\boldsymbol{X}_{t,s} \mathbb{1}_{\|\boldsymbol{X}_{t,s}\|_2\leq B(u)}]}_{:=\hat{\boldsymbol{X}}_{t,s}}$$

By the triangle inequality, we have

$$\|\boldsymbol{Z}_t\|_2 \leq \|\hat{\boldsymbol{Z}}_{t,t}\|_2 + \sum_{s=1}^{t} \big\|\boldsymbol{X}_{t,s} \mathbb{1}_{\|\boldsymbol{X}_{t,s}\|_2 > B(u)} - \mathbb{E}_{s-1}[\boldsymbol{X}_{t,s} \mathbb{1}_{\|\boldsymbol{X}_{t,s}\|_2 > B(u)}]\big\|_2.$$

Here, we have

- *Variance estimate.* We have

$$\|\mathbb{E}_{s-1}[\boldsymbol{X}_{t,s}^2]\|_2 \overset{(a)}{\leq} \mathbb{1}_{s\leq\mathcal{T}} K_U^4 \|\mathbb{E}_{s-1}[\boldsymbol{Y}_s^2]\|_2 \leq \mathbb{1}_{s\leq\mathcal{T}} K_U^4 v_{s-1} \overset{(b)}{\leq} K_U^4 v_*. \tag{F.10}$$

  where we used Lemma 8 in (a), and (F.11) in (b).

- *Clip bias.* We have

$$\big\|\mathbb{E}_{s-1}[\boldsymbol{X}_{t,s} \mathbb{1}_{\|\boldsymbol{X}_{t,s}\|_2 > B(u)}]\big\|_2 \overset{(c)}{\leq} \|\mathbb{E}_{s-1}[\boldsymbol{X}_{t,s}^2]\|_2^{1/2} e^{-u} \overset{(d)}{\leq} K_U^2 \sqrt{v_*} e^{-u}$$

  where we used Cauchy-Schwartz and (F.8) in (c), and (F.10) in (d).

Then, by the union bound for $s \leq t (\leq N)$,

$$\mathbb{P}[\|\boldsymbol{Z}_t\|_2 > \gamma] \leq \mathbb{P}[\|\hat{\boldsymbol{Z}}_{t,t}\|_2 > \gamma - K_U^2 \sqrt{N v_*} \sqrt{N e^{-2u}}] + N e^{-2u}$$

To bound $\hat{\boldsymbol{Z}}_{t,t}$, we observe that $\|\hat{\boldsymbol{X}}_{t,s}\|_2 \leq 2B(u)$ and $\|\mathbb{E}_{s-1}[\hat{\boldsymbol{X}}_{t,s}^2]\|_2 \leq K_U^4 v_*$. By [Tro15] and choosing C in κ_δ accordingly, we have for $t \leq N$,

$$\mathbb{P}\Big[\|\hat{\boldsymbol{Z}}_{t,t}\|_2 > K_U^2 \big(\sqrt{2N v_* \kappa_\delta} + \frac{4}{3} C_\varepsilon \kappa_\delta (4a_* u^2 + 16\eta b_* u^4)\big)\Big] \leq \frac{\delta}{2N}$$

By choosing $u = \kappa_\delta$ and $\gamma = 2K_U^2 \big(\sqrt{N v^* \kappa_\delta} + \frac{4}{3} C_\varepsilon (4a_* \kappa_\delta^3 + 16\eta b_* \kappa_\delta^5)\big)$, we have $\mathbb{P}[\|\boldsymbol{Z}_t\|_2 > \gamma] < \delta/N$. By using union bound and hiding constants in C_ε, we have the result. □

F.2.2 Lower block

In this part, we will bound $\sum_{s=1}^{t}(\boldsymbol{I}_{d-r_u}+\eta\boldsymbol{A}_{22})^{-s}\boldsymbol{\Xi}_{s,22}(\boldsymbol{I}_{d-r_u}+\eta\boldsymbol{A}_{22})^{-s}$ in (D.7).

- For notational convenience, we will use $\boldsymbol{D}\coloneqq\boldsymbol{I}_{d-r_u}+\eta\boldsymbol{A}_{22}$ and $\boldsymbol{Z}_s\coloneqq\boldsymbol{D}^{-s}\boldsymbol{\Xi}_{s,22}\boldsymbol{D}^{-s}$.
- We observe that $\boldsymbol{Z}_{s+1}$ is a $\mathcal{F}_{s+1}$-measurable symmetric random matrix and $\mathbb{E}_s[\boldsymbol{Z}_{s+1}]=0$.

Proposition 18. *We define* $\nu_s^2\coloneqq(\|\boldsymbol{T}^*-\boldsymbol{M}_s\|_F^2+\mathbb{E}_s[\varepsilon_{s+1}^2])$ *and*

$$\mathsf{C}_s\coloneqq\boldsymbol{D}^{-(s+1)}\boldsymbol{\Sigma}_{22}^2\boldsymbol{D}^{-(s+1)},\qquad \mathsf{F}_s\coloneqq\boldsymbol{D}^{-(s+1)}(\boldsymbol{M}_{s,22}^2+\boldsymbol{M}_{s,12}^\top\boldsymbol{M}_{s,12})\boldsymbol{D}^{-(s+1)}$$
$$c_s\coloneqq\nu_s\sqrt{\operatorname{tr}(\mathsf{C}_s)\operatorname{tr}(\mathsf{F}_s)},\qquad f_s\coloneqq\nu_s^2\operatorname{tr}(\mathsf{C}_s)\operatorname{tr}(\boldsymbol{M}_s).$$

For a constant $C_\varepsilon>0$ depending on only the noise distribution, we have

- $\mathbb{E}_s[\boldsymbol{Z}_{s+1}^2]\preceq C_\varepsilon\nu_s^2[(\operatorname{tr}\mathsf{F}_s)\mathsf{C}_s+(\operatorname{tr}\mathsf{C}_s)\mathsf{F}_s]+C_\varepsilon\eta^2\nu_s^4(\operatorname{tr}\mathsf{C}_s)(\operatorname{tr}\boldsymbol{M}_s)^2\mathsf{C}_s$

  $\mathbb{P}_s[\|\boldsymbol{Z}_{s+1}\|_2>C_\varepsilon(c_su^2+\eta f_su^4)]\le e^{-u}$, *for* $u\ge2$.

- $\mathbb{E}_s[\operatorname{tr}(\boldsymbol{Z}_{s+1})^2]\le C_\varepsilon[c_s^2+\eta^2f_s^2]$

  $\mathbb{P}_s[|\operatorname{tr}(\boldsymbol{Z}_{s+1})|>C_\varepsilon(c_su^2+\eta f_su^4)]\le e^{-u}$, *for* $u\ge2$.

Proof. We define

$$\boldsymbol{J}_s\coloneqq\boldsymbol{D}^{-(s+1)}\begin{bmatrix}0 & \boldsymbol{I}_{d-r_u}\end{bmatrix},\qquad\tilde{\boldsymbol{C}}_s\coloneqq\boldsymbol{J}_s\boldsymbol{\Sigma},\qquad\tilde{\boldsymbol{D}}_s\coloneqq\boldsymbol{M}_s\boldsymbol{J}_s^\top.$$

We have

$$\mathsf{C}_s=\tilde{\boldsymbol{C}}_s\tilde{\boldsymbol{C}}_s^\top\ \text{ and }\ \mathsf{F}_s=\tilde{\boldsymbol{D}}_s^\top\tilde{\boldsymbol{D}}_s.$$

We define $\boldsymbol{Z}_{s+1}\coloneqq\boldsymbol{Q}_{s+1}-\mathbb{E}_s[\boldsymbol{Q}_{s+1}]$, where

$$\boldsymbol{Q}_{s+1}=\underbrace{(\langle\boldsymbol{\Delta}_s,\mathrm{He}_2(\boldsymbol{z}_{s+1})\rangle+\varepsilon_{s+1})(\tilde{\boldsymbol{C}}_s\tfrac{1}{\sqrt{2}}\boldsymbol{z}_{s+1}\boldsymbol{z}_{s+1}^\top\tilde{\boldsymbol{D}}_s+\tilde{\boldsymbol{D}}_s^\top\tfrac{1}{\sqrt{2}}\boldsymbol{z}_{s+1}\boldsymbol{z}_{s+1}^\top\tilde{\boldsymbol{C}}_s^\top)}_{\coloneqq\boldsymbol{Q}_{s+1,1}}$$
$$+\eta\underbrace{(\langle\boldsymbol{\Delta}_s,\mathrm{He}_2(\boldsymbol{z}_{s+1})\rangle+\varepsilon_{s+1})^2\tilde{\boldsymbol{C}}_s\tfrac{1}{\sqrt{2}}\boldsymbol{z}_{s+1}\boldsymbol{z}_{s+1}^\top\boldsymbol{M}_s\tfrac{1}{\sqrt{2}}\boldsymbol{z}_{s+1}\boldsymbol{z}_{s+1}^\top\tilde{\boldsymbol{C}}_s^\top}_{\coloneqq\boldsymbol{Q}_{s+1,2}}.$$

The first part of the proof is the same with the proof of Proposition 16. Therefore, we focus on the trace terms. We have

$$\operatorname{tr}(\boldsymbol{Q}_{s+1,1})=\sqrt{2}(\langle\boldsymbol{\Delta}_s,\mathrm{He}_2(\boldsymbol{z}_{s+1})\rangle+\varepsilon_{s+1})\boldsymbol{z}_{s+1}^\top\tilde{\boldsymbol{D}}_s\tilde{\boldsymbol{C}}_s\boldsymbol{z}_{s+1}$$
$$\operatorname{tr}(\boldsymbol{Q}_{s+1,2})=(\langle\boldsymbol{\Delta}_s,\mathrm{He}_2(\boldsymbol{z}_{s+1})\rangle+\varepsilon_{s+1})^2\|\tilde{\boldsymbol{C}}_s\mathrm{He}_2(\boldsymbol{z}_{s+1})\boldsymbol{M}_s^{\frac{1}{2}}\|_F^2.$$

Therefore, we have

$$\mathbb{E}_s[\operatorname{tr}(\boldsymbol{Q}_{s+1,1})^2]\le C\nu_s^2\operatorname{tr}(\mathsf{C}_s)\operatorname{tr}(\mathsf{F}_s)=Cc_s^2,$$
$$\mathbb{E}_s[\operatorname{tr}(\boldsymbol{Q}_{s+1,2})^2]\le C_\varepsilon\nu_s^4\operatorname{tr}(\mathsf{C}_s)^2\operatorname{tr}(\boldsymbol{M}_s)^2=C_\varepsilon f_s^2,$$

which gives the variance bound. Moreover, by Gaussian moments, we have

$$\mathbb{E}_s[|\operatorname{tr}(\boldsymbol{Q}_{s+1,1})|^p]^{\frac{1}{p}}\le C_\varepsilon p^2c_s,\qquad\mathbb{E}_s[|\operatorname{tr}(\boldsymbol{Q}_{s+1,2})|^p]^{\frac{1}{p}}\le C_\varepsilon p^4f_s.$$

By Markov's bound, we have the result. □

For the next statement, we re-introduce

$$v_s=C_\varepsilon(c_s^2+\eta^2f_s^2),\qquad\kappa_\delta\coloneqq\log\frac{CdN}{\delta}.$$

Proposition 19. *For deterministic $v_*,c_*,f_*>0$, we define a hitting time*

$$\mathcal{T}(v_*,c_*,f_*)\coloneqq\inf\{s\ge0:v_s>v_*\ \text{ or }\ c_s>c_*\ \text{ or }\ f_s>f_*\},\tag{F.11}$$

where we will use $\mathcal{T}=\mathcal{T}(v_*,c_*,f_*)$ whenever the arguments can be understood from the context. We have

$$\mathbb{P}\Big[\max_{1\le t\le N\wedge\mathcal{T}}\Big\|\sum_{s=0}^{t-1}\boldsymbol{Z}_{s+1}\Big\|_2\vee\Big|\operatorname{tr}\Big(\sum_{s=0}^{t-1}\boldsymbol{Z}_{s+1}\Big)\Big|\le C_\varepsilon K_U^2(\sqrt{Nv_*\kappa_\delta}+c_*\kappa_\delta^3+\eta f_*\kappa_\delta^5)\Big]\ge1-\delta.$$

Proof. The proof is identical to the proof of Proposition 17. □

G Miscellaneous

G.1 Proof of Proposition 7

Proof of Proposition 7. We will use

$$\begin{aligned}
\boldsymbol{B}_t &\coloneqq \boldsymbol{\Sigma}_{11}(\boldsymbol{T}_{11}^* - \boldsymbol{L}_t), \qquad & \boldsymbol{C}_t &\coloneqq \boldsymbol{\Sigma}_{11}\hat{\boldsymbol{R}}_t\boldsymbol{L}_t,\\
\boldsymbol{P}_t &\coloneqq \boldsymbol{I}_{r_u} - \eta\boldsymbol{B}_t + \eta^2\boldsymbol{B}_t^2, \quad & \boldsymbol{K}_t &\coloneqq \boldsymbol{L}_{t+1}^{-1/2}\boldsymbol{M}_{t,12}\boldsymbol{M}_{t,12}^\top\boldsymbol{L}_{t+1}^{-1/2}.
\end{aligned}$$

We have

$$\boldsymbol{L}_{t+1}^{-1/2}\boldsymbol{M}_{t+1,11}\boldsymbol{L}_{t+1}^{-1/2} = \boldsymbol{P}_t\boldsymbol{L}_t^{-1/2}\boldsymbol{M}_{t+1,11}\boldsymbol{L}_t^{-1/2}\boldsymbol{P}_t$$

We have

$$\begin{aligned}
\boldsymbol{L}_t^{-1/2}\boldsymbol{M}_{t+1,11}\boldsymbol{L}_t^{-1/2} &= \boldsymbol{R}_t + \eta\big(\boldsymbol{A}_{11}\boldsymbol{R}_t + \boldsymbol{R}_t\boldsymbol{A}_{11}\big)\\
&\quad - \eta\big(\boldsymbol{\Sigma}_{11}(\boldsymbol{R}_t\boldsymbol{L}_t\boldsymbol{R}_t + \boldsymbol{K}_t) + (\boldsymbol{R}_t\boldsymbol{L}_t\boldsymbol{R}_t + \boldsymbol{K}_t)\boldsymbol{\Sigma}_{11}\big)\\
&\quad + \eta^2\boldsymbol{L}_t^{-1/2}\boldsymbol{H}_{t,11}\boldsymbol{L}_t^{-1/2} + \eta\tilde{\boldsymbol{\Xi}}_{t+1,11}\\
&= \boldsymbol{R}_t + \eta\Big((\boldsymbol{B}_t - \boldsymbol{C}_t)\boldsymbol{R}_t + \boldsymbol{R}_t(\boldsymbol{B}_t - \boldsymbol{C}_t) - \boldsymbol{\Sigma}_{11}\boldsymbol{K}_t - \boldsymbol{K}_t\boldsymbol{\Sigma}_{11}\Big)\\
&\quad + \eta^2(\boldsymbol{B}_t - \boldsymbol{C}_t)\boldsymbol{R}_t(\boldsymbol{B}_t - \boldsymbol{C}_t) + \eta^2\tilde{\boldsymbol{H}}_{t,11} + \eta\tilde{\boldsymbol{\Xi}}_{t+1,11}\\
&= (\boldsymbol{I}_{r_u} + \eta(\boldsymbol{B}_t - \boldsymbol{C}_t))\boldsymbol{R}_t(\boldsymbol{I}_{r_u} + \eta(\boldsymbol{B}_t - \boldsymbol{C}_t)) - \eta(\boldsymbol{\Sigma}_{11}\boldsymbol{K}_t + \boldsymbol{K}_t\boldsymbol{\Sigma}_{11})\\
&\quad + \eta^2\tilde{\boldsymbol{H}}_{t,11} + \eta\tilde{\boldsymbol{\Xi}}_{t+1,11}
\end{aligned}$$

We define

$$\boldsymbol{V}_t \coloneqq \boldsymbol{P}_t(\boldsymbol{I}_{r_u} + \eta(\boldsymbol{B}_t - \boldsymbol{C}_t)), \qquad \boldsymbol{J}_t \coloneqq \boldsymbol{\Sigma}_{11}\boldsymbol{K}_t + \boldsymbol{K}_t\boldsymbol{\Sigma}_{11}, \qquad \boldsymbol{Q}_t = \boldsymbol{B}_t^2(\boldsymbol{B}_t - \boldsymbol{C}_t).$$

We have

$$\boldsymbol{R}_{t+1} = \boldsymbol{V}_t\boldsymbol{R}_t\boldsymbol{V}_t - \eta\boldsymbol{P}_t\boldsymbol{J}_t\boldsymbol{P}_t + \eta^2\boldsymbol{P}_t\tilde{\boldsymbol{H}}_{t,11}\boldsymbol{P}_t + \eta\boldsymbol{P}_t\tilde{\boldsymbol{\Xi}}_{t+1,11}\boldsymbol{P}_t.$$

We have

$$\begin{aligned}
\boldsymbol{V}_t &= \boldsymbol{P}_t(\boldsymbol{I}_{r_u} + \eta\boldsymbol{B}_t) - \eta\boldsymbol{P}_t\boldsymbol{C}_t\\
&= \boldsymbol{I}_{r_u} + \eta^3\boldsymbol{B}_t^3 - \eta\boldsymbol{C}_t + \eta^2\boldsymbol{B}_t\boldsymbol{C}_t - \eta^3\boldsymbol{B}_t^2\boldsymbol{C}_t = \boldsymbol{I}_{r_u} - \eta\boldsymbol{C}_t + \eta^2\boldsymbol{B}_t\boldsymbol{C}_t - \eta^3\boldsymbol{Q}_t
\end{aligned}$$

We have

$$\boldsymbol{V}_t\boldsymbol{R}_t\boldsymbol{V}_t^\top = \boldsymbol{R}_t - \eta(\boldsymbol{C}_t\boldsymbol{R}_t + \boldsymbol{R}_t\boldsymbol{C}_t^\top) + \eta^2\left(\boldsymbol{B}_t\boldsymbol{C}_t\boldsymbol{R}_t + \boldsymbol{R}_t\boldsymbol{C}_t^\top\boldsymbol{B}_t + \boldsymbol{C}_t\boldsymbol{R}_t\boldsymbol{C}_t^\top\right) + \eta^3\boldsymbol{F}_t^{(R)},$$

where the exact expression for $\boldsymbol{F}_t^{(R)}$ is given below. Similarly,

$$-\eta\boldsymbol{P}_t\boldsymbol{J}_t\boldsymbol{P}_t = -\eta\boldsymbol{J}_t + \eta^2(\boldsymbol{B}_t\boldsymbol{J}_t + \boldsymbol{J}_t\boldsymbol{B}_t) + \eta^3\boldsymbol{F}_t^{(K)}$$

where

$$\boldsymbol{F}_t^{(K)} \coloneqq -(\boldsymbol{B}_t^2\boldsymbol{J}_t + \boldsymbol{B}_t\boldsymbol{J}_t\boldsymbol{B}_t + \boldsymbol{J}_t\boldsymbol{B}_t^2) + \eta(\boldsymbol{B}_t^2\boldsymbol{J}_t\boldsymbol{B}_t + \boldsymbol{B}_t\boldsymbol{J}_t\boldsymbol{B}_t^2) - \eta^2\boldsymbol{B}_t^2\boldsymbol{J}_t\boldsymbol{B}_t^2.$$

and

$$\begin{aligned}
\boldsymbol{F}_t^{(R)} &= \boldsymbol{Q}_t\boldsymbol{R}_t + \boldsymbol{R}_t\boldsymbol{Q}_t^\top - \boldsymbol{B}_t\boldsymbol{C}_t\boldsymbol{R}_t\boldsymbol{C}_t^\top - \boldsymbol{C}_t\boldsymbol{R}_t\boldsymbol{C}_t^\top\boldsymbol{B}_t\\
&\quad + \eta\Big[\boldsymbol{B}_t\boldsymbol{C}_t\boldsymbol{R}_t\boldsymbol{C}_t^\top\boldsymbol{B}_t - \boldsymbol{C}_t\boldsymbol{R}_t\boldsymbol{Q}_t^\top - \boldsymbol{Q}_t\boldsymbol{R}_t\boldsymbol{C}_t^\top\Big]\\
&\quad + \eta^2\Big[\boldsymbol{B}_t\boldsymbol{C}_t\boldsymbol{R}_t\boldsymbol{Q}_t^\top + \boldsymbol{Q}_t\boldsymbol{R}_t\boldsymbol{C}_t^\top\boldsymbol{B}_t\Big] + \eta^3\boldsymbol{Q}_t\boldsymbol{R}_t\boldsymbol{Q}_t^\top.
\end{aligned}$$

We define $\boldsymbol{F}_t \coloneqq \boldsymbol{F}_t^{(R)} + \boldsymbol{F}_t^{(K)}$. We have

$$\begin{aligned}
\boldsymbol{R}_{t+1} &= \boldsymbol{R}_t - \eta(\boldsymbol{C}_t\boldsymbol{R}_t + \boldsymbol{R}_t\boldsymbol{C}_t^\top + \boldsymbol{J}_t)\\
&\quad + \eta^2(\boldsymbol{B}_t\boldsymbol{C}_t\boldsymbol{R}_t + \boldsymbol{R}_t\boldsymbol{C}_t^\top\boldsymbol{B}_t + \boldsymbol{C}_t\boldsymbol{R}_t\boldsymbol{C}_t^\top + \boldsymbol{B}_t\boldsymbol{J}_t + \boldsymbol{J}_t\boldsymbol{B}_t + \boldsymbol{P}_t\tilde{\boldsymbol{H}}_t\boldsymbol{P}_t)\\
&\quad + \eta^3\boldsymbol{F}_t + \eta\boldsymbol{P}_t\tilde{\boldsymbol{\Xi}}_{t+1,11}\boldsymbol{P}_t.
\end{aligned}$$

We have

- $\boldsymbol{C}_t\boldsymbol{R}_t + \boldsymbol{R}_t\boldsymbol{C}_t^\top = \boldsymbol{\Sigma}_{11}\hat{\boldsymbol{R}}_t\boldsymbol{L}_t + \boldsymbol{L}_t\hat{\boldsymbol{R}}_t\boldsymbol{\Sigma}_{11} + \boldsymbol{\Sigma}_{11}\hat{\boldsymbol{R}}_t\boldsymbol{L}_t\hat{\boldsymbol{R}}_t + \hat{\boldsymbol{R}}_t\boldsymbol{L}_t\hat{\boldsymbol{R}}_t\boldsymbol{\Sigma}_{11}$.
- $\boldsymbol{B}_t\boldsymbol{C}_t\boldsymbol{R}_t = \boldsymbol{\Sigma}_{11}^2(\boldsymbol{T}_{11}^* - \boldsymbol{L}_t)\hat{\boldsymbol{R}}_t\boldsymbol{L}_t\boldsymbol{R}_t$
- $\boldsymbol{C}_t\boldsymbol{R}_t\boldsymbol{C}_t^\top = \boldsymbol{\Sigma}_{11}\hat{\boldsymbol{R}}_t\boldsymbol{L}_t\boldsymbol{R}_t\boldsymbol{L}_t\hat{\boldsymbol{R}}_t\boldsymbol{\Sigma}_{11}$

We have

$$\begin{aligned}
\mathcal{D}_t = & -(\boldsymbol{\Sigma}_{11}\hat{\boldsymbol{R}}_t\boldsymbol{L}_t\hat{\boldsymbol{R}}_t + \hat{\boldsymbol{R}}_t\boldsymbol{L}_t\hat{\boldsymbol{R}}_t\boldsymbol{\Sigma}_{11} + \boldsymbol{J}_t) \\
& + \eta(\boldsymbol{B}_t\boldsymbol{C}_t\boldsymbol{R}_t + \boldsymbol{R}_t\boldsymbol{C}_t^\top\boldsymbol{B}_t + \boldsymbol{C}_t\boldsymbol{R}_t\boldsymbol{C}_t^\top + \boldsymbol{B}_t\boldsymbol{J}_t + \boldsymbol{J}_t\boldsymbol{B}_t + \boldsymbol{P}_t\tilde{\boldsymbol{H}}_t\boldsymbol{P}_t) \\
& - \eta\boldsymbol{\Sigma}_{11}\boldsymbol{L}_t\hat{\boldsymbol{R}}_t\boldsymbol{L}_t\boldsymbol{\Sigma}_{11} + \eta^2\boldsymbol{F}_t \\
= & -(\boldsymbol{\Sigma}_{11}\hat{\boldsymbol{R}}_t\boldsymbol{L}_t\hat{\boldsymbol{R}}_t + \hat{\boldsymbol{R}}_t\boldsymbol{L}_t\hat{\boldsymbol{R}}_t\boldsymbol{\Sigma}_{11} + \boldsymbol{\Sigma}_{11}\boldsymbol{K}_t + \boldsymbol{K}_t\boldsymbol{\Sigma}_{11}) \\
& + \eta\big(\boldsymbol{\Sigma}_{11}^2(\boldsymbol{T}_{11}^* - \boldsymbol{L}_t)\hat{\boldsymbol{R}}_t\boldsymbol{L}_t\boldsymbol{R}_t + \boldsymbol{R}_t\boldsymbol{L}_t\hat{\boldsymbol{R}}_t(\boldsymbol{T}_{11}^* - \boldsymbol{L}_t)\boldsymbol{\Sigma}_{11}^2\big) \\
& + \big(\boldsymbol{\Sigma}_{11}\hat{\boldsymbol{R}}_t\boldsymbol{L}_t\boldsymbol{R}_t\boldsymbol{L}_t\hat{\boldsymbol{R}}_t\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{11}\boldsymbol{L}_t\hat{\boldsymbol{R}}_t\boldsymbol{L}_t\boldsymbol{\Sigma}_{11}\big) \\
& + \eta(\boldsymbol{\Sigma}_{11}(\boldsymbol{T}_{11}^* - \boldsymbol{L}_t)(\boldsymbol{\Sigma}_{11}\boldsymbol{K}_t + \boldsymbol{K}_t\boldsymbol{\Sigma}_{11}) + (\boldsymbol{\Sigma}_{11}\boldsymbol{K}_t + \boldsymbol{K}_t\boldsymbol{\Sigma}_{11})(\boldsymbol{T}_{11}^* - \boldsymbol{L}_t)\boldsymbol{\Sigma}_{11}) + \eta\boldsymbol{P}_t\tilde{\boldsymbol{H}}_t\boldsymbol{P}_t \\
& + \eta^2\boldsymbol{F}_t.
\end{aligned}$$

With this, we have

$$\hat{\boldsymbol{R}}_{t+1} = \hat{\boldsymbol{R}}_t - \eta(\boldsymbol{\Sigma}_{11}\hat{\boldsymbol{R}}_t\boldsymbol{L}_t + \boldsymbol{L}_t\hat{\boldsymbol{R}}_t\boldsymbol{\Sigma}_{11}) + \eta\mathcal{D}_t + \eta\boldsymbol{P}_t\tilde{\boldsymbol{\Xi}}_{t+1,11}\boldsymbol{P}_t.$$

In the remaining part, we will bound $\boldsymbol{F}_t$. Under the conditions in (D.9), we observe the followings:

- $\|\boldsymbol{B}_t\|_2 \le 1$ and $\|\boldsymbol{C}_t\|_2 \le 1 + \|\boldsymbol{R}_t\|_2$
- $\|\boldsymbol{Q}_t\|_2 \le 1 + \|\boldsymbol{R}_t\|_2$ and $\|\boldsymbol{J}_t\|_2 \le 2\|\boldsymbol{K}_t\|_2$
- We will use $r_t \coloneqq \|\boldsymbol{R}_t\|_2$ and $w_t \coloneqq 1 + \|\boldsymbol{R}_t\|_2$

Then, we have

$$\|\boldsymbol{F}_t^{(R)}\|_2 \le 2rw + (2 + 3\eta + 2\eta^2 + \eta^3)rw^2 \le 3w^3.$$

Moreover,

$$\|\boldsymbol{F}_t^{(K)}\|_2 \le (3 + 2\eta + \eta^2)\|\boldsymbol{J}_t\|_2 \le (6 + 4\eta + 2\eta^2)\|\boldsymbol{K}_t\|_2 \le 7\|\boldsymbol{K}_t\|_2.$$

Hence, we obtain the bound in (D.10), which concludes our proof. □

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: This is a ML theory paper. The claims made in the abstract and the introduction are all proved by rigorous mathematical arguments. The full details are presented in the appendix.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See Section 6. In particular, we provided three limitations of our analyses: 1) spectral alignment; 2) focus on two-layer NN; 3) SGD instability only via emperical evidence.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: We provide full set of assumptions and theorem statements clearly stated in the main body. We provide all proof steps in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The experiments are performed on two-layer neural networks. The results are merely a demonstration of our our theoretical predictions. We provide all experimental details in the captions of the figures.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: NeurIPS guideline states: "If any of the main contributions of your paper depends on an experimental result,...". The main contributions in this paper is theoretical. The experiments are provided as supporting evidence of our theory.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The experiments are performed on two-layer neural networks. The results are merely a demonstration of our our theoretical predictions. We provide all experimental details in the main text.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The main contributions in this paper is theoretical. The experiments are provided as supporting evidence of our theory, so we do not find the error bars to be necessary.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The experiments are performed on two-layer neural networks. The results are merely a demonstration of our our theoretical predictions. We provide all experimental details in the captions of the figures.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This is a theoretical work explaining a practical phenomenon in machine learning. There are no ethics concerns around our paper.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: This is a theoretical work explaining a practical phenomenon in machine learning. We do not expect any societal impact of our work.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: This is a ML theory paper. We do no expect any misuse of our results.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [N/A]

Justification: this is a ML theory paper. Our analyses are mathematical and do no involve any assets.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: this is a ML theory paper. Our analyses are mathematical and do no involve any assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: this is a ML theory paper. Our analyses are mathematical and do no involve any crowdsoursing or human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: this is a ML theory paper. Our analyses are mathematical and do no involve any human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are only used for writing, editing, or formatting.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.