Feature Learning in High-Dimensions under Structured Covariance: Scaling Laws in Quadratic Networks

Qingyuan Yu[*] QINGYUANYU@CS.TORONTO.EDU
University of Toronto and Vector Institute, Toronto, Canada

Nuri Mert Vural[*] VURAL@CS.TORONTO.EDU
University of Toronto and Vector Institute, Toronto, Canada

Xin T. Tong XIN.T.TONG@NUS.EDU.SG
National University of Singapore, Singapore

Murat A. Erdogdu ERDOGDU@CS.TORONTO.COM
University of Toronto and Vector Institute, Toronto, Canada

Abstract

Recent theoretical work has shown that nonlinear solvable models exhibit scaling laws in the feature-learning regime. However, these results largely rely on the assumption of isotropic inputs, and understanding how these laws extend to anisotropic data remains a central open problem. In this work, we address this gap by analyzing the learning dynamics of two-layer neural networks with quadratic activations under anisotropic Gaussian inputs. We provide a sharp characterization of online stochastic gradient descent (SGD), explicitly quantifying how the covariance spectrum influences both the scaling exponent and sample complexity. Furthermore, we establish that normalization techniques overcome the intrinsic limitations of vanilla SGD, strictly improving sample efficiency in anisotropic settings. Experiments on two-layer networks with general activation functions support our theoretical predictions, suggesting that these insights extend well beyond the quadratic model.

Keywords: scaling laws, stochastic gradient descent, shallow neural network, multi-index model

1. Introduction

The predictable scaling of generalization error with respect to computational resources, dataset size, and model width—commonly referred to as scaling laws—has become a defining feature of modern machine learning [10, 11]. Because empirical losses consistently exhibit robust power-law decay, practitioners can reliably extrapolate large-scale performance from a relatively small number of measurements [2, 12]. This phenomenon has sparked significant theoretical interest in deriving scaling laws from first principles [18, 21]. While earlier work primarily focused on linear models [16, 22], recent efforts have shifted toward nonlinear solvable models, which are necessary to capture phenomena such as feature learning [7] and and the heavy spectral tails observed in real-world representations [13].

Despite this progress, existing scaling analyses for nonlinear networks largely rely on the restrictive assumption of isotropic inputs [3, 23]. In this work, we study the anisotropic setting. Specifically, we analyze the learning dynamics of a two-layer neural network with quadratic activations under anisotropic Gaussian data:

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

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

∗. Equal contribution.

matrix, which may exhibit a power-law spectrum. The constant terms $\langle \boldsymbol{\theta}_j, \boldsymbol{\Sigma}\boldsymbol{\theta}_j \rangle$ are included to ensure the labels remain centered, i.e., $\mathbb{E}[y] = 0$. Our goal is to learn this teacher network using a two-layer student network of width r_s trained via gradient-based optimization.

A central challenge in this setting is understanding the fundamental statistical and computational limits of feature learning under anisotropy. Existing work provides sample complexity guarantees bounded by the effective dimension $r_{\text{eff}}(\boldsymbol{\Sigma}) \coloneqq \operatorname{tr}(\boldsymbol{\Sigma})/\|\boldsymbol{\Sigma}\|_2$: scaling quadratically as $(r_{\text{eff}}(\boldsymbol{\Sigma}))^2$ for kernel methods [8, 24], and scaling linearly as $r_{\text{eff}}(\boldsymbol{\Sigma})$ for infinite-width mean-field models [20]. However, the exact risk trajectory of stochastic gradient descent (SGD) applied to narrow networks—and the corresponding scaling behavior—remains poorly understood. More broadly, theoretical analyses of nonlinear multi-index models under anisotropic covariates remain limited. Motivated by these gaps, we investigate the following question:

Can we characterize the exact risk trajectory and the resulting computational and statistical requirements of SGD for quadratic networks under anisotropic data?

1.1. Our contributions

In this paper, we study the model in (1.1) under structured power-law conditions on both the signal strength and the covariance spectrum. Specifically, letting σ_i denote the eigenvalues of $\boldsymbol{\Sigma}$, 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 existing literature, α and β is closely related to the *source* and *capacity* conditions, respectively [5, 24]. Our main contributions are summarized below:

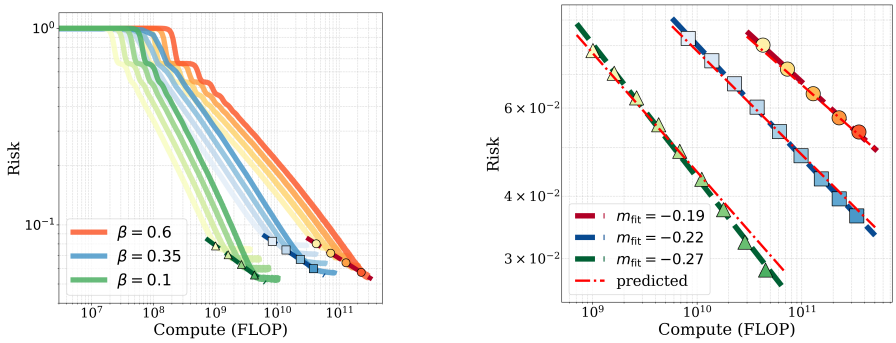
1. **Scaling exponents:** In Section 3, we derive a sharp characterization of the excess risk under online vanilla SGD dynamics. In particular, we show that the risk exhibits an explicit power-law decay jointly governed by the source and capacity exponents α and β. We validate these predictions empirically, observing a strong agreement between theory and simulations (Figure 1).
2. **Algorithmic lower bound:** In Section 3.1, we study the limitations of vanilla online SGD under anisotropic data. Combining theoretical results with supporting empirical evidence, we show that unconstrained SGD is restricted to conservative learning rates, which force its sample complexity to $\Omega(d)$. This prevents vanilla SGD from attaining the near-optimal sample complexity $r_{\text{eff}}(\boldsymbol{\Sigma})$.
3. **Near-optimal sample complexity:** In Section 3.2, we propose and analyze an online Stiefel SGD algorithm. We prove that it achieves a near-optimal sample complexity of $\tilde{\mathcal{O}}(r_{\text{eff}}(\boldsymbol{\Sigma}))$, matching the effective dimension of the data up to logarithmic factors while preserving the exact scaling behavior. This establishes a sharp separation in sample efficiency between SGD learning in feature-learning regime and lazy training regimes under anisotropic data.

2. Problem Setting

We consider learning a two-layer teacher network with a quadratic activation when the input is anisotropic Gaussian given in (1.1), under the following assumption:

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 $\boldsymbol{x}$ [9]. 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 [17].



(a) (b)

Figure 1: (a) Population risk vs. compute for two-layer quadratic networks trained via one-pass SGD for $\beta \in \{0.1, 0.35, 0.6\}$ and $r_s \in \{128, 152, 181, 215\}$ ($d = 2048$, $\alpha = 0.7 - \beta$). (b) Empirical scaling exponents across model widths, with theoretical predictions (red lines) closely matching observed slopes.

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

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

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

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

By substituting the teacher and student models in (1.1) and , the population risk admits the following closed-form expression:

$$R(\boldsymbol{W}, b) = \Big\| \tfrac{1}{r_s} \boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{W}\boldsymbol{W}^\top \boldsymbol{\Sigma}^{\frac{1}{2}} - \boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{\Theta}\boldsymbol{\Lambda}\boldsymbol{\Theta}^\top \boldsymbol{\Sigma}^{\frac{1}{2}} \Big\|_{\mathrm{F}}^2 + \tfrac{1}{2}\big(b - \tfrac{1}{r_s}\operatorname{tr}(\boldsymbol{\Sigma}\boldsymbol{W}\boldsymbol{W}^\top)\big)^2 + \tfrac{1}{2}\mathbb{E}[\epsilon^2]. \quad (2.1)$$

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 structured covariance

3.1. Online vanilla SGD

We first analyze the learning dynamics of online (one-pass) SGD. Given a fresh sample $(\boldsymbol{x}_{t+1}, y_{t+1})$, we define the instantaneous loss:

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

where $\hat{y}(\boldsymbol{x}_{t+1}; \boldsymbol{W}, b)$ denotes the network output. To learn the weights and bias simultaneously, we employ two-time-scale gradient descent, where the bias updates significantly 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 (2.1) that keeps the predictions centered. Specifically, the parameter updates are given by:

$$\boldsymbol{W}_{t+1} = \boldsymbol{W}_t - \eta \nabla \mathcal{L}_{t+1}(\boldsymbol{W}_t, b_t), \qquad b_{t+1} = b_t - \tfrac{\eta}{\delta} \nabla \mathcal{L}_{t+1}(\boldsymbol{W}_t, b_t) \qquad \text{(SGD)}$$

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 $\boldsymbol{W}_{0,ij} \sim_{\text{i.i.d.}} \mathcal{N}(0, 1/d)$. The following theorem characterizes the resulting risk trajectory.

Theorem 1 *Let $\alpha, \beta > 0$ denote the source and capacity exponents defined in (1.2), and assume the teacher width $r \gg 1$. We consider the regime $2\alpha + 2\beta > 1$ and a mildly overparameterized student network with $r_s \gg \operatorname{polylog} d$. We consider step size η and timescale parameter δ such that $\eta \asymp o(1/d)$ and $\delta \in o(1/\log d)$. Almost surely, we have*

$$\mathcal{R}(\eta t \log d) - \tfrac{1}{2}\mathbb{E}[\epsilon^2] \asymp \Theta\big((\eta t)^{-\frac{2(\alpha+\beta)-1}{\alpha+2\beta}} + r_s^{-2(\alpha+\beta)+1}\big) \tag{3.1}$$

wherever the limit of ηt is well-defined.

Remark 2 *We highlight the following:*

- *This asymptotic risk exhibits a decomposition similar to the neural scaling laws in [11, 12], where the error cleanly separates into optimization and approximation terms. Here, the capacity exponent β enters both contributions additively due to the alignment structure. While the limit $\beta \to 0$ recovers the known isotropic scaling behavior [3], our result establishes strictly faster risk decay for any level of anisotropy ($\beta > 0$).*
- *However, the resulting sample complexity satisfies $T \simeq d \operatorname{polylog}(d)$. Because the algorithm operates in a strict one-pass regime, runtime and sample complexity are identical. This rate is highly suboptimal: the intrinsic statistical complexity is determined by the effective dimension $r_{\mathit{eff}}(\mathbf{\Sigma}) := \operatorname{tr}(\mathbf{\Sigma})/\|\mathbf{\Sigma}\|_2$, which scales as $d^{1-\beta}$ for $\beta \in (0,1)$ and as $\mathcal{O}(1)$ for $\beta > 1$.*

This suboptimal rate suggests a potential algorithmic gap across all anisotropic regimes ($\beta > 0$). Does the $\tilde{\Theta}(d)$ sample requirement reflect a fundamental limitation of the dynamics in (SGD), or is it merely an artifact of the theoretical analysis? To provide evidence that vanilla SGD inherently requires $\Omega(d)$ sample complexity, we analyze an offline empirical estimator for the teacher directions:

$$\hat{R}(\boldsymbol{W}) = \Big\| \frac{1}{r_s} \mathbf{\Sigma}^{\frac{1}{2}} \boldsymbol{W}\boldsymbol{W}^\top \mathbf{\Sigma}^{\frac{1}{2}} - \hat{\boldsymbol{S}} \Big\|_{\mathrm{F}}^2, \quad \text{where} \quad \hat{\boldsymbol{S}} = \frac{1}{T}\sum_{t=1}^{T} y_t \boldsymbol{x}_t \boldsymbol{x}_t^\top. \tag{3.2}$$

This objective coincides with the population risk in (2.1), up to the bias and the noise variance terms. The key difference is that the population covariance is replaced by its unbiased empirical estimator: $\mathbb{E}[\hat{\boldsymbol{S}}] = \mathbf{\Sigma}^{\frac{1}{2}} \mathbf{\Theta} \mathbf{\Lambda} \mathbf{\Theta}^\top \mathbf{\Sigma}^{\frac{1}{2}}$. We observe that (SGD) is the online counterpart of minimizing this empirical objective. Since their sample complexities are known to match in the isotropic case ($\mathbf{\Sigma} = \boldsymbol{I}_d$), analyzing $\hat{R}(\boldsymbol{W})$ provides a reliable proxy for the online dynamics [4]. In the following proposition, we show that minimizing $\hat{R}(\boldsymbol{W})$ requires $\Omega(d)$ samples to achieve a non-vanishing distance with the true teacher directions. Consequently, when $T \ll d$, the estimator fundamentally fails to recover the relevant signal. This limitation stems from the $\mathbf{\Sigma}^{1/2}$ factors in (3.2), which implicitly offset the statistical advantages induced by anisotropy in $\hat{\boldsymbol{S}}$ and forces the optimal weights $\hat{\boldsymbol{W}}_T$ to align with the top eigenspace of the whitened matrix $\mathbf{\Sigma}^{-1/2} \hat{\boldsymbol{S}} \mathbf{\Sigma}^{-1/2}$.

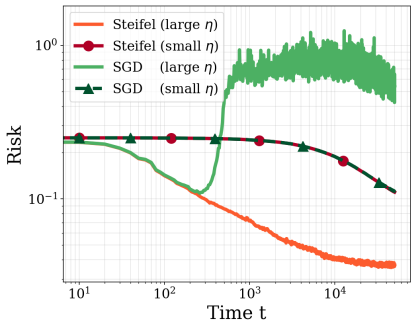


Figure 2: Population risk vs training iteration of (SGD) and Stiefel SGD. Risks are plotted after the alignment step. We set $d = 16384$, $r = 2048$, $r_s = 512$, and $\alpha = \beta = 0.35$. Small η scales as $\tilde{\mathcal{O}}(1/d)$, while large η scales as $\tilde{\mathcal{O}}(1/r_{\mathrm{eff}}(\mathbf{\Sigma}))$.

Proposition 3 (Offline Estimator Lower Bound) *Let $\hat{\boldsymbol{W}}_T$ be the minimizer of $\hat{R}(\boldsymbol{W})$ in (3.2). If $T \ll d$, then for a sufficiently small constant $\varepsilon > 0$ and d sufficiently large,*

$$\mathbb{P}\left[\left\|\frac{1}{r_s}\boldsymbol{\Sigma}^{\frac{1}{2}}\hat{\boldsymbol{W}}_T\hat{\boldsymbol{W}}_T^\top\boldsymbol{\Sigma}^{\frac{1}{2}} - \boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\Theta}\boldsymbol{\Lambda}\boldsymbol{\Theta}^\top\boldsymbol{\Sigma}^{\frac{1}{2}}\right\|_F > \varepsilon\right] \geq 1 - o_d(1).$$

By showing that the population risk remains lower-bounded when $T \ll d$, Proposition 3 provides evidence that the sample complexity requirement in Theorem 1 cannot be substantially improved. Beyond this theoretical argument, we provide empirical observations supporting this conclusion. As illustrated in Figure 2, (SGD) becomes unstable for learning rates $\eta \asymp \tilde{\mathcal{O}}(1/r_{\text{eff}}(\boldsymbol{\Sigma}))$, whereas the normalized Stiefel (which is introduced in the subsequent section) remain stable in this regime. Because learning rates of this magnitude are required to achieve the statistically optimal sample complexity in one-pass settings, this instability provides an alternative argument that (SGD) is fundamentally restricted to conservative step sizes.

3.2. Improved sample complexity via normalized dynamics

In this section, we introduce normalized Stiefel dynamics to achieve near-optimal sample complexity under anisotropic data. The fundamental issue with the standard quadratic objective (3.2) is that it implicitly counteracts the data's anisotropic structure. Because the network variance scales as $\mathbb{E}[\hat{y}^2] \propto \|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{W}\boldsymbol{W}^\top\boldsymbol{\Sigma}^{\frac{1}{2}}\|_{\text{F}}^2$, unconstrained dynamics are naturally biased toward whitening the covariance spectrum, forcing the weights to align with the inverse covariance geometry.

To bypass this whitening effect and preserve the statistical advantages of anisotropy, we optimize a correlation objective while constraining the weight matrix to the Stiefel manifold $\text{St}(r_s, d) := \{\mathbf{V} \in \mathbb{R}^{d\times r_s} \mid \mathbf{V}^\top\mathbf{V} = \boldsymbol{I}_{r_s}\}$, which makes sure that the weights are unit norm and orthogonal. Given a fresh sample $(\boldsymbol{x}_{t+1}, y_{t+1})$, we define this instantaneous objective and its corresponding Riemannian gradient as:

$$\mathcal{J}_{t+1}(\boldsymbol{W}, b) = -y_{t+1}\hat{y}(\boldsymbol{x}_{t+1}; \boldsymbol{W}, b), \qquad \nabla_{\text{St}}\mathcal{J}_{t+1}(\boldsymbol{W}, b) = (\boldsymbol{I}_d - \boldsymbol{W}\boldsymbol{W}^\top)y_{t+1}\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}^\top\boldsymbol{W}.$$

Following layer-wise training strategies from prior work [1, 6, 15], the training proceeds in two distinct stages:

- Phase 1 (Direction Learning): We initialize $\boldsymbol{W}_0 \sim \text{Unif}(\text{St}(r_s, d))$ and $b_0 = 0$. For $t = 1, 2, \ldots,$ we update the weights using:

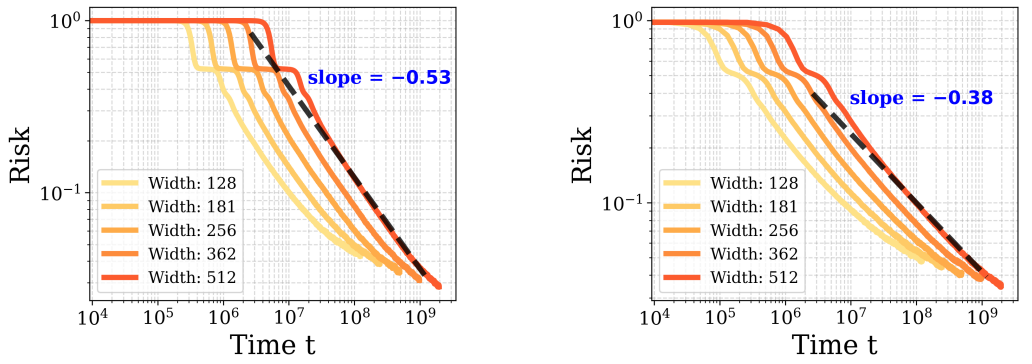
$$\widetilde{\boldsymbol{W}}_{t+1} = \boldsymbol{W}_t - \eta\nabla_{\text{St}}\mathcal{J}_{t+1}(\boldsymbol{W}_t, b_0), \qquad \boldsymbol{W}_{t+1} = \widetilde{\boldsymbol{W}}_{t+1}(\widetilde{\boldsymbol{W}}_{t+1}^\top\widetilde{\boldsymbol{W}}_{t+1})^{-1/2}.$$

- Phase 2 (Weight Alignment): Using a set of T_{align} fresh samples $\{(\tilde{\boldsymbol{x}}_j, \tilde{y}_j)\}$, we freeze the learned features $\boldsymbol{W}_t$ and optimize an alignment matrix $\boldsymbol{\Omega}$ and scalar bias b:

$$(\boldsymbol{\Omega}_*, b_*) = \text{argmin}_{\boldsymbol{\Omega}\in\mathbb{R}^{r_s\times r_s},\, b\in\mathbb{R}} \sum_{j=1}^{T_{\text{align}}} \left(\tilde{y}_j - \hat{y}(\tilde{\boldsymbol{x}}_j; \boldsymbol{W}_t\boldsymbol{\Omega}, b)\right)^2 + \lambda\|\boldsymbol{\Omega}\|_{\text{F}}^2.$$

The final parameters for the model are then set to $\boldsymbol{W}_t^{\text{final}} := \boldsymbol{W}_t\boldsymbol{\Omega}_*$ and $b_t^{\text{final}} := b_*$.

Let $\mathcal{R}(t) := R(\boldsymbol{W}_t^{\text{final}}, b_t^{\text{final}})$ denote the population risk of the final predictor. The following theorem establishes that these normalized dynamics achieve near-optimal sample complexity while matching the exact asymptotic scaling trajectory established in Theorem 1.



(a) $(\alpha, \beta) = (0.6, 0.1)$. ideal exponent: -0.5 (b) $(\alpha, \beta) = (0.35, 0.35)$. ideal exponent: -0.33

Figure 3: Population loss versus compute for a two-layer squared ReLU network. The model is trained using (SGD) with the learning rate regime in Theorem 1. We use $d = 2048$ and $r = 1024$. The resulting risk trajectories follow a power-law decay, where the empirical exponents (dashed lines) closely align with our theoretical prediction of $\frac{1-2(\alpha+\beta)}{\alpha+2\beta}$ for the quadratic setting.

Theorem 4 *Consider the teacher model with parameters α, β, and r as defined in Theorem 1. Let the hyperparameters satisfy: $\eta \ll 1/\tilde{\mathcal{O}}(r_{\mathit{eff}}(\mathbf{\Sigma}))$, and $r_s \gg d/\mathrm{polylog}(d)$, and $T_{\mathit{align}} \gg \mathrm{polylog}(d)$. Then there exists a regularization parameter $\lambda > 0$, such that the risk of the aligned predictor $\mathcal{R}$ satisfies the same guarantee as in (3.1).*

We conclude by emphasizing a critical distinction: unlike (SGD), Stiefel dynamics remain stable at significantly larger learning rates. Consequently, the required sample complexity improves to $\tilde{\mathcal{O}}(r_{\text{eff}}(\mathbf{\Sigma}))$, which matches the effective dimension of the input distribution and is information-theoretically optimal up to logarithmic factors.

4. Conclusion

In this work, we characterized the feature learning dynamics of SGD under anisotropic data and specific spectral alignment conditions. We demonstrated that vanilla SGD requires a sample complexity of $n \asymp T \simeq d\,\mathrm{polylog}(d)$, which is intrinsically suboptimal relative to the effective dimension $r_{\text{eff}}(\mathbf{\Sigma})$ due to instabilities in the empirical objective. To resolve this fundamental $\Omega(d)$ bottleneck, we introduced a normalized algorithm that optimizes a correlation objective under Stiefel constraints. This approach achieves statistical optimality by maintaining stability at larger learning rates ($\eta \asymp 1/r_{\text{eff}}(\mathbf{\Sigma})$). Future directions include relaxing the spectral alignment assumption and more rigorously formalizing the instability regime of vanilla SGD under anisotropic data.

Universality of exponents. A natural next step is extending this theoretical framework beyond two-layer quadratic networks to general activation functions. Because SGD is conjectured to learn in hierarchical phases dictated by Hermite polynomials, standard nonlinearities (e.g., GeLU, $\tanh$) with a nonzero second-order (He_2) component should exhibit an intermediate phase governed by the exact scaling exponents derived in our analysis. Preliminary experiments with squared ReLU activations confirm the existence of this intermediate phase (Figures 3a and 3b), and we leave the formal generalization of this framework as exciting future work.

References

[1] 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.

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

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

[4] Gérard Ben Arous, Cédric Gerbelot, and Vanessa Piccolo. Stochastic gradient descent in high dimensions for multi-spiked tensor pca, 2025. URL https://arxiv.org/abs/2410.18162.

[5] Andrea Caponnetto and Ernesto de Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7:331–368, 2007. URL https://api.semanticscholar.org/CorpusID:207063850.

[6] Alexandru Damian, Jason Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent. In *Conference on Learning Theory*, pages 5413–5452. PMLR, 2022.

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

[8] 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.

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

[10] 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.

[11] 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.

[12] 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.

[13] Juno Kim, Eshaan Nichani, Denny Wu, Alberto Bietti, and Jason D. Lee. Sharp capacity scaling of spectral optimizers in learning associative memory, 2026. URL https://arxiv.org/abs/2603.26554.

[14] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000. ISSN 00905364, 21688966. URL http://www.jstor.org/stable/2674095.

[15] Jason D Lee, Kazusato Oko, Taiji Suzuki, and Denny Wu. Neural network learns low-dimensional polynomials with sgd near the information-theoretic limit. *arXiv preprint arXiv:2406.01581*, 2024.

[16] 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.

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

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

[19] Marco Mondelli and Andrea Montanari. Fundamental limits of weak recovery with applications to phase retrieval. *Foundations of Computational Mathematics*, 19:703–773, 2017.

[20] 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.

[21] 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.

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

[23] 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.

[24] Arie Wortsman and Bruno Loureiro. Kernel ridge regression under power-law data: spectrum and generalization. *ArXiv*, abs/2510.04780, 2025. URL https://api.semanticscholar.org/CorpusID:281842750.

Contents

Appendix A. Preliminaries for Proofs

Proof organization. Section A introduces the additional notation and definitions used throughout, and collects several auxiliary lemma. Section B proves the discrete-time online SGD result in Theorem 1. Section C proves our lower bound argument in Proposition 3. Finally, Section D proves the online Steifel SGD result in Theorem 4.

Additional notation and definitions. We use a reduced time parameter $\bar{t} > 0$ and relate it to the physical time scales by

$$T = \bar{t}\, T_{\text{eff}}, \qquad T_{\text{eff}} := \frac{Z_r}{2} \log \frac{d}{Z_r}, \qquad N = \lfloor \bar{t}\, N_{\text{eff}} \rfloor, \qquad N_{\text{eff}} := \frac{Z_r}{2\eta} \log \frac{d}{Z_r}.$$

We work in the common eigenbasis of the covariance and the teacher:

$$\boldsymbol{\Sigma} = \operatorname{diag}(\sigma_1, \dots, \sigma_d), \qquad \boldsymbol{T} = \operatorname{diag}(\lambda_1, \dots, \lambda_r, 0, \dots, 0),$$

where

$$\lambda_1 \geq \cdots \geq \lambda_r > 0, \qquad \sigma_1 \geq \cdots \geq \sigma_d > 0, \qquad \lambda_i \asymp i^{-\alpha}, \qquad \sigma_i \asymp i^{-\beta}. \tag{A.1}$$

We assume

$$2\alpha + 2\beta > 1, \qquad \text{equivalently} \qquad \alpha + \beta > \frac{1}{2}.$$

In particular, under the usual fixed spectral envelope in (A.1), one has $\sigma_1 = O(1)$.

Define

$$Z_r := \|\boldsymbol{T}\boldsymbol{\Sigma}\|_F = \left(\sum_{i=1}^{r} \sigma_i^2 \lambda_i^2 \right)^{1/2}, \qquad \frac{d}{Z_r} \to \infty.$$

Set

$$\boldsymbol{S}^\star := \frac{1}{Z_r} \boldsymbol{\Sigma}^{1/2} \boldsymbol{T} \boldsymbol{\Sigma}^{1/2}, \qquad \boldsymbol{T}_\star := \boldsymbol{\Sigma}\boldsymbol{S}^\star = \boldsymbol{S}^\star \boldsymbol{\Sigma} = \frac{1}{Z_r} \boldsymbol{\Sigma} \boldsymbol{T} \boldsymbol{\Sigma}.$$

Then

$$\boldsymbol{T}_\star = \operatorname{diag}(\tau_1, \dots, \tau_r, 0, \dots, 0), \qquad \tau_i = \frac{\sigma_i^2 \lambda_i}{Z_r}, \qquad \kappa_i := \sigma_i^2 \lambda_i, \qquad s_i^\star := \frac{\tau_i}{\sigma_i} = \frac{\sigma_i \lambda_i}{Z_r}.$$

The normalization is

$$\|\boldsymbol{S}^\star\|_F^2 = \sum_{i=1}^{r} (s_i^\star)^2 = 1.$$

And

$$\|\boldsymbol{T}_\star\|_F^2 = \frac{\sum_{i=1}^{r} \sigma_i^4 \lambda_i^2}{Z_r^2} \leq \sigma_1^2 \|\boldsymbol{S}^\star\|_F^2 = O(1). \tag{A.2}$$

Fix $\bar{t} > 0$ and define

$$m_\star(\bar{t}) := \max\{ i \leq r : \bar{t}\, \kappa_i > 1 \},$$

with the convention that the maximum of the empty set is 0. We write

$$k := m_\star(\bar{t}).$$

If $k < r$, set

$$a_{\mathrm{s}} := \bar{t}\,\kappa_{k+1} \in [0, 1),$$

and if $k = r$, set $a_{\mathrm{s}} := 0$.

Fix

$$0 < \rho < \min\{\beta, 1\}.$$

When $k < r$ and the slow block is nonempty, we additionally assume

$$\rho < a_{\mathrm{s}}.$$

Define

$$m_\rho(\bar{t}) := \max\{i \le r : \bar{t}\,\kappa_i > \rho\},$$

again with the convention that the maximum of the empty set is 0.

We assume an off-transition gap: there exists $\delta_{\mathrm{gap}} \in (0, 1)$, independent of d, such that

$$\bar{t}\,\kappa_i \notin [1 - \delta_{\mathrm{gap}}, 1 + \delta_{\mathrm{gap}}] \qquad \text{for all } i \le m_\rho(\bar{t}).$$

Consequently,

$$k = \max\{i \le m_\rho(\bar{t}) : \bar{t}\,\kappa_i \ge 1 + \delta_{\mathrm{gap}}\},$$

and, if $k < r$,

$$a_{\mathrm{s}} \le 1 - \delta_{\mathrm{gap}}.$$

We use the diagonal projectors

$$\boldsymbol{P}_{\mathrm{f}} := \mathrm{diag}(\mathbb{1}\{i \le k\})_{i=1}^{d}, \qquad \boldsymbol{P}_{\mathrm{r}} := \boldsymbol{I}_d - \boldsymbol{P}_{\mathrm{f}}.$$

Inside the rest block, we further write

$$\boldsymbol{P}_{\mathrm{s}} := \mathrm{diag}(\mathbb{1}\{k < i \le m_\rho(\bar{t})\})_{i=1}^{d}, \qquad \boldsymbol{P}_{\mathrm{t}} := \boldsymbol{I}_d - \boldsymbol{P}_{\mathrm{f}} - \boldsymbol{P}_{\mathrm{s}}.$$

Since $\kappa_i \asymp i^{-(\alpha+2\beta)}$, for every fixed $\bar{t}$ and ρ,

$$k = O_{\bar{t}}(1), \qquad m_\rho(\bar{t}) = O_{\bar{t},\rho}(1).$$

Unless stated otherwise, constants may depend on the fixed spectral regularity constants, on $\bar{t}$, ρ, δ_{gap}, K, and σ_ε, but not on d, N, η, δ, r_s.

Lemma 5 (Rank floor) *For every matrix $\boldsymbol{M}$ with $r_u(\boldsymbol{M}) \le r_s$,*

$$\|\boldsymbol{S}^\star - \boldsymbol{M}\|_F^2 \ge \sum_{i > r_s} (s_i^\star)^2. \tag{A.3}$$

If $\alpha + \beta > 1/2$, then, in the regime $r_s \to \infty$ with $r_s < r$,

$$\sum_{i > r_s} (s_i^\star)^2 \asymp r_s^{-2(\alpha+\beta)+1}. \tag{A.4}$$

Proof The nonzero singular values of $\boldsymbol{S}^\star$ are precisely $(s_i^\star)_{i \le r}$, in nonincreasing order. The lower bound (A.3) is the Eckart–Young theorem. The tail estimate (A.4) follows from $(s_i^\star)^2 \asymp i^{-2(\alpha+\beta)}$ and the integral test. ■

Appendix B. SGD comparison proof

B.1. Setup, exact recursion, and centered ideal discrete flow

Let

$$\boldsymbol{x}_n \sim \mathcal{N}(0, \boldsymbol{\Sigma}) \quad \text{i.i.d.}, \qquad \varepsilon_n \perp\!\!\!\perp \boldsymbol{x}_n, \qquad \mathbb{E}[\varepsilon_n] = 0, \qquad \|\varepsilon_n\|_{\psi_2} \le \sigma_\varepsilon.$$

Set

$$\boldsymbol{g}_n := \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}_n \sim \mathcal{N}(0, \boldsymbol{I}_d), \qquad \boldsymbol{\chi}_n := \boldsymbol{g}_n\boldsymbol{g}_n^\top, \qquad \boldsymbol{\xi}_n := \boldsymbol{\chi}_n - \boldsymbol{I}_d.$$

The label is

$$y_n = \left\langle \frac{\boldsymbol{T}}{Z_r}, \boldsymbol{x}_n\boldsymbol{x}_n^\top - \boldsymbol{\Sigma} \right\rangle + \varepsilon_n = \langle \boldsymbol{S}^\star, \boldsymbol{\xi}_n \rangle + \varepsilon_n.$$

Define

$$\boldsymbol{G}_n := \frac{1}{r_s}\boldsymbol{W}_n\boldsymbol{W}_n^\top, \qquad \boldsymbol{M}_n := \boldsymbol{\Sigma}^{1/2}\boldsymbol{G}_n\boldsymbol{\Sigma}^{1/2}, \qquad e_n := b_n - \mathrm{tr}(\boldsymbol{M}_n), \qquad \boldsymbol{\Delta}_n := \boldsymbol{S}^\star - \boldsymbol{M}_n.$$

The online prediction is

$$\widehat{y}_{n+1} = \langle \boldsymbol{M}_n, \boldsymbol{\chi}_{n+1} \rangle - b_n.$$

To avoid conflict with the two-timescale parameter δ, the prediction error is denoted by

$$\mathfrak{r}_{n+1} := y_{n+1} - \widehat{y}_{n+1} = \langle \boldsymbol{\Delta}_n, \boldsymbol{\xi}_{n+1} \rangle + e_n + \varepsilon_{n+1}. \tag{B.1}$$

The online SGD updates are

$$\boldsymbol{W}_{n+1} = \boldsymbol{W}_n + \frac{\eta}{2}\mathfrak{r}_{n+1}\boldsymbol{x}_{n+1}\boldsymbol{x}_{n+1}^\top\boldsymbol{W}_n, \tag{B.2}$$

$$b_{n+1} = b_n - \frac{\eta}{4\delta}\mathfrak{r}_{n+1}. \tag{B.3}$$

Let

$$N_{\mathrm{eff}} := \frac{Z_r}{2\eta}\log\frac{d}{Z_r}, \qquad N := \lfloor \bar{t}\, N_{\mathrm{eff}} \rfloor, \qquad \Lambda_d := \log\frac{ed}{Z_r}.$$

Let

$$\mathscr{F}_n := \sigma(\boldsymbol{W}_0, b_0, (\boldsymbol{x}_t, \varepsilon_t)_{1\le t\le n})\,.$$

We also set

$$p_d := 40(K+2)\log(edN), \qquad R_s := 1 + \sqrt{r_s}, \qquad L_2 := \mathrm{tr}(\boldsymbol{\Sigma}^2).$$

Proposition 6 (Exact sandwich recursion and first-order expansion) *For every $n \ge 0$, $\boldsymbol{M}_n \succeq 0$ and*

$$\boldsymbol{M}_{n+1} = \left(\boldsymbol{I}_d + \frac{\eta}{2}\mathfrak{r}_{n+1}\boldsymbol{\Sigma}\boldsymbol{\chi}_{n+1}\right)\boldsymbol{M}_n\left(\boldsymbol{I}_d + \frac{\eta}{2}\mathfrak{r}_{n+1}\boldsymbol{\chi}_{n+1}\boldsymbol{\Sigma}\right). \tag{B.4}$$

Equivalently,

$$\boldsymbol{M}_{n+1} = \boldsymbol{M}_n + \eta\mathcal{F}^\sharp(\boldsymbol{M}_n, e_n) + \eta\boldsymbol{Z}_{n+1} + \eta^2\boldsymbol{Q}_{n+1}, \tag{B.5}$$

where

$$\mathcal{F}^\sharp(\boldsymbol{M}, e) := \mathcal{F}(\boldsymbol{M}) + e\mathcal{B}(\boldsymbol{M}),$$

and

$$\boldsymbol{Z}_{n+1} := \frac{1}{2}\mathfrak{r}_{n+1}\left(\boldsymbol{\Sigma}\boldsymbol{\chi}_{n+1}\boldsymbol{M}_n + \boldsymbol{M}_n\boldsymbol{\chi}_{n+1}\boldsymbol{\Sigma}\right) - \mathcal{F}^\sharp(\boldsymbol{M}_n, e_n),$$
$$\boldsymbol{Q}_{n+1} := \frac{1}{4}\mathfrak{r}_{n+1}^2\boldsymbol{\Sigma}\boldsymbol{\chi}_{n+1}\boldsymbol{M}_n\boldsymbol{\chi}_{n+1}\boldsymbol{\Sigma}.$$

Moreover,

$$\mathbb{E}[\boldsymbol{Z}_{n+1} \mid \mathscr{F}_n] = 0. \tag{B.6}$$

Proof Since $\boldsymbol{x}_{n+1}\boldsymbol{x}_{n+1}^\top = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\chi}_{n+1}\boldsymbol{\Sigma}^{1/2}$, (B.2) gives (B.4); PSD is therefore preserved. Expanding the sandwich gives (B.5).

It remains to identify the conditional drift. By (B.1),

$$\mathbb{E}[\mathfrak{r}_{n+1}\boldsymbol{\chi}_{n+1} \mid \mathscr{F}_n] = \mathbb{E}[\mathfrak{r}_{n+1}\boldsymbol{\xi}_{n+1} \mid \mathscr{F}_n] + e_n\boldsymbol{I}_d.$$

We have

$$\mathbb{E}[\langle \boldsymbol{A}, \boldsymbol{\xi}_{n+1}\rangle\boldsymbol{\xi}_{n+1}] = 2\boldsymbol{A} \qquad (\boldsymbol{A} = \boldsymbol{A}^\top).$$

Taking $\boldsymbol{A} = \boldsymbol{\Delta}_n$ yields

$$\mathbb{E}[\mathfrak{r}_{n+1}\boldsymbol{\chi}_{n+1} \mid \mathscr{F}_n] = 2\boldsymbol{\Delta}_n + e_n\boldsymbol{I}_d.$$

Substitution into the linear part of (B.4) gives exactly $\mathcal{F}^\sharp(\boldsymbol{M}_n, e_n)$, proving (B.6). ■

The centered ideal discrete flow is the diagonal recursion

$$\boldsymbol{L}_{n+1} = \boldsymbol{L}_n + \eta\left(\boldsymbol{T}_\star\boldsymbol{L}_n + \boldsymbol{L}_n\boldsymbol{T}_\star - 2\boldsymbol{L}_n\boldsymbol{\Sigma}\boldsymbol{L}_n\right) + \eta^2(\boldsymbol{T}_\star - \boldsymbol{\Sigma}\boldsymbol{L}_n)\boldsymbol{L}_n(\boldsymbol{T}_\star - \boldsymbol{L}_n\boldsymbol{\Sigma}), \qquad \boldsymbol{L}_0 = \frac{1}{d}\boldsymbol{\Sigma}.$$

Equivalently,

$$\ell_{i,n+1} = \ell_{i,n}\left(1 + \eta(\tau_i - \sigma_i\ell_{i,n})\right)^2, \qquad i \le r,$$
$$\ell_{i,n+1} = \ell_{i,n}\left(1 - \eta\sigma_i\ell_{i,n}\right)^2, \qquad i > r. \tag{B.7}$$

B.2. Scalar logistic facts

Lemma 7 (Basic scalar logistic facts) *Fix $c \in (0, 1/2]$, and let*

$$a_{n+1} = a_n(1 + c(1 - a_n))^2, \qquad 0 \le a_0 \le 1. \tag{B.8}$$

Then:

(i) $0 \le a_n \le 1$ *and* $a_{n+1} \ge a_n$ *for all* n.

(ii) $a_n \le a_0(1 + c)^{2n}$ *for all* n.

(iii) *If* $a_m \le \theta < 1$ *for all* $0 \le m \le n$, *then*

$$a_n \ge a_0(1 + c(1 - \theta))^{2n}.$$

(iv) *If* $a_n \ge \theta > 0$, *then*

$$1 - a_{n+1} \le (1 - 2c\theta)(1 - a_n). \tag{B.9}$$

Proof Let $\phi(a) := a(1 + c(1 - a))^2$. Then $\phi([0, 1]) \subseteq [0, 1]$ and

$$\phi(a) - a = ac(1 - a)(2 + c(1 - a)) \geq 0,$$

which proves (i). Item (ii) follows from $1+c(1-a_n) \leq 1+c$. Item (iii) follows from $1-a_m \geq 1-\theta$. Finally,

$$1 - a_{n+1} = 1 - a_n(1 + c(1 - a_n))^2 = (1 - a_n)\left(1 - 2ca_n - c^2a_n(1 - a_n)\right).$$

If $a_n \geq \theta$, the last factor is at most $1 - 2c\theta$, proving (B.9). ■

Lemma 8 (Robust discrete threshold law) *Fix $i \leq r$ and set*

$$c_i := \eta\tau_i, \qquad n(\bar{t}) := \lfloor \bar{t}\, N_{\text{eff}} \rfloor.$$

Assume

$$\eta\tau_1 \leq \frac{1}{2}, \qquad \eta \log^2 \frac{d}{Z_r} \to 0. \tag{B.10}$$

Let a_n satisfy (B.8) with $c = c_i$ and initial data

$$c_- \frac{Z_r}{d} \leq a_0 \leq c_+ \frac{Z_r}{d},$$

where $0 < c_- < c_+ < \infty$ are independent of d. For every fixed $\delta_{\text{thr}} \in (0, 1)$:
(i) *If $\bar{t}\,\kappa_i \leq 1 - \delta_{\text{thr}}$, then $a_{n(\bar{t})} \to 0$.*
(ii) *If $\bar{t}\,\kappa_i \geq 1 + \delta_{\text{thr}}$, then $a_{n(\bar{t})} \to 1$.*

Proof Write $n = n(\bar{t})$. If $\bar{t}\,\kappa_i \leq 1 - \delta_{\text{thr}}$, Lemma 7(ii) gives

$$a_n \leq c_+ \frac{Z_r}{d} \exp(2nc_i).$$

Since

$$2nc_i = \bar{t}\,\kappa_i \log \frac{d}{Z_r} + o(1),$$

we have

$$a_n \leq c_+ \left(\frac{d}{Z_r}\right)^{-1+\bar{t}\,\kappa_i+o(1)} \leq c_+ \left(\frac{d}{Z_r}\right)^{-\delta_{\text{thr}}+o(1)} \to 0.$$

Now assume $\bar{t}\,\kappa_i \geq 1 + \delta_{\text{thr}}$. Set

$$\varepsilon := \frac{\delta_{\text{thr}}}{4}, \qquad \gamma := 1 - \frac{\delta_{\text{thr}}}{2(1 + \delta_{\text{thr}})}, \qquad n_- := \lfloor \gamma\bar{t}\, N_{\text{eff}} \rfloor.$$

Then $(1 - \varepsilon)\gamma(1 + \delta_{\text{thr}}) > 1$. If $a_m \leq \varepsilon$ for all $m \leq n_-$, Lemma 7(iii) yields

$$a_{n_-} \geq a_0(1 + c_i(1 - \varepsilon))^{2n_-}.$$

Using $\log(1 + x) \geq x - x^2$ for $x \in [0, 1/2]$ and (B.10),

$$\log a_{n_-} \geq \log a_0 + ((1 - \varepsilon)\gamma\bar{t}\,\kappa_i + o(1)) \log \frac{d}{Z_r}.$$

The exponent is strictly larger than 1 by a fixed margin, contradicting $a_{n_-} \le 1$. Hence some $m \le n_-$ has $a_m \ge \varepsilon$, and monotonicity gives $a_{n_-} \ge \varepsilon$. Lemma 7(iv) implies

$$1 - a_n \le \exp(-2c_i\varepsilon(n - n_-)).$$

Finally,

$$2c_i(n - n_-) = (1-\gamma)\bar{t}\,\kappa_i \log \frac{d}{Z_r} + o(1) \to \infty,$$

so $a_n \to 1$. ∎

B.3. Initialization, localization, and one-step bounds

Lemma 9 (Initialization event for SGD) *Assume*

$$(\boldsymbol{W}_0)_{ia} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1/d), \qquad b_0 = 0.$$

Then there exists an event $\mathcal{G}_{\text{init}}$ such that

$$\mathbb{P}(\mathcal{G}_{\text{init}}) \ge 1 - Cd^{-K},$$

and on $\mathcal{G}_{\text{init}}$,

$$\|\boldsymbol{M}_0\|_F \le 2, \qquad |e_0| \le 2.$$

Moreover, for every deterministic diagonal projector $\boldsymbol{P}$,

$$\operatorname{tr}(\boldsymbol{P}\boldsymbol{M}_0) \le \frac{C}{d}\operatorname{tr}(\boldsymbol{P}\boldsymbol{\Sigma}). \tag{B.11}$$

If $k \ge 1$, then for every $1 \le i \le k$,

$$\left|(\boldsymbol{H}_0)_{ii} - \frac{\sigma_i}{d}\right| \le \frac{C\sigma_i}{d}\sqrt{\frac{p_d}{r_s}}, \qquad (\boldsymbol{H}_0)_{ii} \ge \frac{c}{d}, \tag{B.12}$$

and

$$\delta_0^{\text{coh}} := \max_{1\le i\ne j\le k} \frac{|(\boldsymbol{H}_0)_{ij}|}{\sqrt{(\boldsymbol{H}_0)_{ii}(\boldsymbol{H}_0)_{jj}}} \le C\sqrt{\frac{p_d}{r_s}}. \tag{B.13}$$

Define the localization stopping time

$$\tau_\Delta := \inf\{n \ge 0 : \|\boldsymbol{\Delta}_n\|_F > 4\},$$

and the one-step localized event

$$\mathcal{G}_{\text{loc},n} := \{\|\boldsymbol{\Delta}_n\|_F \le 4,\ |e_n| \le 4,\ \|\boldsymbol{M}_n\|_F \le 5\}.$$

Lemma 10 (Gaussian chaos bounds) *For every $p \ge 2$, the following hold.*

(i) *For every symmetric $\boldsymbol{A}$,*

$$\mathbb{E}[\langle \boldsymbol{A}, \boldsymbol{\xi}_n\rangle^2] = 2\|\boldsymbol{A}\|_F^2, \qquad \|\langle \boldsymbol{A}, \boldsymbol{\xi}_n\rangle\|_{L^p} \le Cp\|\boldsymbol{A}\|_F.$$

(ii) *For symmetric* $\boldsymbol{A}, \boldsymbol{B}$*,*

$$\|\langle \boldsymbol{A}, \boldsymbol{\xi}_n\rangle\langle \boldsymbol{B}, \boldsymbol{\xi}_n\rangle\|_{L^p} \le Cp^2\|\boldsymbol{A}\|_F\|\boldsymbol{B}\|_F.$$

(iii) *If* $\boldsymbol{A} \succeq 0$*, then*

$$\|\boldsymbol{g}_n^\top \boldsymbol{A}\boldsymbol{g}_n\|_{L^p} \le C\left(\operatorname{tr}(\boldsymbol{A}) + p\|\boldsymbol{A}\|_{\mathrm{op}}\right). \tag{B.14}$$

(iv) *If* $\boldsymbol{M} \succeq 0$*,* $r_u(\boldsymbol{M}) \le r_s$*, and* $\|\boldsymbol{M}\|_F \le 5$*, then for every diagonal projector* $\boldsymbol{P}$*,*

$$\|\operatorname{tr}(\boldsymbol{P}\boldsymbol{\Sigma}\boldsymbol{\chi}_n\boldsymbol{M}\boldsymbol{\chi}_n\boldsymbol{\Sigma})\|_{L^p} \le Cp^2R_s\left(1 + \operatorname{tr}(\boldsymbol{P}\boldsymbol{\Sigma}^2)\right). \tag{B.15}$$

Proof Items (i) and (ii) are standard hypercontractive estimates for Gaussian chaoses of order two and four. Item (iii) follows by diagonalization and moment bounds for weighted sums of centered χ^2 variables. For (iv),

$$\operatorname{tr}(\boldsymbol{P}\boldsymbol{\Sigma}\boldsymbol{\chi}_n\boldsymbol{M}\boldsymbol{\chi}_n\boldsymbol{\Sigma}) = (\boldsymbol{g}_n^\top\boldsymbol{M}\boldsymbol{g}_n)(\boldsymbol{g}_n^\top\boldsymbol{P}\boldsymbol{\Sigma}^2\boldsymbol{P}\boldsymbol{g}_n).$$

Apply Hölder and (B.14). Since $\boldsymbol{M} \succeq 0$,

$$\operatorname{tr}(\boldsymbol{M}) \le \sqrt{r_u(\boldsymbol{M})}\|\boldsymbol{M}\|_F \le CR_s.$$

This proves (B.15). ∎

Lemma 11 (Martingale maximal and stable convolution bounds) *Let* $(X_n)_{n\ge1}$ *be a martingale difference sequence with respect to* $(\mathscr{G}_n)_{n\ge0}$*. Then, for* $p \ge 2$*,*

$$\left\|\sup_{1\le m\le N}\left|\sum_{n=1}^m X_n\right|\right\|_{L^p} \le C\left[\sqrt{p}\left\|\left(\sum_{n=1}^N \mathbb{E}[X_n^2 \mid \mathscr{G}_{n-1}]\right)^{1/2}\right\|_{L^p} + p\left(\sum_{n=1}^N \|X_n\|_{L^p}^p\right)^{1/p}\right]. \tag{B.16}$$

If $0 < \gamma < 1$ *and*

$$S_n := \sum_{m=0}^{n-1}\gamma^{n-1-m}X_{m+1},$$

with

$$\mathbb{E}[X_{m+1}^2 \mid \mathscr{F}_m] \le v^2, \qquad \|X_{m+1}\|_{L^{p_d}(\mathbb{P}(\cdot\mid\mathscr{F}_m))} \le A_{p_d},$$

then, with probability at least $1 - Cd^{-K}$*,*

$$\sup_{0\le n\le N}|S_n| \le C\left(\sqrt{\frac{p_dv^2}{1-\gamma}} + p_dA_{p_d}\right).$$

Proof The first estimate is the Burkholder–Freedman maximal inequality in L^p. For the stable convolution, apply (B.16) to the weighted martingale for fixed n, using

$$\sum_{j\ge0}\gamma^{2j} \le \frac{1}{1-\gamma}, \qquad \left(\sum_{j\ge0}\gamma^{p_dj}\right)^{1/p_d} \le C.$$

A union bound over $0 \le n \le N$ is absorbed into the definition of p_d. ∎

For a diagonal projector $\boldsymbol{P}$, define

$$u_n^{(\boldsymbol{P})} := \operatorname{tr}(\boldsymbol{P}\boldsymbol{M}_n), \qquad \tau(\boldsymbol{P}) := \max_{i:P_{ii}=1} \tau_i, \qquad \sigma(\boldsymbol{P}) := \max_{i:P_{ii}=1} \sigma_i,$$

with empty maxima equal to 0. Also set

$$d_{n+1}^{(\boldsymbol{P})} := \operatorname{tr}(\boldsymbol{P}\boldsymbol{Z}_{n+1}), \qquad q_{n+1}^{(\boldsymbol{P})} := \operatorname{tr}(\boldsymbol{P}\boldsymbol{Q}_{n+1}),$$

and

$$\overline{q}_n^{(\boldsymbol{P})} := \mathbb{E}[q_{n+1}^{(\boldsymbol{P})} \mid \mathscr{F}_n], \qquad \widehat{q}_{n+1}^{(\boldsymbol{P})} := q_{n+1}^{(\boldsymbol{P})} - \overline{q}_n^{(\boldsymbol{P})}.$$

Lemma 12 (Localized one-step bounds) *On $\mathcal{G}_{\mathrm{loc},n}$, the following estimates hold for every $p \geq 2$.*

(i) *For every diagonal projector $\boldsymbol{P}$,*

$$\mathbb{E}[d_{n+1}^{(\boldsymbol{P})} \mid \mathscr{F}_n] = 0, \qquad \mathbb{E}[\widehat{q}_{n+1}^{(\boldsymbol{P})} \mid \mathscr{F}_n] = 0,$$

and

$$\|d_{n+1}^{(\boldsymbol{P})}\|_{L^p(\mathbb{P}(\cdot|\mathscr{F}_n))} \leq Cp^2 \left(1 + \operatorname{tr}(\boldsymbol{P}\boldsymbol{\Sigma}^2)\right)^{1/2}, \tag{B.17}$$

$$\overline{q}_n^{(\boldsymbol{P})} \leq CR_s \left(1 + \operatorname{tr}(\boldsymbol{P}\boldsymbol{\Sigma}^2)\right), \tag{B.18}$$

$$\|\widehat{q}_{n+1}^{(\boldsymbol{P})}\|_{L^p(\mathbb{P}(\cdot|\mathscr{F}_n))} \leq CR_s p^4 \left(1 + \operatorname{tr}(\boldsymbol{P}\boldsymbol{\Sigma}^2)\right). \tag{B.19}$$

(ii) *If $\boldsymbol{P}$ has bounded support contained in $\{1, \dots, r\}$, then*

$$\|d_{n+1}^{(\boldsymbol{P})}\|_{L^p(\mathbb{P}(\cdot|\mathscr{F}_n))} \leq Cp^2 \left(u_n^{(\boldsymbol{P})}\right)^{1/2}, \tag{B.20}$$

$$\overline{q}_n^{(\boldsymbol{P})} \leq CR_s, \qquad \|\widehat{q}_{n+1}^{(\boldsymbol{P})}\|_{L^p(\mathbb{P}(\cdot|\mathscr{F}_n))} \leq CR_s p^4.$$

(iii) *Define*

$$u_{n+1} := \operatorname{tr}(\boldsymbol{\Sigma}\boldsymbol{\chi}_{n+1}\boldsymbol{M}_n), \qquad s_n := \operatorname{tr}(\boldsymbol{\Sigma}\boldsymbol{M}_n),$$

and

$$\zeta_{n+1}^e := -\frac{1}{4\delta}(\mathfrak{r}_{n+1} - e_n) - (\mathfrak{r}_{n+1}u_{n+1} - \mathbb{E}[\mathfrak{r}_{n+1}u_{n+1} \mid \mathscr{F}_n]). \tag{B.21}$$

Then

$$\mathbb{E}[\zeta_{n+1}^e \mid \mathscr{F}_n] = 0, \qquad \|\zeta_{n+1}^e\|_{L^p(\mathbb{P}(\cdot|\mathscr{F}_n))} \leq C\left(\frac{p}{\delta} + p^2\right). \tag{B.22}$$

For

$$q_{n+1}^e := q_{n+1}^{(\boldsymbol{I}_d)} = \operatorname{tr}(\boldsymbol{Q}_{n+1}),$$

one has

$$\mathbb{E}[q_{n+1}^e \mid \mathscr{F}_n] \leq CR_s(1+L_2), \qquad \|q_{n+1}^e - \mathbb{E}[q_{n+1}^e \mid \mathscr{F}_n]\|_{L^p(\mathbb{P}(\cdot|\mathscr{F}_n))} \leq CR_s p^4(1+L_2). \tag{B.23}$$

Proof For (i), write

$$d_{n+1}^{(\boldsymbol{P})} = \mathfrak{r}_{n+1}\langle \boldsymbol{K}_n^{(\boldsymbol{P})}, \boldsymbol{\chi}_{n+1}\rangle - \mathbb{E}[\mathfrak{r}_{n+1}\langle \boldsymbol{K}_n^{(\boldsymbol{P})}, \boldsymbol{\chi}_{n+1}\rangle \mid \mathscr{F}_n],$$

where

$$\boldsymbol{K}_n^{(\boldsymbol{P})} := \frac{1}{2}(\boldsymbol{M}_n\boldsymbol{P}\boldsymbol{\Sigma} + \boldsymbol{\Sigma}\boldsymbol{P}\boldsymbol{M}_n).$$

On $\mathcal{G}_{\mathrm{loc},n}$,

$$\|\boldsymbol{K}_n^{(\boldsymbol{P})}\|_F \le C\left(1 + \operatorname{tr}(\boldsymbol{P}\boldsymbol{\Sigma}^2)\right)^{1/2}.$$

Using (B.1), $\|\boldsymbol{\Delta}_n\|_F \le 4$, $|e_n| \le 4$, and Lemma 10, we get (B.17). The second-order term satisfies

$$q_{n+1}^{(\boldsymbol{P})} = \frac{1}{4}\mathfrak{r}_{n+1}^2 \operatorname{tr}(\boldsymbol{P}\boldsymbol{\Sigma}\boldsymbol{\chi}_{n+1}\boldsymbol{M}_n\boldsymbol{\chi}_{n+1}\boldsymbol{\Sigma}),$$

so (B.18) and (B.19) follow from conditional Hölder and (B.15).

If $\boldsymbol{P}$ has bounded support in $\{1, \dots, r\}$, PSD and bounded support give

$$\|\boldsymbol{K}_n^{(\boldsymbol{P})}\|_F \le C\left(u_n^{(\boldsymbol{P})}\right)^{1/2},$$

which proves (ii).

For (iii), note that

$$u_{n+1} = s_n + \left\langle \frac{1}{2}(\boldsymbol{\Sigma}\boldsymbol{M}_n + \boldsymbol{M}_n\boldsymbol{\Sigma}), \boldsymbol{\xi}_{n+1} \right\rangle.$$

Lemma 10 gives

$$\|\mathfrak{r}_{n+1} - e_n\|_{L^p(\mathbb{P}(\cdot|\mathscr{F}_n))} \le Cp$$

and

$$\|\mathfrak{r}_{n+1}u_{n+1} - \mathbb{E}[\mathfrak{r}_{n+1}u_{n+1} \mid \mathscr{F}_n]\|_{L^p(\mathbb{P}(\cdot|\mathscr{F}_n))} \le Cp^2.$$

This proves (B.22). The estimates for q_{n+1}^e are (B.18)–(B.19) with $\boldsymbol{P} = \boldsymbol{I}_d$. ■

B.4. Discrete bias tracking

Define

$$a_n^e := 2\operatorname{tr}(\boldsymbol{\Sigma}\boldsymbol{M}_n^2) - 2\operatorname{tr}(\boldsymbol{T}_\star\boldsymbol{M}_n).$$

Lemma 13 (Discrete bias recursion) *One has*

$$e_{n+1} = \left(1 - \frac{\eta}{4\delta} - \eta s_n\right) e_n + \eta a_n^e + \eta\zeta_{n+1}^e - \eta^2 q_{n+1}^e. \tag{B.24}$$

Proof Taking traces in (B.5),

$$\operatorname{tr}(\boldsymbol{M}_{n+1}) = \operatorname{tr}(\boldsymbol{M}_n) + \eta\mathfrak{r}_{n+1}u_{n+1} + \eta^2 q_{n+1}^e.$$

Together with (B.3), this gives

$$e_{n+1} = e_n - \frac{\eta}{4\delta}\mathfrak{r}_{n+1} - \eta\mathfrak{r}_{n+1}u_{n+1} - \eta^2 q_{n+1}^e.$$

The Gaussian identity used in Proposition 6 gives

$$\mathbb{E}[\mathfrak{r}_{n+1}u_{n+1} \mid \mathscr{F}_n] = s_n e_n - a_n^e.$$

Substituting (B.21) proves (B.24). ■

Set

$$\mathfrak{e}_N := \delta + R_s\eta\delta(1+L_2) + \sqrt{\frac{p_d\eta}{\delta}} + \frac{\eta p_d^3}{\delta} + R_s(1+L_2)\sqrt{p_d\eta^3\delta} + R_s(1+L_2)\eta^2 p_d^5. \quad \text{(B.25)}$$

Proposition 14 (Discrete bias tracking) *Assume*

$$\frac{\eta}{4\delta} \le \frac{1}{4}. \quad \text{(B.26)}$$

There exists an event $\mathcal{G}_e$ with

$$\mathbb{P}(\mathcal{G}_e) \ge 1 - Cd^{-K}$$

such that on $\mathcal{G}_{\mathrm{init}} \cap \mathcal{G}_e$, up to τ_Δ,

$$|e_n| \le \exp\left(-\frac{n\eta}{8\delta}\right)|e_0| + C\mathfrak{e}_N, \qquad 0 \le n \le N \wedge \tau_\Delta, \quad \text{(B.27)}$$

and

$$\eta\sum_{m=0}^{n-1}|e_m| \le C\delta|e_0| + C\Lambda_d\mathfrak{e}_N, \qquad 0 \le n \le N \wedge \tau_\Delta. \quad \text{(B.28)}$$

Proof On $\{n < \tau_\Delta\}$,

$$\|\boldsymbol{M}_n\|_F \le 5, \qquad |a_n^e| + s_n \le C,$$

where the bound on a_n^e uses $\|\boldsymbol{T}_\star\|_F = O(1)$ from (A.2). For sufficiently large d, (B.26) and $\delta \to 0$ imply

$$0 \le 1 - \frac{\eta}{4\delta} - \eta s_n \le 1 - \frac{\eta}{8\delta}.$$

Iterating (B.24) with this stable factor gives a deterministic contribution $O(\delta)$ from a_n^e and $O(R_s\eta\delta(1+L_2))$ from the predictable part of q_{n+1}^e. Lemma 11 applied to the stable convolutions of ζ_{n+1}^e and of the centered part of q_{n+1}^e, using (B.22) and (B.23), gives the remaining terms in (B.25). This proves (B.27). Summing the same bound over m and using $N\eta \le C\Lambda_d$ gives (B.28). ■

B.5. Discrete Frobenius localization with explicit remainders

Define the localization remainder scale

$$\mathfrak{l}_N := \sqrt{p_d\eta\Lambda_d} + \eta p_d^3 + R_s(1+L_2)\left(\eta\Lambda_d + \sqrt{p_d\eta^3\Lambda_d} + \eta^2 p_d^5\right).$$

Proposition 15 (Discrete Frobenius localization) *There exists an event $\mathcal{G}_{\mathrm{en}}$ with*

$$\mathbb{P}(\mathcal{G}_{\mathrm{en}}) \ge 1 - Cd^{-K}$$

such that on $\mathcal{G}_{\mathrm{init}} \cap \mathcal{G}_e \cap \mathcal{G}_{\mathrm{en}}$, if

$$\Lambda_d\mathfrak{e}_N \to 0, \qquad \mathfrak{l}_N \to 0, \quad \text{(B.29)}$$

then

$$\tau_\Delta > N, \qquad \sup_{0\le n\le N}\|\boldsymbol{M}_n\|_F \le 5, \qquad I_N := \eta\sum_{n=0}^{N-1}|e_n| = o(1).$$

Proof By Proposition 14,

$$I_N \le C\delta|e_0| + C\Lambda_d \mathfrak{e}_N = o(1).$$

Let

$$\mathcal{E}_n := \|\boldsymbol{\Delta}_n\|_F^2.$$

On $\{n < \tau_\Delta\}$, write

$$\mathcal{D}_n := \mathcal{F}(\boldsymbol{M}_n) + e_n \mathcal{B}(\boldsymbol{M}_n).$$

Then (B.5) gives

$$\boldsymbol{\Delta}_{n+1} = \boldsymbol{\Delta}_n - \eta \mathcal{D}_n - \eta \boldsymbol{Z}_{n+1} - \eta^2 \boldsymbol{Q}_{n+1}.$$

Expanding the square,

$$\mathcal{E}_{n+1} - \mathcal{E}_n = -2\eta\langle \boldsymbol{\Delta}_n, \mathcal{F}(\boldsymbol{M}_n)\rangle - 2\eta e_n \langle \boldsymbol{\Delta}_n, \mathcal{B}(\boldsymbol{M}_n)\rangle - 2\eta\langle \boldsymbol{\Delta}_n, \boldsymbol{Z}_{n+1}\rangle + \mathcal{R}_{n+1},$$

where the full second-order remainder is

$$\begin{aligned}\mathcal{R}_{n+1} := \eta^2\|\mathcal{D}_n\|_F^2 + 2\eta^2\langle \mathcal{D}_n, \boldsymbol{Z}_{n+1}\rangle + \eta^2\|\boldsymbol{Z}_{n+1}\|_F^2 - 2\eta^2\langle \boldsymbol{\Delta}_n, \boldsymbol{Q}_{n+1}\rangle \\ + 2\eta^3\langle \mathcal{D}_n + \boldsymbol{Z}_{n+1}, \boldsymbol{Q}_{n+1}\rangle + \eta^4\|\boldsymbol{Q}_{n+1}\|_F^2.\end{aligned}$$

The monotonicity identity (**??**) gives

$$\langle \boldsymbol{\Delta}_n, \mathcal{F}(\boldsymbol{M}_n)\rangle = 2\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Delta}_n \boldsymbol{M}_n^{1/2}\|_F^2 \ge 0.$$

Moreover,

$$|e_n\langle \boldsymbol{\Delta}_n, \mathcal{B}(\boldsymbol{M}_n)\rangle| \le C|e_n|.$$

The remainder terms are controlled by Lemmas 10, 12, and 11, and the scale conditions (B.29) ensure that their cumulative contribution is $o(1)$ uniformly over $n \le N$ on $\mathcal{G}_{\mathrm{en}}$. Hence

$$\sup_{m \le N\wedge\tau_\Delta} \mathcal{E}_m \le \mathcal{E}_0 + CI_N + o(1).$$

On $\mathcal{G}_{\mathrm{init}}$, $\mathcal{E}_0 \le 9$. Thus the right-hand side is strictly less than 16 for all large d, so $\tau_\Delta > N$. Finally,

$$\|\boldsymbol{M}_n\|_F \le \|\boldsymbol{S}^\star\|_F + \|\boldsymbol{\Delta}_n\|_F \le 5$$

for all $0 \le n \le N$. ■

B.6. Rest-block reduction for SGD

Write

$$\boldsymbol{M}_n = \begin{pmatrix} \boldsymbol{H}_n & \boldsymbol{R}_n \\ \boldsymbol{R}_n^\top & \boldsymbol{J}_n \end{pmatrix}, \qquad \theta_n := \mathrm{tr}(\boldsymbol{J}_n).$$

Inside the rest block, define

$$a_n := \mathrm{tr}(\boldsymbol{P}_{\mathrm{s}}\boldsymbol{M}_n), \qquad c_n := \mathrm{tr}(\boldsymbol{P}_{\mathrm{t}}\boldsymbol{M}_n), \qquad \theta_n = a_n + c_n, \qquad \Theta_N := \eta\sum_{n=0}^{N-1}\theta_n.$$

Lemma 16 (Projector trace recursion) *For every diagonal projector $\boldsymbol{P}$,*

$$u_{n+1}^{(\boldsymbol{P})} \le (1 + 2\eta\tau(\boldsymbol{P}) + \eta\sigma(\boldsymbol{P})|e_n|)\, u_n^{(\boldsymbol{P})} + \eta d_{n+1}^{(\boldsymbol{P})} + \eta^2 q_{n+1}^{(\boldsymbol{P})}. \tag{B.30}$$

Proof Apply $\mathrm{tr}(\boldsymbol{P}\cdot)$ to (B.5). Since $\boldsymbol{P}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{T}_\star$ are diagonal and $\boldsymbol{M}_n \succeq 0$,

$$\mathrm{tr}(\boldsymbol{P}\mathcal{F}^\sharp(\boldsymbol{M}_n, e_n)) \le (2\tau(\boldsymbol{P}) + \sigma(\boldsymbol{P})|e_n|)\, u_n^{(\boldsymbol{P})}.$$

This proves (B.30). ∎

Set

$$\tau_{\mathrm{t}} := \tau(\boldsymbol{P}_{\mathrm{t}}), \qquad \sigma_{\mathrm{t}} := \sigma(\boldsymbol{P}_{\mathrm{t}}), \qquad \gamma_{\mathrm{t}} := 1 + 2\eta\tau_{\mathrm{t}}.$$

Proposition 17 (Tail trace bound) *There exists an event $\mathcal{G}_{\mathrm{t}}$ with*

$$\mathbb{P}(\mathcal{G}_{\mathrm{t}}) \ge 1 - Cd^{-K}$$

such that on $\mathcal{G}_{\mathrm{init}} \cap \mathcal{G}_e \cap \mathcal{G}_{\mathrm{en}} \cap \mathcal{G}_{\mathrm{t}}$,

$$\sup_{0\le n\le N} c_n \le e^{\sigma_{\mathrm{t}} I_N} \left[\gamma_{\mathrm{t}}^N c_0 + CR_s\gamma_{\mathrm{t}}^N \eta^2 N(1+L_2) + \mathfrak{M}_{\mathrm{t}}\right], \tag{B.31}$$

where

$$\mathfrak{M}_{\mathrm{t}} := C\gamma_{\mathrm{t}}^N \Big[\sqrt{p_d\eta(1+L_2)\Lambda_d} + \eta p_d^3(1+L_2)^{1/2} + R_s(1+L_2)\left(\sqrt{p_d\eta^3\Lambda_d} + \eta^2 p_d^5\right)\Big].$$

If

$$\left(\frac{d}{Z_r}\right)^{\rho} \frac{1}{d} \sum_{i > m_\rho(\bar{t})} \sigma_i \to 0, \qquad \left(\frac{d}{Z_r}\right)^{2\rho} R_s\eta(1+L_2)\Lambda_d^8 \to 0, \tag{B.32}$$

then

$$\sup_{0\le n\le N} c_n = o(1), \qquad \eta \sum_{n=0}^{N-1} c_n = o(1).$$

Proof Apply Lemma 16 with $\boldsymbol{P} = \boldsymbol{P}_{\mathrm{t}}$:

$$c_{n+1} \le (\gamma_{\mathrm{t}} + \eta\sigma_{\mathrm{t}}|e_n|)c_n + \eta d_{n+1}^{(\boldsymbol{P}_{\mathrm{t}})} + \eta^2 q_{n+1}^{(\boldsymbol{P}_{\mathrm{t}})}.$$

Iterating and using

$$\prod_{\ell=m+1}^{n-1} (\gamma_{\mathrm{t}} + \eta\sigma_{\mathrm{t}}|e_\ell|) \le \gamma_{\mathrm{t}}^{n-1-m} e^{\sigma_{\mathrm{t}} I_N}$$

gives (B.31), after applying Lemmas 12 and 11 to the martingale terms.

Since

$$\gamma_{\mathrm{t}}^N \le \left(\frac{d}{Z_r}\right)^{\rho+o(1)}$$

by the definition of $\boldsymbol{P}_{\mathrm{t}}$, and

$$c_0 \le \frac{C}{d} \sum_{i > m_\rho(\bar{t})} \sigma_i$$

by (B.11), the initial term is $o(1)$. The martingale and predictable second-order terms are $o(1)$ under (B.32). Summing the same recursion over n gives the integrated bound. ■

Set

$$\tau_{\mathrm{s}} := \tau(\boldsymbol{P}_{\mathrm{s}}), \qquad \sigma_{\mathrm{s}} := \sigma(\boldsymbol{P}_{\mathrm{s}}), \qquad \gamma_{\mathrm{s}} := 1 + 2\eta\tau_{\mathrm{s}}.$$

Proposition 18 (Slow trace bound) *If $k < r$, there exists an event $\mathcal{G}_{\mathrm{s}}$ with*

$$\mathbb{P}(\mathcal{G}_{\mathrm{s}}) \ge 1 - Cd^{-K}$$

such that on $\mathcal{G}_{\mathrm{init}} \cap \mathcal{G}_e \cap \mathcal{G}_{\mathrm{en}} \cap \mathcal{G}_{\mathrm{s}}$,

$$\sup_{0 \le n \le N} a_n \le C e^{\sigma_{\mathrm{s}} I_N} \gamma_{\mathrm{s}}^N \left(\frac{1}{d} + R_s \eta p_d \right), \tag{B.33}$$

$$\eta \sum_{n=0}^{N-1} a_n \le C e^{\sigma_{\mathrm{s}} I_N} \gamma_{\mathrm{s}}^N \left(\frac{1}{d} + R_s \eta p_d \right). \tag{B.34}$$

If

$$R_s \eta p_d \left(\frac{d}{Z_r} \right)^{a_{\mathrm{s}}} \to 0,$$

then both quantities in (B.33)–(B.34) are $o(1)$. If $k = r$, then $a_n \equiv 0$.

Proof Assume $k < r$. Apply Lemma 16 with $\boldsymbol{P} = \boldsymbol{P}_{\mathrm{s}}$. Let

$$R_\sharp := K_0 \left(\frac{1}{d} + R_s \eta p_d \right),$$

where K_0 is a sufficiently large constant, and stop the normalized process $\gamma_{\mathrm{s}}^{-n} a_n$ at its first exit above $R_\sharp$. On the stopped interval,

$$a_m \le R_\sharp e^{\sigma_{\mathrm{s}} I_N} \gamma_{\mathrm{s}}^m.$$

The bounded-support estimate (B.20) and Lemma 11 give

$$\left\| \sup_{n \le N} \left| \eta \sum_{m=0}^{n-1} \gamma_{\mathrm{s}}^{-m-1} d_{m+1}^{(\boldsymbol{P}_{\mathrm{s}})} \right| \right\|_{L^{p_d}} \le C \left(\sqrt{p_d \eta R_\sharp} + \eta p_d^3 R_\sharp^{1/2} \right),$$

and the centered second-order part is bounded by

$$C R_s \left(\sqrt{p_d \eta^3 \Lambda_d} + \eta^2 p_d^5 \right).$$

The predictable second-order contribution is $O(R_s \eta)$. These terms are absorbed into $R_\sharp$ with probability at least $1 - Cd^{-K}$ if K_0 is large. Since $a_0 \le C/d$, the stopped process does not exit before N. Finally,

$$\gamma_{\mathrm{s}}^N \le \left(\frac{d}{Z_r} \right)^{a_{\mathrm{s}} + o(1)},$$

which proves the result. If $k = r$, then $\boldsymbol{P}_{\mathrm{s}} = 0$. ■

Corollary 19 (Rest block and cross block for SGD) *On the intersection of the good events constructed above,*

$$\sup_{0\le n\le N}\theta_n = o(1), \qquad \Theta_N = o(1), \qquad \sup_{0\le n\le N}\|\boldsymbol{J}_n\|_F = o(1), \qquad \sup_{0\le n\le N}\|\boldsymbol{R}_n\|_F = o(1).$$

B.7. Fast-block bootstrap for SGD

Assume $k \ge 1$. Write

$$\boldsymbol{H}_n = \boldsymbol{D}_n + \boldsymbol{E}_n, \qquad \boldsymbol{D}_n := \mathrm{diag}(h_{1,n},\dots,h_{k,n}), \qquad (\boldsymbol{E}_n)_{ii} = 0, \qquad \boldsymbol{\Gamma}_n := \boldsymbol{R}_n\boldsymbol{R}_n^\top.$$

For $1 \le i \le k$, define the matched logistic sequence

$$\widetilde{\ell}_{i,n+1} = \widetilde{\ell}_{i,n}\left(1+\eta(\tau_i - \sigma_i\widetilde{\ell}_{i,n})\right)^2, \qquad \widetilde{\ell}_{i,0} = h_{i,0}, \tag{B.35}$$

and

$$\widetilde{\boldsymbol{L}}_n^{\mathrm{f}} := \mathrm{diag}(\widetilde{\ell}_{1,n},\dots,\widetilde{\ell}_{k,n}).$$

For $i \ne j$, define

$$Y_{ij,n} := \frac{(\boldsymbol{E}_n)_{ij}}{\sqrt{\widetilde{\ell}_{i,n}\widetilde{\ell}_{j,n}}}, \qquad X_n := \max_{i\ne j}\sup_{0\le m\le n}|Y_{ij,m}|,$$

and

$$W_n := \max_{1\le i\le k}\sup_{0\le m\le n}\left|\frac{h_{i,m}}{\widetilde{\ell}_{i,m}} - 1\right|.$$

Lemma 20 (Reciprocal sums for the matched logistics) *On $\mathcal{G}_{\mathrm{init}}$, for every $1 \le i \le k$,*

$$\eta\sum_{n=0}^{N-1}\widetilde{\ell}_{i,n}^{-1} \le Cd, \tag{B.36}$$

and

$$\eta\sum_{n=0}^{N-1}\widetilde{\ell}_{i,n}^{-2} \le Cd^2. \tag{B.37}$$

Proof Fix $i \le k$ and set $s^\star := \tau_i/\sigma_i = s_i^\star = \sigma_i\lambda_i/Z_r$. Define

$$z_n := \frac{\widetilde{\ell}_{i,n}}{s^\star} \in (0,1].$$

Then

$$z_{n+1} = z_n\big(1+\eta\tau_i(1-z_n)\big)^2, \qquad 0 < z_0 = \frac{h_{i,0}}{s^\star}.$$

On $\mathcal{G}_{\mathrm{init}}$, $h_{i,0} \asymp \sigma_i/d$, hence for fixed i,

$$z_0 \asymp \frac{\sigma_i/d}{\sigma_i\lambda_i/Z_r} = \frac{Z_r}{d\,\lambda_i} \asymp \frac{Z_r}{d}.$$

Let $n_{1/2} := \inf\{n \geq 0 : z_n \geq 1/2\}$. For $n < n_{1/2}$, $1 - z_n \geq 1/2$ and hence

$$z_{n+1} \geq z_n\Big(1 + \tfrac{1}{2}\eta\tau_i\Big)^2.$$

Therefore

$$\eta \sum_{n=0}^{(n_{1/2}-1)\wedge(N-1)} z_n^{-1} \leq \eta z_0^{-1} \sum_{n\geq 0}\Big(1 + \tfrac{1}{2}\eta\tau_i\Big)^{-2n} \leq C z_0^{-1} \asymp C\frac{d}{Z_r}.$$

For $n \geq n_{1/2}$, $z_n^{-1} \leq 2$, hence

$$\eta \sum_{n=n_{1/2}\wedge N}^{N-1} z_n^{-1} \leq 2\eta N \leq C \log \frac{d}{Z_r}.$$

Combining and using $\widetilde{\ell}_{i,n}^{-1} = (s^\star)^{-1} z_n^{-1}$ with $(s^\star)^{-1} = Z_r/(\sigma_i\lambda_i) = O(Z_r)$ (for fixed i), we obtain (B.36).

The bound (B.37) is analogous, using z_n^{-2} and the same splitting at $n_{1/2}$, which yields $\eta \sum z_n^{-2} \lesssim z_0^{-2} + \eta N \asymp d^2/Z_r^2 + \log(d/Z_r)$ and multiplying by $(s^\star)^{-2} \asymp Z_r^2$. ∎

Lemma 21 (Fast-block recursion) *On $\mathcal{G}_{\mathrm{loc},n}$, for $1 \leq i \leq k$,*

$$\begin{aligned} h_{i,n+1} &= h_{i,n} + 2\eta\tau_i h_{i,n} - 2\eta\sigma_i h_{i,n}^2 - \eta q_{i,n}^{\mathrm{f}} + \eta\sigma_i e_n h_{i,n} + \eta z_{i,n+1}^{\mathrm{d}} + \eta^2 b_{i,n+1}^{\mathrm{d}}, \qquad \text{(B.38)} \\ q_{i,n}^{\mathrm{f}} &:= 2\sigma_i \sum_{j\neq i} (\boldsymbol{E}_n)_{ij}^2 + 2\sigma_i(\boldsymbol{\Gamma}_n)_{ii} \geq 0. \end{aligned}$$

For $i \neq j$,

$$\begin{aligned} (\boldsymbol{E}_{n+1})_{ij} &= (\boldsymbol{E}_n)_{ij} + \eta c_{ij,n}(\boldsymbol{E}_n)_{ij} - \eta(\sigma_i+\sigma_j)(\boldsymbol{E}_n^2)_{ij} - \eta(\sigma_i+\sigma_j)(\boldsymbol{\Gamma}_n)_{ij} \\ &\quad + \frac{\eta}{2}(\sigma_i+\sigma_j) e_n (\boldsymbol{E}_n)_{ij} + \eta z_{ij,n+1}^{\mathrm{od}} + \eta^2 b_{ij,n+1}^{\mathrm{od}}, \qquad \text{(B.39)} \\ c_{ij,n} &:= \tau_i + \tau_j - (\sigma_i+\sigma_j)(h_{i,n} + h_{j,n}). \end{aligned}$$

Here

$$z_{i,n+1}^{\mathrm{d}} := (\boldsymbol{Z}_{n+1})_{ii}, \quad b_{i,n+1}^{\mathrm{d}} := (\boldsymbol{Q}_{n+1})_{ii}, \quad z_{ij,n+1}^{\mathrm{od}} := (\boldsymbol{Z}_{n+1})_{ij}, \quad b_{ij,n+1}^{\mathrm{od}} := (\boldsymbol{Q}_{n+1})_{ij}.$$

Proof Take the (i,i) and (i,j) components of (B.5). The algebra is left to the reader. ∎

Fix

$$0 < \varepsilon_\star < \frac{\sigma_{\min}}{2\sigma_{\max}},$$

and define

$$\nu_\star := \inf\{m \geq 0 : W_m > \varepsilon_\star/2\}, \qquad \mu_\theta := \inf\{m \geq 0 : \theta_m > 1\}.$$

Lemma 22 (Fast noise bounds) *On* $\{n < \nu_\star \wedge \mu_\theta\} \cap \mathcal{G}_{\mathrm{loc},n}$, *for all* $p \geq 2$,

$$\|z^{\mathrm{d}}_{i,n+1}\|_{L^p(\mathbb{P}(\cdot|\mathscr{F}_n))} \leq Cp^2\widetilde{\ell}^{1/2}_{i,n}, \tag{B.40}$$

$$\|z^{\mathrm{od}}_{ij,n+1}\|_{L^p(\mathbb{P}(\cdot|\mathscr{F}_n))} \leq Cp^2\left(\widetilde{\ell}^{1/2}_{i,n} + \widetilde{\ell}^{1/2}_{j,n}\right). \tag{B.41}$$

Writing

$$b^{\mathrm{d}}_{i,n+1} = \overline{b}^{\mathrm{d}}_{i,n} + \widehat{b}^{\mathrm{d}}_{i,n+1}, \qquad b^{\mathrm{od}}_{ij,n+1} = \overline{b}^{\mathrm{od}}_{ij,n} + \widehat{b}^{\mathrm{od}}_{ij,n+1},$$

with conditional means as predictable parts,

$$|\overline{b}^{\mathrm{d}}_{i,n}| + |\overline{b}^{\mathrm{od}}_{ij,n}| \leq C, \qquad \|\widehat{b}^{\mathrm{d}}_{i,n+1}\|_{L^p(\mathbb{P}(\cdot|\mathscr{F}_n))} + \|\widehat{b}^{\mathrm{od}}_{ij,n+1}\|_{L^p(\mathbb{P}(\cdot|\mathscr{F}_n))} \leq Cp^4.$$

Proof For the diagonal first-order term,

$$z^{\mathrm{d}}_{i,n+1} = \mathfrak{r}_{n+1}\langle \boldsymbol{K}^{\mathrm{d}}_{i,n}, \boldsymbol{\chi}_{n+1}\rangle - \mathbb{E}[\mathfrak{r}_{n+1}\langle \boldsymbol{K}^{\mathrm{d}}_{i,n}, \boldsymbol{\chi}_{n+1}\rangle \mid \mathscr{F}_n],$$

where

$$\boldsymbol{K}^{\mathrm{d}}_{i,n} = \frac{1}{2}(\boldsymbol{M}_n\boldsymbol{e}_i\boldsymbol{e}_i^\top\boldsymbol{\Sigma} + \boldsymbol{\Sigma}\boldsymbol{e}_i\boldsymbol{e}_i^\top\boldsymbol{M}_n).$$

On $\{n < \nu_\star\}$, $h_{i,n} \asymp \widetilde{\ell}_{i,n}$; PSD gives

$$\|\boldsymbol{K}^{\mathrm{d}}_{i,n}\|_F \leq C\widetilde{\ell}^{1/2}_{i,n}.$$

The off-diagonal case is identical. Lemma 10 proves (B.40)–(B.41). The second-order bounds follow from conditional Hölder, using $k = O_{\bar{t}}(1)$, $\theta_n \leq 1$, and $\|\boldsymbol{M}_n\|_F \leq 5$. ∎

Lemma 23 (Fast-block algebra) *On* $\{W_n \leq \varepsilon_\star/2\}$,

$$\left(1 - \frac{\varepsilon_\star}{2}\right)\widetilde{\ell}_{i,n} \leq h_{i,n} \leq \left(1 + \frac{\varepsilon_\star}{2}\right)\widetilde{\ell}_{i,n}. \tag{B.42}$$

Moreover,

$$\frac{q^{\mathrm{f}}_{i,n}}{\widetilde{\ell}_{i,n}} \leq C(X_n^2 + \theta_n), \tag{B.43}$$

$$\frac{|(\boldsymbol{E}_n^2)_{ij}|}{\sqrt{\widetilde{\ell}_{i,n}\widetilde{\ell}_{j,n}}} \leq CX_n^2, \tag{B.44}$$

$$\frac{|(\boldsymbol{\Gamma}_n)_{ij}|}{\sqrt{\widetilde{\ell}_{i,n}\widetilde{\ell}_{j,n}}} + \frac{(\boldsymbol{\Gamma}_n)_{ii}}{\widetilde{\ell}_{i,n}} \leq C\theta_n. \tag{B.45}$$

Finally, for all sufficiently large d,

$$0 \leq \frac{1 + \eta c_{ij,n}}{\left(1 + \eta(\tau_i - \sigma_i\widetilde{\ell}_{i,n})\right)\left(1 + \eta(\tau_j - \sigma_j\widetilde{\ell}_{j,n})\right)} \leq 1. \tag{B.46}$$

Proof The first display is the definition of W_n. The estimate (B.44) follows from

$$(\boldsymbol{E}_n)_{ab} = Y_{ab,n}\sqrt{\widetilde{\ell}_{a,n}\widetilde{\ell}_{b,n}}$$

and $k = O_{\bar{t}}(1)$. We have

$$\|\operatorname{row}_i(\boldsymbol{R}_n)\|_2^2 \le h_{i,n}\theta_n,$$

which gives (B.45); (B.43) follows from the definition of $q^{\mathrm{f}}_{i,n}$. The last claim follows by expanding numerator and denominator in (B.46), using (B.42), $\varepsilon_\star < \sigma_{\min}/(2\sigma_{\max})$, and $\eta\tau_1 \le 1/4$. ∎

Define

$$\mathfrak{M}_{\mathrm{h}} := C\left(\sqrt{dp_d\eta} + \sqrt{d}\,\eta p_d^3\right), \qquad \mathfrak{B}_{\mathrm{h}} := C\left(d\sqrt{p_d\eta^3} + d\eta^2 p_d^5\right).$$

Proposition 24 (Fast martingale event) *There exists an event $\mathcal{G}_{\mathrm{h}}$ with*

$$\mathbb{P}(\mathcal{G}_{\mathrm{h}}) \ge 1 - Cd^{-K}$$

such that on $\mathcal{G}_{\mathrm{h}}$, for all $1 \le i \le k$ and $i \ne j$,

$$\sup_{0\le n\le N}\left|\sum_{\ell=0}^{(n-1)\wedge(\nu_\star\wedge\mu_\theta-1)} \eta\frac{z^{\mathrm{d}}_{i,\ell+1}}{\widetilde{\ell}_{i,\ell}}\right| + \sup_{0\le n\le N}\left|\sum_{\ell=0}^{(n-1)\wedge(\nu_\star\wedge\mu_\theta-1)} \eta\frac{z^{\mathrm{od}}_{ij,\ell+1}}{\sqrt{\widetilde{\ell}_{i,\ell}\widetilde{\ell}_{j,\ell}}}\right| \le \mathfrak{M}_{\mathrm{h}}, \tag{B.47}$$

$$\sup_{0\le n\le N}\left|\sum_{\ell=0}^{(n-1)\wedge(\nu_\star\wedge\mu_\theta-1)} \eta^2\frac{\widehat{b}^{\mathrm{d}}_{i,\ell+1}}{\widetilde{\ell}_{i,\ell}}\right| + \sup_{0\le n\le N}\left|\sum_{\ell=0}^{(n-1)\wedge(\nu_\star\wedge\mu_\theta-1)} \eta^2\frac{\widehat{b}^{\mathrm{od}}_{ij,\ell+1}}{\sqrt{\widetilde{\ell}_{i,\ell}\widetilde{\ell}_{j,\ell}}}\right| \le \mathfrak{B}_{\mathrm{h}}. \tag{B.48}$$

Proof Apply Lemma 11 to the martingale sums in (B.47)–(B.48). Lemma 22 gives the one-step moments, and Lemma 20 gives the reciprocal variance sums. A union bound over the fixed fast block completes the proof. ∎

Proposition 25 (Fast bootstrap for SGD) *Work on the intersection of all good events needed above. Set*

$$\Xi_N := e^{\sigma_{\max}I_N}\left(\delta_0^{\mathrm{coh}} + C\Theta_N + Cd\eta + C\mathfrak{M}_{\mathrm{h}} + C\mathfrak{B}_{\mathrm{h}}\right),$$
$$R_W := C\left(N\eta\,\Xi_N^2 + \Theta_N + d\eta + \mathfrak{M}_{\mathrm{h}} + \mathfrak{B}_{\mathrm{h}} + I_N\right).$$

If

$$4CN\eta\,\Xi_N \le 1, \qquad R_W \le \frac{\varepsilon_\star}{2}, \tag{B.49}$$

then, for all $0 \le n \le N$,

$$X_n \le 2\Xi_N, \qquad W_n \le R_W.$$

Consequently,

$$\sup_{0\le n\le N}\|\boldsymbol{E}_n\|_F \le C\Xi_N, \qquad \sup_{0\le n\le N}\|\boldsymbol{D}_n - \widetilde{\boldsymbol{L}}^{\mathrm{f}}_n\|_F \le CR_W.$$

Proof Since $\Theta_N = o(1)$ on the good event, $\mu_\theta > N$ for all large d. Let

$$\nu := \inf\{n \geq 0 : W_n > R_W\}.$$

On $n < \nu$, Lemma 23 applies. Divide (B.39) by $\sqrt{\widetilde{\ell}_{i,n+1}\widetilde{\ell}_{j,n+1}}$. Using (B.46),

$$|Y_{ij,n+1}| \leq (1 + C\eta|e_n|)|Y_{ij,n}| + C\eta X_n^2 + C\eta\theta_n + \xi_{ij,n+1},$$

where the cumulative normalized martingale and centered second-order contributions are bounded by

$$\sup_{0\leq n\leq N}\left|\sum_{m=0}^{n-1}\xi_{ij,m+1}\right| \leq C(d\eta + \mathfrak{M}_{\mathrm{h}} + \mathfrak{B}_{\mathrm{h}})$$

thanks to Proposition 24; the predictable second-order contribution is $O(d\eta)$ by Lemma 20. Hence

$$X_n \leq \Xi_N + C\eta\sum_{m=0}^{n-1} X_m^2 \qquad (n \leq \nu).$$

The quadratic induction under (B.49) gives $X_n \leq 2\Xi_N$ for $n \leq \nu$.

For the diagonal part, set

$$w_{i,n} := \frac{h_{i,n}}{\widetilde{\ell}_{i,n}} - 1.$$

Subtract (B.35) from (B.38), divide by $\widetilde{\ell}_{i,n+1}$, and use Lemma 23. The deterministic logistic term is contracting up to harmless $O(\eta|e_n|)$ factors, while

$$\frac{q^{\mathrm{f}}_{i,n}}{\widetilde{\ell}_{i,n}} \leq C(X_n^2 + \theta_n)$$

by (B.43). Lemma 20 and Proposition 24 control the normalized stochastic and second-order terms. Consequently,

$$W_n \leq C\left(N\eta\,\Xi_N^2 + \Theta_N + d\eta + \mathfrak{M}_{\mathrm{h}} + \mathfrak{B}_{\mathrm{h}} + I_N\right) = R_W \qquad (n \leq \nu).$$

Thus the bootstrap improves strictly, so $\nu > N$. The matrix bounds follow from the definitions of X_n, W_n and $k = O_{\bar{t}}(1)$. ■

Lemma 26 (Matched and canonical logistics in discrete time) *Assume*

$$\frac{p_d}{r_s} \to 0.$$

Then

$$\|\widetilde{\boldsymbol{L}}^{\mathrm{f}}_N - \boldsymbol{L}^{\mathrm{f}}_N\|_F = o(1) \tag{B.50}$$

with probability $1 - o(1)$.

Proof Fix $i \le k$ and define

$$a_{i,n} := \frac{\ell_{i,n}}{s_i^\star}, \qquad \widetilde{a}_{i,n} := \frac{\widetilde{\ell}_{i,n}}{s_i^\star}.$$

Both sequences satisfy

$$u_{n+1} = u_n(1 + \eta\tau_i(1 - u_n))^2.$$

By (B.12),

$$\frac{d\, h_{i,0}}{\sigma_i} = 1 + O_{\mathbb{P}}\left(\sqrt{\frac{p_d}{r_s}}\right),$$

so $a_{i,0}$ and $\widetilde{a}_{i,0}$ have the same Z_r/d scale. Since $i \le k$ implies $\bar{t}\kappa_i \ge 1 + \delta_{\mathrm{gap}}$, Lemma 8 yields

$$a_{i,N} \to 1, \qquad \widetilde{a}_{i,N} \to 1.$$

Therefore $|\widetilde{\ell}_{i,N} - \ell_{i,N}| = o(1)$. Summing over $O_{\bar{t}}(1)$ fast indices proves (B.50). ∎

Theorem 27 (Comparison with the centered ideal discrete flow) *Assume*

$$\eta\tau_1 \le \frac{1}{4}, \qquad \frac{\eta}{4\delta} \le \frac{1}{4}, \qquad r_s \gg p_d\Lambda_d^2.$$

Assume also

$$\Lambda_d \mathfrak{e}_N \to 0, \qquad d\eta\Lambda_d^8 \to 0, \tag{B.51}$$

and

$$\left(\frac{d}{Z_r}\right)^{2\rho} R_s\eta(1 + L_2)\Lambda_d^8 \to 0, \qquad \left(\frac{d}{Z_r}\right)^{\rho} \frac{1}{d} \sum_{i > m_\rho(\bar{t})} \sigma_i \to 0. \tag{B.52}$$

If $k < r$, assume further

$$R_s\eta p_d \left(\frac{d}{Z_r}\right)^{a_{\mathrm{s}}} \to 0.$$

Then there exists an event $\mathcal{G}$ with

$$\mathbb{P}(\mathcal{G}) \ge 1 - Cd^{-K} \tag{B.53}$$

such that on $\mathcal{G}$,

$$\|\boldsymbol{M}_N - \boldsymbol{L}_N\|_F = o(1), \qquad |e_N| = o(1). \tag{B.54}$$

Proof Let $\mathcal{G}$ be the intersection of $\mathcal{G}_{\mathrm{init}}$, $\mathcal{G}_e$, $\mathcal{G}_{\mathrm{en}}$, $\mathcal{G}_{\mathrm{t}}$, together with $\mathcal{G}_{\mathrm{s}}$ when $k < r$, and $\mathcal{G}_{\mathrm{h}}$ when $k \ge 1$. Then (B.53) holds.

The assumptions (B.51)–(B.52) imply $\mathfrak{l}_N = o(1)$, so Proposition 15 gives

$$I_N = o(1), \qquad e_N = o(1), \qquad \sup_{0 \le n \le N} \|\boldsymbol{M}_n\|_F \le 5.$$

Corollary 19 gives

$$\sup_{0 \le n \le N} \|\boldsymbol{J}_n\|_F = o(1), \qquad \sup_{0 \le n \le N} \|\boldsymbol{R}_n\|_F = o(1), \qquad \Theta_N = o(1).$$

If $k = 0$, the fast block is empty. By the off-transition gap, choose $\delta_{\mathrm{thr}} > 0$ such that $\bar{t}\kappa_i \le 1 - \delta_{\mathrm{thr}}$ for all $i \le r$ in the relevant finite set. Lemma 8 gives $\ell_{i,N} \to 0$ for all $i \le r$, and (B.7) is nonincreasing for $i > r$. Hence $\|\boldsymbol{L}_N\|_F = o(1)$, and Lemma **??** gives $\|\boldsymbol{M}_N - \boldsymbol{L}_N\|_F = o(1)$.

Assume $k \ge 1$. By (B.13) and $r_s \gg p_d \Lambda_d^2$,

$$\delta_0^{\mathrm{coh}} = O\left(\sqrt{\frac{p_d}{r_s}}\right) = o(\Lambda_d^{-1}).$$

The scale conditions imply

$$\mathfrak{M}_{\mathrm{h}} = o(1), \qquad \mathfrak{B}_{\mathrm{h}} = o(1), \qquad d\eta = o(1), \qquad \Theta_N = o(1), \qquad I_N = o(1).$$

Therefore

$$\Xi_N = o(1), \qquad R_W = o(1), \qquad N\eta\,\Xi_N = o(1).$$

Proposition 25 gives

$$\|\boldsymbol{H}_N - \widetilde{\boldsymbol{L}}_N^{\mathrm{f}}\|_F = o(1),$$

and Lemma 26 gives

$$\|\widetilde{\boldsymbol{L}}_N^{\mathrm{f}} - \boldsymbol{L}_N^{\mathrm{f}}\|_F = o(1).$$

The rest-block estimates and Lemma **??** complete the proof of (B.54). ∎

For the discrete dynamics, define the one-step conditional excess risk

$$R_n := \frac{1}{8}\mathbb{E}\big[(y_{n+1} - \widehat{y}_{n+1})^2 \mid \mathscr{F}_n\big] - \frac{1}{8}\mathbb{E}[\varepsilon_1^2].$$

Then

$$R_n = \frac{1}{4}\|\boldsymbol{S}^\star - \boldsymbol{M}_n\|_F^2 + \frac{1}{8}e_n^2. \tag{B.55}$$

We also write $R(\eta n) := R_n$.

Corollary 28 (Discrete risk asymptotics and main-text form) *Under the assumptions of Theorem 27, the following hold.*

(i) *For fixed reduced time $\bar{t}$,*

$$\|\boldsymbol{S}^\star - \boldsymbol{M}_N\|_F^2 = \sum_{i > m_\star(\bar{t})} (s_i^\star)^2 + o_{\mathbb{P}}(1), \tag{B.56}$$

and therefore

$$R_N = \frac{1}{4}\sum_{i > m_\star(\bar{t})} (s_i^\star)^2 + o_{\mathbb{P}}(1). \tag{B.57}$$

(ii) *Away from the transition set,*

$$R_N = \Theta\left(\bar{t}^{-\frac{2(\alpha+\beta)-1}{\alpha+2\beta}}\right). \tag{B.58}$$

(iii) *If $Z_r = \Theta(1)$, write $n = \lfloor t \log d \rfloor$ and suppose the corresponding reduced time satisfies $\bar{t} \asymp \eta t$ and remains off the transition set. In the mildly overparameterized regime $r_s \gg \operatorname{polylog} d$,*

$$R(\eta t \log d) \asymp \Theta\left((\eta t)^{-\frac{2(\alpha+\beta)-1}{\alpha+2\beta}} + r_s^{-2(\alpha+\beta)+1}\right) \tag{B.59}$$

with high probability, and almost surely after Borel–Cantelli.

Proof Theorem 27 gives

$$\|\boldsymbol{M}_N - \boldsymbol{L}_N\|_F = o_{\mathbb{P}}(1), \qquad e_N = o_{\mathbb{P}}(1).$$

Thus

$$\|\boldsymbol{S}^\star - \boldsymbol{M}_N\|_F^2 = \|\boldsymbol{S}^\star - \boldsymbol{L}_N\|_F^2 + o_{\mathbb{P}}(1).$$

By Lemma 8 and the off-transition gap, the coordinates with $\bar{t}\kappa_i > 1$ converge to $s_i^\star$, while the subcritical coordinates are negligible. Hence (B.56). Equation (B.57) follows from (B.55).

The power law (B.58) is the same integral-test estimate as in Corollary **??**. Finally, when $Z_r = \Theta(1)$, $N_{\text{eff}} \asymp \eta^{-1} \log d$, so $n = t \log d$ corresponds to reduced time $\bar{t} \asymp \eta t$. Combining the dynamical term with the rank floor in Lemma 5 gives (B.59). The almost-sure statement follows from Borel–Cantelli because all high-probability estimates are available with probability at least $1 - C_K d^{-K}$ for arbitrary fixed K. ■

Appendix C. Proof of Proposition 3

Let $\boldsymbol{G} \coloneqq \frac{1}{r_s}\boldsymbol{W}\boldsymbol{W}^\top$ and redefine $\hat{R}$ in (3.2) in terms of $\boldsymbol{G}$, i.e., $\hat{R}(\boldsymbol{G}) = \|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{G}\boldsymbol{\Sigma}^{\frac{1}{2}} - \hat{\boldsymbol{S}}\|_F^2$. We observe that the minimizer of $\hat{R}(\boldsymbol{G})$ is low-rank approximation of $\boldsymbol{\Sigma}^{\frac{-1}{2}}\hat{\boldsymbol{S}}\boldsymbol{\Sigma}^{\frac{-1}{2}}$. For strong anisotropy ($\beta \gg 1/2$), driving $\hat{R}(\boldsymbol{G})$ to zero strictly requires recovering the principal eigenvector of $\boldsymbol{\Sigma}$, which aligns with $\boldsymbol{\theta}_1$. Because estimating this principal direction reduces to a standard phase retrieval problem via second-order Stein-based spectral estimators, the desired sample complexity lower bound follows immediately from [19, Theorem 3].

Appendix D. Normalized SGD Dynamics

We consider

$$y_{t+1} = \left\langle \boldsymbol{T}, \boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}^\top - \boldsymbol{\Sigma} \right\rangle + \epsilon_{t+1} \quad \text{and} \quad \hat{y}(\boldsymbol{W}_t; \boldsymbol{x}_{t+1}) = \left\langle \tfrac{1}{r_s}\boldsymbol{W}_t\boldsymbol{W}_t^\top, \boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}^\top \right\rangle$$

where $\|\epsilon_{t+1}\|_{\psi_2} \le \sigma$, $\boldsymbol{W}_t \in \mathbb{R}^{d \times r_s}$ and

$$r_s \gg \frac{d}{\operatorname{polylog} d}.$$

We use $\hat{y}_{t+1} \coloneqq \hat{y}(\boldsymbol{W}_t; \boldsymbol{x}_{t+1})$ and consider

- The loss function is $\mathcal{L}(\boldsymbol{W}_t; (\boldsymbol{x}_{t+1}, y_{t+1})) = -y_{t+1}\hat{y}_{t+1}$
- The Euclidean gradient is $\nabla\mathcal{L}(\boldsymbol{W}_t) = \frac{-1}{r_s} y_{t+1}\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}^\top \boldsymbol{W}_t$. Therefore, we have

$$\nabla_{\text{St}}\mathcal{L}(\boldsymbol{W}_t) = \frac{-1}{r_s}\left(\boldsymbol{I}_d - \boldsymbol{W}_t\boldsymbol{W}_t^\top\right) y_{t+1}\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}^\top \boldsymbol{W}_t.$$

- We recall that $\boldsymbol{G}_t = \boldsymbol{W}_t\boldsymbol{W}_t^\top$.

Then, (SGD) reads

$$\hat{\boldsymbol{W}}_{t+1} = \boldsymbol{W}_t + \frac{\eta}{2r_s} \underbrace{\left(\boldsymbol{I}_d - \boldsymbol{W}_t\boldsymbol{W}_t^\top\right) y_{t+1}\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}^\top \boldsymbol{W}_t}_{:=\nabla_{\mathrm{St}}\boldsymbol{L}_{t+1}}$$

$$\boldsymbol{W}_{t+1} = \hat{\boldsymbol{W}}_{t+1}\left(\boldsymbol{I}_{r_s} + \frac{\eta^2}{4r_s^2}\underbrace{\nabla_{\mathrm{St}}\boldsymbol{L}_{t+1}^\top\nabla_{\mathrm{St}}\boldsymbol{L}_{t+1}}_{:=\boldsymbol{\mathcal{P}}_{t+1}}\right)^{-1/2}. \tag{SGD}$$

We observe that

$$\begin{aligned}\frac{\eta^2}{4r_s^2}\boldsymbol{\mathcal{P}}_{t+1} &= \frac{\eta^2}{r_s^2}y_{t+1}^2\boldsymbol{W}_t^\top\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}^\top\left(\boldsymbol{I}_d - \boldsymbol{W}_t\boldsymbol{W}_t^\top\right)\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}^\top\boldsymbol{W}_t \\ &= \frac{\eta^2}{4r_s^2}y_{t+1}^2\|\left(\boldsymbol{I}_d - \boldsymbol{W}_t\boldsymbol{W}_t^\top\right)\boldsymbol{x}_{t+1}\|_2^2\boldsymbol{W}_t^\top\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}^\top\boldsymbol{W}_t.\end{aligned}$$

Let

$$c_{t+1}^2 := \frac{\eta^2}{4r_s^2}\|\boldsymbol{\mathcal{P}}_{t+1}\|_2 = \frac{\eta^2}{4r_s^2}y_{t+1}^2\|\left(\boldsymbol{I}_d - \boldsymbol{W}_t\boldsymbol{W}_t^\top\right)\boldsymbol{x}_{t+1}\|_2^2\|\boldsymbol{W}_t^\top\boldsymbol{x}_{t+1}\|_2^2.$$

We define $\boldsymbol{P}_{t+1} := \left(\boldsymbol{I}_{r_s} + \frac{\eta^2}{4r_s^2}\boldsymbol{\mathcal{P}}_{t+1}\right)^{-1/2}$ and since $\boldsymbol{\mathcal{P}}_{t+1}$ is 1-rank, we have

$$\boldsymbol{P}_{t+1}^2 = \boldsymbol{I}_{r_s} - \frac{\eta^2}{4r_s^2}\frac{\boldsymbol{\mathcal{P}}_{t+1}}{1+c_{t+1}^2}.$$

By recalling that $\boldsymbol{G}_t = \boldsymbol{W}_t\boldsymbol{W}_t^\top$, we have

$$\begin{aligned}\boldsymbol{G}_{t+1} &= \hat{\boldsymbol{W}}_{t+1}\hat{\boldsymbol{W}}_{t+1}^\top + \hat{\boldsymbol{W}}_{t+1}(\boldsymbol{P}_{t+1}^2 - \boldsymbol{I}_{r_s})\hat{\boldsymbol{W}}_{t+1}^\top \\ &= \boldsymbol{G}_t + \frac{\eta}{2r_s}\left(\boldsymbol{I}_d - \boldsymbol{G}_t\right)y_{t+1}\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}^\top\boldsymbol{G}_t + \frac{\eta}{2r_s}\boldsymbol{G}_t y_{t+1}\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}^\top\left(\boldsymbol{I}_d - \boldsymbol{G}_t\right) \\ &+ \frac{\eta^2}{4r_s^2}\left(\boldsymbol{I}_d - \boldsymbol{G}_t\right)y_{t+1}\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}^\top\boldsymbol{G}_t y_{t+1}\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}^\top\left(\boldsymbol{I}_d - \boldsymbol{G}_t\right) - \frac{\eta^2}{4r_s^2}\frac{\hat{\boldsymbol{W}}_{t+1}\boldsymbol{\mathcal{P}}_{t+1}\hat{\boldsymbol{W}}_{t+1}^\top}{1+c_{t+1}^2}.\end{aligned}$$

For $\mathsf{T}_\star := \boldsymbol{\Sigma}\boldsymbol{T}\boldsymbol{\Sigma}$ and $\boldsymbol{N}_{t+1} := \frac{1}{2}y_{t+1}\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}^\top - \mathsf{T}_\star$, we write

$$\begin{aligned}\boldsymbol{G}_{t+1} &= \boldsymbol{G}_t + \frac{\eta}{r_s}\left(\boldsymbol{I}_d - \boldsymbol{G}_t\right)(\mathsf{T}_\star + \boldsymbol{N}_{t+1})\boldsymbol{G}_t + \frac{\eta}{r_s}\boldsymbol{G}_t(\mathsf{T}_\star + \boldsymbol{N}_{t+1})\left(\boldsymbol{I}_d - \boldsymbol{G}_t\right) \\ &+ \frac{\eta^2}{4r_s^2}\nabla_{\mathrm{St}}\boldsymbol{L}_{t+1}\nabla_{\mathrm{St}}\boldsymbol{L}_{t+1}^\top - \frac{\eta^2}{4r_s^2}\frac{\hat{\boldsymbol{W}}_{t+1}\boldsymbol{\mathcal{P}}_{t+1}\hat{\boldsymbol{W}}_{t+1}^\top}{1+c_{t+1}^2}.\end{aligned}$$

On the other hand,

$$\begin{aligned}\hat{\boldsymbol{W}}_{t+1}\boldsymbol{\mathcal{P}}_{t+1}\hat{\boldsymbol{W}}_{t+1}^\top &= \left(\boldsymbol{W}_t + \frac{\eta}{2r_s}\nabla_{\mathrm{St}}\boldsymbol{L}_{t+1}\right)\boldsymbol{\mathcal{P}}_{t+1}\left(\boldsymbol{W}_t + \frac{\eta}{2r_s}\nabla_{\mathrm{St}}\boldsymbol{L}_{t+1}\right)^\top \\ &= \boldsymbol{W}_t\boldsymbol{\mathcal{P}}_{t+1}\boldsymbol{W}_t^\top + \frac{\eta}{r_s}\,\mathrm{sym}\left(\nabla_{\mathrm{St}}\boldsymbol{L}_{t+1}\boldsymbol{\mathcal{P}}_{t+1}\boldsymbol{W}_t^\top\right) + \frac{\eta^2}{4r_s^2}\nabla_{\mathrm{St}}\boldsymbol{L}_{t+1}\boldsymbol{\mathcal{P}}_{t+1}\nabla_{\mathrm{St}}\boldsymbol{L}_{t+1}^\top.\end{aligned}$$

We collect the higher order terms in a single term defined as follows:

$$R_{\text{so}}[\boldsymbol{G}_t] \coloneqq \frac{\eta^2}{4r_s^2}\mathbb{E}_t\left[\nabla_{\text{St}}\boldsymbol{L}_{t+1}\nabla_{\text{St}}\boldsymbol{L}_{t+1}^\top\right] - \frac{\eta^2}{4r_s^2}\boldsymbol{W}_t\mathbb{E}_t\left[\frac{\boldsymbol{\mathcal{P}}_{t+1}}{1+c_{t+1}^2}\right]\boldsymbol{W}_t^\top \\ - \frac{\eta^3}{4r_s^3}\,\text{sym}\left(\mathbb{E}_t\left[\frac{\nabla_{\text{St}}\boldsymbol{L}_{t+1}\boldsymbol{\mathcal{P}}_{t+1}}{1+c_{t+1}^2}\right]\boldsymbol{W}_t^\top\right) - \frac{\eta^4}{16r_s^4}\mathbb{E}_t\left[\frac{\nabla_{\text{St}}\boldsymbol{L}_{t+1}\boldsymbol{\mathcal{P}}_{t+1}\nabla_{\text{St}}\boldsymbol{L}_{t+1}^\top}{1+c_{t+1}^2}\right] \quad \text{(D.2)}$$

We collect the noise terms in a single term defined as follows:

$$\frac{\eta}{r_s}\boldsymbol{\xi}_{t+1} \coloneqq \frac{\eta}{r_s}\left(\boldsymbol{I}_d - \boldsymbol{G}_t\right)\boldsymbol{N}_{t+1}\boldsymbol{G}_t + \frac{\eta}{r_s}\boldsymbol{G}_t\boldsymbol{N}_{t+1}\left(\boldsymbol{I}_d - \boldsymbol{G}_t\right) \\ - \frac{\eta^2}{4r_s^2}\boldsymbol{W}_t\left(\frac{\boldsymbol{\mathcal{P}}_{t+1}}{1+c_{t+1}^2} - \mathbb{E}_t\left[\frac{\boldsymbol{\mathcal{P}}_{t+1}}{1+c_{t+1}^2}\right]\right)\boldsymbol{W}_t^\top \\ + \frac{\eta^2}{4r_s^2}\left(\nabla_{\text{St}}\boldsymbol{L}_{t+1}\nabla_{\text{St}}\boldsymbol{L}_{t+1}^\top - \mathbb{E}_t\left[\nabla_{\text{St}}\boldsymbol{L}_{t+1}\nabla_{\text{St}}\boldsymbol{L}_{t+1}^\top\right]\right) \\ - \frac{\eta^3}{4r_s^3}\,\text{sym}\left(\left(\frac{\nabla_{\text{St}}\boldsymbol{L}_{t+1}\boldsymbol{\mathcal{P}}_{t+1}}{1+c_{t+1}^2} - \mathbb{E}_t\left[\frac{\nabla_{\text{St}}\boldsymbol{L}_{t+1}\boldsymbol{\mathcal{P}}_{t+1}}{1+c_{t+1}^2}\right]\right)\boldsymbol{W}_t^\top\right) \\ - \frac{\eta^4}{16r_s^4}\left(\frac{\nabla_{\text{St}}\boldsymbol{L}_{t+1}\boldsymbol{\mathcal{P}}_{t+1}\nabla_{\text{St}}\boldsymbol{L}^\top}{1+c_{t+1}^2} - \mathbb{E}_t\left[\frac{\nabla_{\text{St}}\boldsymbol{L}_{t+1}\boldsymbol{\mathcal{P}}_{t+1}\nabla_{\text{St}}\boldsymbol{L}_{t+1}^\top}{1+c_{t+1}^2}\right]\right).$$

With these definitions in hand, we have

$$\boldsymbol{G}_{t+1} = \boldsymbol{G}_t + \frac{\eta}{r_s}\left(\mathsf{T}_\star\boldsymbol{G}_t + \boldsymbol{G}_t\mathsf{T}_\star - 2\boldsymbol{G}_t\mathsf{T}_\star\boldsymbol{G}_t\right) + R_{\text{so}}[\boldsymbol{G}_t] + \frac{\eta}{2r_s}\boldsymbol{\xi}_{t+1}.$$

Proposition 29 (Including second order terms) *For $\frac{\eta}{r_s} \ll \frac{1}{\sqrt{\text{tr}(\boldsymbol{\Sigma})\,\text{tr}(\boldsymbol{G}_t\boldsymbol{\Sigma})}}$, by including the bounds in (D.5), we get*

$$\boldsymbol{G}_{t+1} \preceq \boldsymbol{G}_t + \frac{\eta}{r_s}\Big(\mathsf{T}_\star\boldsymbol{G}_t + \boldsymbol{G}_t\mathsf{T}_\star - 2\boldsymbol{G}_t\mathsf{T}_\star\boldsymbol{G}_t\Big) + \frac{C\eta^2\,\text{tr}(\boldsymbol{G}_t\boldsymbol{\Sigma})}{2r_s^2}(\boldsymbol{I}_d - \boldsymbol{G}_t)\boldsymbol{\Sigma}(\boldsymbol{I}_d - \boldsymbol{G}_t) + \frac{\eta}{r_s}\boldsymbol{\xi}_{t+1} \quad \text{(D.3)}$$

$$\boldsymbol{G}_{t+1} \succeq \boldsymbol{G}_t + \frac{\eta}{r_s}\Big(\mathsf{T}_\star\boldsymbol{G}_t + \boldsymbol{G}_t\mathsf{T}_\star - 2\boldsymbol{G}_t(\mathsf{T}_\star + \tfrac{C\eta\,\text{tr}(\boldsymbol{\Sigma})}{r_s}\boldsymbol{\Sigma})\boldsymbol{G}_t\Big) \\ + \frac{c\eta^2\,\text{tr}(\boldsymbol{G}_t\boldsymbol{\Sigma})}{2r_s^2}(\boldsymbol{I}_d - \boldsymbol{G}_t)\boldsymbol{\Sigma}(\boldsymbol{I}_d - \boldsymbol{G}_t) + \frac{\eta}{r_s}\boldsymbol{\xi}_{t+1} \quad \text{(D.4)}$$

Proof By using $\text{tr}(\boldsymbol{G}_t\boldsymbol{\Sigma}) \le r_s$, the result is immediate use of the bound (D.5). ■

Proposition 30 (Sufficient statistics) *Let*

$$\mathsf{T} =: \boldsymbol{\Theta}\boldsymbol{\Lambda}\boldsymbol{\Theta}^\top, \qquad \boldsymbol{\Sigma}_1 = \boldsymbol{\Theta}^\top\boldsymbol{\Sigma}\boldsymbol{\Theta}, \qquad \boldsymbol{M}_t \coloneqq \boldsymbol{\Sigma}_1^{\frac{1}{2}}\boldsymbol{G}_t\boldsymbol{\Sigma}_1^{\frac{1}{2}}, \qquad \boldsymbol{\zeta}_t \coloneqq \boldsymbol{\Sigma}_1^{\frac{1}{2}}\boldsymbol{\xi}_t\boldsymbol{\Sigma}_1^{\frac{1}{2}}.$$

We have

$$\boldsymbol{M}_{t+1} \preceq \boldsymbol{M}_t + \frac{\eta}{r_s}\Big(\boldsymbol{\Lambda}\boldsymbol{\Sigma}_1\boldsymbol{M}_t + \boldsymbol{M}_t\boldsymbol{\Sigma}_1\boldsymbol{\Lambda} - 2\boldsymbol{M}_t\boldsymbol{\Lambda}\boldsymbol{M}_t\Big) + \frac{C\eta^2\,\text{tr}(\boldsymbol{G}_t\boldsymbol{\Sigma})}{2r_s^2}(\boldsymbol{\Sigma}_1 - \boldsymbol{M}_t)^2 + \frac{\eta}{r_s}\boldsymbol{\zeta}_{t+1}$$

$$\boldsymbol{M}_{t+1} \succeq \boldsymbol{M}_t + \frac{\eta}{r_s}\Big(\boldsymbol{\Lambda}\boldsymbol{\Sigma}_1\boldsymbol{M}_t + \boldsymbol{M}_t\boldsymbol{\Sigma}_1\boldsymbol{\Lambda} - 2\boldsymbol{M}_t\boldsymbol{\Lambda}\boldsymbol{M}_t\Big) + \frac{c\eta^2\operatorname{tr}(\boldsymbol{G}_t\boldsymbol{\Sigma})}{2r_s^2}(\boldsymbol{\Sigma}_1 - \boldsymbol{M}_t)^2 + \frac{\eta}{r_s}\boldsymbol{\zeta}_{t+1}$$

Proof The proof immediately follows from(D.3) and (D.4). The additional term in the quadratic term of lower bound can be ignored since it does not contribute in asymptotic limit. ■

We introduce block matrix notation:

$$\boldsymbol{M}_t =: \begin{bmatrix}\boldsymbol{M}_{t,11} & \boldsymbol{M}_{t,12}\\ \boldsymbol{M}_{t,12}^\top & \boldsymbol{M}_{t,22}\end{bmatrix}, \quad \boldsymbol{\Lambda} =: \begin{bmatrix}\boldsymbol{\Lambda}_1 & 0\\ 0 & \boldsymbol{\Lambda}_2\end{bmatrix}, \quad \boldsymbol{\zeta}_t =: \begin{bmatrix}\boldsymbol{\zeta}_{t,11} & \boldsymbol{\zeta}_{t,12}\\ \boldsymbol{\zeta}_{t,12}^\top & \boldsymbol{\zeta}_{t,22}.\end{bmatrix}, \quad \boldsymbol{\Sigma}_1 =: \begin{bmatrix}\boldsymbol{\Sigma}_{11} & 0\\ 0 & \boldsymbol{\Sigma}_{12}.\end{bmatrix}$$

where $\boldsymbol{M}_{t,11},\ \boldsymbol{\Lambda}_1,\ \boldsymbol{\zeta}_{t,11},\ \boldsymbol{\Sigma}_{11} \in \mathbb{R}^{r_u\times r_u}$.

Proposition 31 (Sparsification) *We define*

$$\begin{aligned}&\boldsymbol{\Lambda}_{u_1} \coloneqq \boldsymbol{\Lambda}_1\boldsymbol{\Sigma}_{11} - C\eta\operatorname{tr}(\boldsymbol{\Sigma})\big(\boldsymbol{\Sigma}_{11} - \|\boldsymbol{\Sigma}_{12}\|\boldsymbol{I}_{r_u}\big), \quad \textit{and} \quad \boldsymbol{\Lambda}_{u_2} \coloneqq \boldsymbol{\Lambda}_1 - C\eta\operatorname{tr}(\boldsymbol{\Sigma})\boldsymbol{I}_{r_u}\\ &\boldsymbol{\Lambda}_{\ell_1} \coloneqq \boldsymbol{\Lambda}_1\boldsymbol{\Sigma}_{11} - \|\boldsymbol{\Sigma}_{12}\|_2\|\boldsymbol{\Lambda}_2\|_2\boldsymbol{I}_{r_u}, \qquad\qquad \textit{and} \quad \boldsymbol{\Lambda}_{\ell_2} \coloneqq \boldsymbol{\Lambda}_1,\end{aligned}$$

and

$$\boldsymbol{V}_t^+ \coloneqq 2\boldsymbol{\Lambda}_{u_2}^{\frac{1}{2}}\boldsymbol{M}_{t,11}\boldsymbol{\Lambda}_{u_2}^{\frac{1}{2}} - \boldsymbol{\Lambda}_{u_1} \ \textit{and}\ \boldsymbol{V}_t^- \coloneqq 2\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{M}_{t,11}\boldsymbol{\Lambda}^{\frac{1}{2}} - \boldsymbol{\Lambda}.$$

For $\mathfrak{y} = \frac{\eta}{r_s}$*, we have*

$$\boldsymbol{V}_{t+1}^+ \preceq \boldsymbol{V}_t^+\Big(\boldsymbol{I}_{r_u} + \frac{\mathfrak{y}}{1+1.1\mathfrak{y}}\boldsymbol{V}_t^+\Big)^{-1} + \mathfrak{y}\boldsymbol{\Lambda}_{u_1}^2 + \mathfrak{y}^2\operatorname{tr}(\boldsymbol{\Sigma})\boldsymbol{\Lambda}_{u_2}\boldsymbol{\Sigma}_{11}^2 + 2\mathfrak{y}\boldsymbol{\Lambda}_{u_2}^{\frac{1}{2}}\boldsymbol{\zeta}_{t+1,11}\boldsymbol{\Lambda}_{u_2}^{\frac{1}{2}}$$

$$\boldsymbol{V}_{t+1}^- \succeq \boldsymbol{V}_t^-\Big(\boldsymbol{I}_{r_u} + \frac{\mathfrak{y}}{1-1.1\mathfrak{y}}\boldsymbol{V}_t^-\Big)^{-1} + \mathfrak{y}\boldsymbol{\Lambda}_{\ell_2}^2 + 2\mathfrak{y}\boldsymbol{\Lambda}_{\ell_2}^{\frac{1}{2}}\boldsymbol{\zeta}_{t+1,11}\boldsymbol{\Lambda}_{\ell_2}^{\frac{1}{2}}$$

Proof For the upper bound, we observe that

$$(\boldsymbol{\Sigma}_1 - \boldsymbol{M}_t)^2 = \boldsymbol{\Sigma}_1^2 - \boldsymbol{\Sigma}_1\boldsymbol{M}_t - \boldsymbol{M}_t\boldsymbol{\Sigma}_1 + \boldsymbol{M}_t^2$$

Note that $\boldsymbol{M}_{t,12} = \boldsymbol{\Sigma}_{11}^{\frac{1}{2}}\boldsymbol{G}_{t,12}\boldsymbol{\Sigma}_{12}^{\frac{1}{2}}$. Therefore,

$$\boldsymbol{M}_{t,12}\boldsymbol{M}_{t,12}^\top = \boldsymbol{\Sigma}_{11}^{\frac{1}{2}}\boldsymbol{G}_{t,12}\boldsymbol{\Sigma}_{12}^{\frac{1}{2}}\boldsymbol{\Sigma}_{12}^{\frac{1}{2}}\boldsymbol{G}_{t,12}^\top\boldsymbol{\Sigma}_{11}^{\frac{1}{2}} \preceq \|\boldsymbol{\Sigma}_{12}\|_2\boldsymbol{\Sigma}_{11}^{\frac{1}{2}}\boldsymbol{G}_{t,12}\boldsymbol{G}_{t,12}^\top\boldsymbol{\Sigma}_{11}^{\frac{1}{2}} \preceq \|\boldsymbol{\Sigma}_{12}\|_2\boldsymbol{M}_{t,11}.$$

For the lower bound,

$$\boldsymbol{M}_{t,12}\boldsymbol{\Lambda}_2\boldsymbol{M}_{t,12}^\top = \boldsymbol{\Sigma}_{11}^{\frac{1}{2}}\boldsymbol{G}_{t,12}\boldsymbol{\Sigma}_{12}^{\frac{1}{2}}\boldsymbol{\Lambda}_2\boldsymbol{\Sigma}_{12}^{\frac{1}{2}}\boldsymbol{G}_{t,12}^\top\boldsymbol{\Sigma}_{11}^{\frac{1}{2}} \preceq \|\boldsymbol{\Sigma}_{12}\|_2\|\boldsymbol{\Lambda}_2\|_2\boldsymbol{M}_{t,11}.$$

Therefore, the coefficients $\boldsymbol{\Lambda}_{u_1}, \boldsymbol{\Lambda}_{u_2}, \boldsymbol{\Lambda}_{\ell_1}, \boldsymbol{\Lambda}_{\ell_2}$ follow. The bounds can be derived by ordering the higher order terms in $\boldsymbol{V} \to \boldsymbol{V}(\boldsymbol{I}_{r_u} + \eta\boldsymbol{V})^{-1}$ ■

D.1. Definitions and bounding systems

Throughout the proof, we will also use a constant $\kappa_d \in o_d(1)$ that will be specified later. Moreover, we make the following definitions:

- **Noise sequence.** For $\boldsymbol{\nu}_0 = 0$, we define the noise sequence $\boldsymbol{\nu}_{t+1} \coloneqq \boldsymbol{\nu}_t + \eta\boldsymbol{\zeta}_{t+1,11}$.
- **Reference sequence.** For $\boldsymbol{R}_0 = \frac{\kappa_d r_s}{d}\boldsymbol{\Sigma}_{11}$, we define the reference sequence

$$\boldsymbol{R}_{t+1} = \boldsymbol{R}_t + 2(1-2\kappa_d)\eta\left(\boldsymbol{\Lambda}_{\ell_1}\boldsymbol{R}_t - \frac{3\kappa_d+1}{\kappa_d(1-2\kappa_d)}\boldsymbol{\Lambda}_{u_1}\boldsymbol{R}_t^2\right).$$

- **Bounding systems.** We define the lower and upper bounding recursions as

$$\underline{\boldsymbol{V}}_{t+1} = \underline{\boldsymbol{V}}_t\left(\boldsymbol{I}_{r_u} + \frac{\eta(1+2\kappa_d)}{1-1.2\eta}\underline{\boldsymbol{V}}_t\right)^{-1} + \frac{\eta(1+2\kappa_d)}{1-1.2\eta}\left(\frac{\boldsymbol{\Lambda}_{\ell_1}^2}{(1+2\kappa_d)^2} - C(\eta\boldsymbol{\Lambda}_{\ell_1}^2 + \eta^2\boldsymbol{\Lambda}_{\ell_1})\right),$$
$$\bar{\boldsymbol{V}}_{t+1} = \bar{\boldsymbol{V}}_t\left(\boldsymbol{I}_{r_u} + \frac{\eta(1-2\kappa_d)}{1+1.2\eta}\bar{\boldsymbol{V}}_t\right)^{-1} + \frac{\eta(1-2\kappa_d)}{1+1.2\eta}\left(\frac{\boldsymbol{\Lambda}_{u_1}^2}{(1-2\kappa_d)^2} + C(\eta\operatorname{tr}(\boldsymbol{\Sigma})\boldsymbol{\Lambda}_{u_2}\boldsymbol{\Sigma}_{11}^2 + \eta\boldsymbol{\Lambda}_{u_2}^2 + \eta^2\boldsymbol{\Lambda}_{u_2})\right).$$

  where the iterates $\{\underline{\boldsymbol{V}}_{t+1}\}_{t\in\mathbb{N}}$ and $\{\bar{\boldsymbol{V}}_{t+1}\}_{t\in\mathbb{N}}$ are functions of the bounding sequences $\{\underline{\boldsymbol{M}}_t\}_{t\in\mathbb{N}}$ and $\{\bar{\boldsymbol{M}}_t\}_{t\in\mathbb{N}}$ as following:

$$\underline{\boldsymbol{V}}_t = 2\boldsymbol{\Lambda}_{u_2}^{\frac{1}{2}}\underline{\boldsymbol{M}}_t\boldsymbol{\Lambda}_{u_2}^{\frac{1}{2}} - \frac{\boldsymbol{\Lambda}_{u_1}}{1+2\kappa_d} \text{ and } \underline{\boldsymbol{M}}_0 \preceq \boldsymbol{M}_{0,11} - \boldsymbol{R}_0,$$

$$\bar{\boldsymbol{V}}_t = 2\boldsymbol{\Lambda}_{\ell_2}^{\frac{1}{2}}\bar{\boldsymbol{M}}_t\boldsymbol{\Lambda}_{\ell_2}^{\frac{1}{2}} - \frac{\boldsymbol{\Lambda}_{\ell_1}}{1-2\kappa_d} \text{ and } \bar{\boldsymbol{M}}_0 \succeq \boldsymbol{M}_{0,11} + \boldsymbol{R}_0.$$

- **Stopping times.** We define and a sequence of events $\{\mathcal{E}_t\}_{t\geq 0}$

$$\mathcal{E}_t \coloneqq \Big\{ -\kappa_d r_u^{\frac{-\alpha-\beta}{2}}\boldsymbol{R}_t \preceq \boldsymbol{\nu}_t \preceq \kappa_d r_u^{\frac{-\alpha-\beta}{2}}\boldsymbol{R}_t\Big\} \cap \Big\{ -\frac{\kappa_d^2}{4} r_u^{\frac{-\alpha-\beta}{2}}\boldsymbol{R}_t \preceq \boldsymbol{\Sigma}_{11}^{\frac{1}{2}}\boldsymbol{\Lambda}_1^{\frac{1}{2}}\boldsymbol{\nu}_t\boldsymbol{\Lambda}_1^{\frac{1}{2}}\boldsymbol{\Sigma}_{11}^{\frac{1}{2}} \preceq \frac{\kappa_d^2}{4} r_u^{\frac{-\alpha-\beta}{2}}\boldsymbol{R}_t\Big\}.$$

  We define the stopping times

$$\mathcal{T}_{\text{noise}}(\omega) \coloneqq \inf\{t \geq 0 \mid \omega \notin \mathcal{E}_t\} \wedge d^3 \text{ and } \mathcal{T}_{\text{bounded}} \coloneqq \inf\left\{t \geq 0 \;\middle|\; \|\boldsymbol{\Sigma}_{11}^{-1}\underline{\boldsymbol{M}}_t\|_2 \vee \|\boldsymbol{\Sigma}_{11}^{-1}\bar{\boldsymbol{M}}_t\|_2 > 1.5\right\},$$

  and

$$\mathcal{T}_{\text{bad}} \coloneqq \mathcal{T}_{\text{noise}} \wedge \mathcal{T}_{\text{bounded}} \wedge \{t \geq 0 : \|\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{R}_t\|_2 > 1.2\kappa_d\}.$$

We start with the followng lemma:

Lemma 32 *We consider $\kappa_d \ll \frac{r_u^{-1}}{\log d}$. The event $\mathcal{E}_t$ implies for $d \geq \Omega(1)$ and $t \leq \mathcal{T}_{\text{bounded}} \wedge \{t : \|\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{R}_t\|_2 > 1.2\kappa_d\}$ that*

1. $-3\kappa_d\boldsymbol{\Lambda}_{\ell_1}^{\frac{1}{2}}\boldsymbol{R}_t\boldsymbol{\Lambda}_{\ell_1}^{\frac{1}{2}} \preceq \boldsymbol{\Lambda}_{\ell_1}\boldsymbol{\nu}_t + \boldsymbol{\nu}_t\boldsymbol{\Lambda}_{\ell_1}$

2. $\boldsymbol{\Lambda}_{u_1}\boldsymbol{\nu}_t + \boldsymbol{\nu}_t\boldsymbol{\Lambda}_{u_1} \preceq 3\kappa_d\boldsymbol{\Lambda}_{u_1}^{\frac{1}{2}}\boldsymbol{R}_t\boldsymbol{\Lambda}_{u_1}^{\frac{1}{2}}$
3. $\left(\boldsymbol{\Lambda}_{\ell_2}^{\frac{1}{2}}\boldsymbol{\nu}_t\boldsymbol{\Lambda}_{\ell_2}^{\frac{1}{2}}\right)^2 \preceq \frac{\kappa_d^2}{4}\boldsymbol{\Lambda}_{\ell_1}\boldsymbol{R}_t\boldsymbol{\Lambda}_{\ell_1}$
4. $\left(\boldsymbol{\Lambda}_{u_2}^{\frac{1}{2}}\boldsymbol{\nu}_t\boldsymbol{\Lambda}_{u_2}^{\frac{1}{2}}\right)^2 \preceq \frac{\kappa_d^2}{4}\boldsymbol{\Lambda}_{u_1}\boldsymbol{R}_t\boldsymbol{\Lambda}_{u_1}$

Proof For notational convenience, we define $\widetilde{\boldsymbol{\nu}}_t \coloneqq \boldsymbol{R}_t^{\frac{-1}{2}}\boldsymbol{\nu}_t\boldsymbol{R}_t^{\frac{-1}{2}}$. We observe that

$$\begin{aligned}\boldsymbol{\Sigma}_{11}\boldsymbol{\Lambda}_1\widetilde{\boldsymbol{\nu}}_t + \widetilde{\boldsymbol{\nu}}_t\boldsymbol{\Lambda}_1\boldsymbol{\Sigma}_{11} &\preceq \frac{\kappa_d^2}{4}\boldsymbol{\Lambda}_1\boldsymbol{\Sigma}_{11} + \frac{4}{\kappa_d^2}\boldsymbol{\Sigma}_{11}^{\frac{-1}{2}}\boldsymbol{\Lambda}_1^{\frac{-1}{2}}\left(\boldsymbol{\Sigma}_{11}^{\frac{1}{2}}\boldsymbol{\Lambda}_1^{\frac{1}{2}}\widetilde{\boldsymbol{\nu}}_t\boldsymbol{\Lambda}_1^{\frac{1}{2}}\boldsymbol{\Sigma}_{11}^{\frac{1}{2}}\right)^2\boldsymbol{\Lambda}_1^{\frac{-1}{2}}\boldsymbol{\Sigma}_{11}^{\frac{-1}{2}}\\ &\preceq \frac{\kappa_d^2}{4}\boldsymbol{\Lambda}_1\boldsymbol{\Sigma}_{11} + \frac{\kappa_d^2}{4}r_u^{-\alpha-\beta}\boldsymbol{I}_{r_u}\end{aligned}$$

$$\begin{aligned}\boldsymbol{\Sigma}_{11}\boldsymbol{\Lambda}_1\widetilde{\boldsymbol{\nu}}_t + \widetilde{\boldsymbol{\nu}}_t\boldsymbol{\Lambda}_1\boldsymbol{\Sigma}_{11} &\succeq -\frac{\kappa_d^2}{4}\boldsymbol{\Lambda}_1\boldsymbol{\Sigma}_{11} - \frac{4}{\kappa_d^2}\boldsymbol{\Sigma}_{11}^{\frac{-1}{2}}\boldsymbol{\Lambda}_1^{\frac{-1}{2}}\left(\boldsymbol{\Sigma}_{11}^{\frac{1}{2}}\boldsymbol{\Lambda}_1^{\frac{1}{2}}\widetilde{\boldsymbol{\nu}}_t\boldsymbol{\Lambda}_1^{\frac{1}{2}}\boldsymbol{\Sigma}_{11}^{\frac{1}{2}}\right)^2\boldsymbol{\Lambda}_1^{\frac{-1}{2}}\boldsymbol{\Sigma}_{11}^{\frac{-1}{2}}\\ &\succeq -\frac{\kappa_d^2}{4}\boldsymbol{\Lambda}_1\boldsymbol{\Sigma}_{11} - \frac{\kappa_d^2}{4}r_u^{-\alpha-\beta}\boldsymbol{I}_{r_u}\end{aligned}$$

The items follows the use of matrix Young's inequality with the given bounds. ■

The main result of this section is the following:

Proposition 33 *Consider d large enough so that $\kappa_d \le \frac{1}{50}$ and η small enough to make sure that*

$$\Big(1 - \frac{C}{\log d}\Big)\boldsymbol{I}_{r_u} \preceq \boldsymbol{\Lambda}_{u_1}\boldsymbol{\Lambda}_{u_2}^{-1}, \boldsymbol{\Lambda}_{\ell_1}\boldsymbol{\Lambda}_{\ell_2}^{-1} \preceq \Big(1 + \frac{C}{\log d}\Big)\boldsymbol{I}_{r_u}.$$

We have

$$\underline{\boldsymbol{M}}_{t\wedge\mathcal{T}_{bad}} + \boldsymbol{R}_{t\wedge\mathcal{T}_{bad}} + \boldsymbol{\nu}_{t\wedge\mathcal{T}_{bad}} \preceq \boldsymbol{M}_{t\wedge\mathcal{T}_{bad},11} \preceq \bar{\boldsymbol{M}}_{t\wedge\mathcal{T}_{bad}} - \boldsymbol{R}_{t\wedge\mathcal{T}_{bad}} + \boldsymbol{\nu}_{t\wedge\mathcal{T}_{bad}}.$$

Proof The proof is identical with [3, Proposition 15]. ■

D.2. Bounding the second order terms

Proposition 34 *There exists a universal constant $C > 0$ such that*

$$R_{so}[\boldsymbol{G}_t] \sim \frac{\eta^2\operatorname{tr}(\boldsymbol{\Sigma})}{r_s^2}\boldsymbol{G}_t\boldsymbol{\Sigma}\boldsymbol{G}_t + \Big(\frac{\eta^2\operatorname{tr}(\boldsymbol{G}_t\boldsymbol{\Sigma})}{r_s^2} + \frac{\eta^4\operatorname{tr}(\boldsymbol{\Sigma})\operatorname{tr}(\boldsymbol{G}_t\boldsymbol{\Sigma})^2}{r_s^4}\Big)(\boldsymbol{I}_d - \boldsymbol{G}_t)\boldsymbol{\Sigma}(\boldsymbol{I}_d - \boldsymbol{G}_t). \tag{D.5}$$

Proof We will bound each term in (D.2) separately. For the first term,

$$\mathbb{E}_t\Big[\nabla_{\mathrm{St}}\boldsymbol{L}_{t+1}\nabla_{\mathrm{St}}\boldsymbol{L}_{t+1}^\top\Big] = (\boldsymbol{I}_d - \boldsymbol{G}_t)\,\mathbb{E}_t\left[y_{t+1}^2\|\boldsymbol{W}_t^\top\boldsymbol{x}_{t+1}\|_2^2\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}^\top\right](\boldsymbol{I}_d - \boldsymbol{G}_t)\,.$$

Therefore,

$$0 \preceq \frac{\eta^2}{4r_s^2}\mathbb{E}_t\left[\nabla_{\mathrm{St}}\boldsymbol{L}_{t+1}\nabla_{\mathrm{St}}\boldsymbol{L}_{t+1}^\top\right] \preceq \frac{C\eta^2}{r_s^2}\operatorname{tr}(\boldsymbol{G}_t\boldsymbol{\Sigma})(\boldsymbol{I}_d - \boldsymbol{G}_t)\boldsymbol{\Sigma}(\boldsymbol{I}_d - \boldsymbol{G}_t).$$

For the second term, we have

$$\boldsymbol{W}_t\mathbb{E}_t\left[\frac{\boldsymbol{\mathcal{P}}_{t+1}}{1+c_{t+1}^2}\right]\boldsymbol{W}_t^\top=\boldsymbol{G}_t\mathbb{E}_t\left[\frac{y_{t+1}^2\|(\boldsymbol{I}_d-\boldsymbol{G}_t)\,\boldsymbol{x}_{t+1}\|_2^2\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}^\top}{1+c_{t+1}^2}\right]\boldsymbol{G}_t.$$

Therefore,

$$0\preceq\frac{\eta^2}{4r_s^2}\boldsymbol{W}_t\mathbb{E}_t\left[\frac{\boldsymbol{\mathcal{P}}_{t+1}}{1+c_{t+1}^2}\right]\boldsymbol{W}_t^\top\preceq C\frac{\eta^2\operatorname{tr}(\boldsymbol{\Sigma})}{r_s^2}\boldsymbol{G}_t\boldsymbol{\Sigma}\boldsymbol{G}_t.$$

For the third term, we have

$$\mathbb{E}_t\left[\frac{\nabla_{\mathrm{St}}\boldsymbol{L}_{t+1}\boldsymbol{\mathcal{P}}_{t+1}}{1+c_{t+1}^2}\right]\boldsymbol{W}_t^\top=(\boldsymbol{I}_d-\boldsymbol{G}_t)\,\mathbb{E}_t\left[\frac{y_{t+1}^3}{1+c_{t+1}^2}\|(\boldsymbol{I}_d-\boldsymbol{G}_t)\,\boldsymbol{x}_{t+1}\|_2^2\|\boldsymbol{W}_t^\top\boldsymbol{x}_{t+1}\|_2^2\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}^\top\right]\boldsymbol{G}_t.$$

Then, by using Cauchy-Schwartz inequality, we can show that

$$\left\|\mathbb{E}_t\left[\frac{y_{t+1}^3}{1+c_{t+1}^2}\|(\boldsymbol{I}_d-\boldsymbol{G}_t)\,\boldsymbol{x}_{t+1}\|_2^2\|\boldsymbol{W}_t^\top\boldsymbol{x}_{t+1}\|_2^2\boldsymbol{\Sigma}^{\frac{-1}{2}}\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}^\top\boldsymbol{\Sigma}^{\frac{-1}{2}}\right]\right\|_2\leq C\operatorname{tr}(\boldsymbol{\Sigma})\operatorname{tr}(\boldsymbol{G}_t\boldsymbol{\Sigma}).$$

Therefore,

$$\frac{\eta^3}{4r_s^3}\operatorname{sym}\left(\mathbb{E}_t\left[\frac{\nabla_{\mathrm{St}}\boldsymbol{L}_{t+1}\boldsymbol{\mathcal{P}}_{t+1}}{1+c_{t+1}^2}\right]\boldsymbol{W}_t^\top\right)\preceq C\left(\frac{\eta^4\operatorname{tr}(\boldsymbol{\Sigma})\operatorname{tr}(\boldsymbol{G}_t\boldsymbol{\Sigma})^2}{r_s^4}(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{\Sigma}(\boldsymbol{I}_d-\boldsymbol{G}_t)+\frac{\eta^2\operatorname{tr}(\boldsymbol{\Sigma})}{r_s^2}\boldsymbol{G}_t\boldsymbol{\Sigma}\boldsymbol{G}_t\right)$$

By repeating the argument for the lower bound,

$$\frac{\eta^3}{4r_s^3}\operatorname{sym}\left(\mathbb{E}_t\left[\frac{\nabla_{\mathrm{St}}\boldsymbol{L}_{t+1}\boldsymbol{\mathcal{P}}_{t+1}}{1+c_{t+1}^2}\right]\boldsymbol{W}_t^\top\right)\succeq -C\left(\frac{\eta^4\operatorname{tr}(\boldsymbol{\Sigma})\operatorname{tr}(\boldsymbol{G}_t\boldsymbol{\Sigma})^2}{r_s^4}(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{\Sigma}(\boldsymbol{I}_d-\boldsymbol{G}_t)+\frac{\eta^2\operatorname{tr}(\boldsymbol{\Sigma})}{r_s^2}\boldsymbol{G}_t\boldsymbol{\Sigma}\boldsymbol{G}_t\right).$$

For the last term, we write

$$\mathbb{E}_t\left[\frac{\nabla_{\mathrm{St}}\boldsymbol{L}_{t+1}\boldsymbol{\mathcal{P}}_{t+1}\nabla_{\mathrm{St}}\boldsymbol{L}_{t+1}^\top}{1+c_{t+1}^2}\right]=(\boldsymbol{I}_d-\boldsymbol{G}_t)\,\mathbb{E}_t\left[\frac{y_{t+1}^4}{1+c_{t+1}^2}\|\boldsymbol{W}_t^\top\boldsymbol{x}_{t+1}\|_2^4\|(\boldsymbol{I}_d-\boldsymbol{G}_t)\,\boldsymbol{x}_{t+1}\|_2^2\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}^\top\right](\boldsymbol{I}_d-\boldsymbol{G}_t).$$

We have

$$\left\|\mathbb{E}_t\Big[\frac{y_{t+1}^4}{1+c_{t+1}^2}\|\boldsymbol{W}_t^\top\boldsymbol{x}_{t+1}\|_2^4\|(\boldsymbol{I}_d-\boldsymbol{G}_t)\,\boldsymbol{x}_{t+1}\|_2^2\boldsymbol{\Sigma}^{\frac{-1}{2}}\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}^\top\boldsymbol{\Sigma}^{\frac{-1}{2}}\Big]\right\|_2\leq C\operatorname{tr}(\boldsymbol{\Sigma})\operatorname{tr}(\boldsymbol{G}_t\boldsymbol{\Sigma})^2.$$

Therefore,

$$0\preceq\frac{\eta^4}{16r_s^4}\mathbb{E}_t\left[\frac{\nabla_{\mathrm{St}}\boldsymbol{L}_{t+1}\boldsymbol{\mathcal{P}}_{t+1}\nabla_{\mathrm{St}}\boldsymbol{L}_{t+1}^\top}{1+c_{t+1}^2}\right]\preceq C\frac{\eta^4\operatorname{tr}(\boldsymbol{\Sigma})\operatorname{tr}(\boldsymbol{G}_t\boldsymbol{\Sigma})^2}{r_s^4}(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{\Sigma}(\boldsymbol{I}_d-\boldsymbol{G}_t).$$

■

D.3. Noise characterization

For $\boldsymbol{K}_1, \boldsymbol{K}_2 \in \mathbb{R}^{d\times m}$, we define

$$\mathcal{A}_{t+1}(\boldsymbol{K}_1,\boldsymbol{K}_2) \equiv \left\{ \left\| \boldsymbol{K}_1^\top (\boldsymbol{I}_d - \boldsymbol{G}_t)\boldsymbol{N}_{t+1}\boldsymbol{G}_t\boldsymbol{K}_2 \right\|_2 \le \frac{L^2}{2}\sqrt{\operatorname{tr}\left(\boldsymbol{K}_1\boldsymbol{K}_1^\top(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{\Sigma}(\boldsymbol{I}_d-\boldsymbol{G}_t)\right)\operatorname{tr}(\boldsymbol{K}_2\boldsymbol{K}_2^\top\boldsymbol{G}_t\boldsymbol{\Sigma}\boldsymbol{G}_t)} \right\}$$

$$\mathcal{B}_{t+1}(\boldsymbol{K}_1,\boldsymbol{K}_2) \equiv \left\{ \left\| \boldsymbol{K}_1^\top \boldsymbol{W}_t\boldsymbol{\mathcal{P}}_{t+1}\boldsymbol{W}_t^\top\boldsymbol{K}_2 \right\|_2 \le \frac{L^4\operatorname{tr}(\boldsymbol{\Sigma})}{2}\sqrt{\operatorname{tr}(\boldsymbol{K}_1\boldsymbol{K}_1^\top\boldsymbol{G}_t\boldsymbol{\Sigma}\boldsymbol{G}_t)\operatorname{tr}(\boldsymbol{K}_2\boldsymbol{K}_2^\top\boldsymbol{G}_t\boldsymbol{\Sigma}\boldsymbol{G}_t)} \right\}$$

$$\mathcal{C}_{t+1}(\boldsymbol{K}_1,\boldsymbol{K}_2) \equiv \left\{ \left\| \boldsymbol{K}_1^\top \nabla_{\mathrm{St}}\boldsymbol{L}_{t+1}\nabla_{\mathrm{St}}\boldsymbol{L}_{t+1}^\top\boldsymbol{K}_2 \right\|_2 \le \frac{L^4\operatorname{tr}(\boldsymbol{\Sigma})}{2}\sqrt{\prod_{i=1}^{2}\operatorname{tr}\left(\boldsymbol{K}_i\boldsymbol{K}_i^\top(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{\Sigma}(\boldsymbol{I}_d-\boldsymbol{G}_t)\right)} \right\}$$

$$\mathcal{D}_{t+1}(\boldsymbol{K}_1,\boldsymbol{K}_2) \equiv \left\{ \left\| \boldsymbol{K}_1^\top \nabla_{\mathrm{St}}\boldsymbol{L}_{t+1}\boldsymbol{\mathcal{P}}_{t+1}\boldsymbol{W}_t^\top\boldsymbol{K}_2 \right\|_2 \le \frac{L^6\operatorname{tr}(\boldsymbol{\Sigma})^2}{2}\sqrt{\operatorname{tr}\left(\boldsymbol{K}_1\boldsymbol{K}_1^\top(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{\Sigma}(\boldsymbol{I}_d-\boldsymbol{G}_t)\right)\operatorname{tr}\left(\boldsymbol{K}_2\boldsymbol{K}_2^\top\boldsymbol{G}_t\boldsymbol{\Sigma}\boldsymbol{G}_t\right)} \right\}$$

$$\mathcal{F}_{t+1}(\boldsymbol{K}_1,\boldsymbol{K}_2) \equiv \left\{ \left\| \boldsymbol{K}_1^\top \nabla_{\mathrm{St}}\boldsymbol{L}_{t+1}\boldsymbol{\mathcal{P}}_{t+1}\nabla_{\mathrm{St}}\boldsymbol{L}_{t+1}^\top\boldsymbol{K}_2 \right\|_2 \le \frac{L^8\operatorname{tr}(\boldsymbol{\Sigma})^3}{2}\sqrt{\prod_{i=1}^{2}\operatorname{tr}\left(\boldsymbol{K}_i\boldsymbol{K}_i^\top(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{\Sigma}(\boldsymbol{I}_d-\boldsymbol{G}_t)\right)} \right\}.$$

We start with the following statement:

Proposition 35 *There exists a universal constant $C>0$ such that for $L \ge e(\sqrt{8}+\sigma)$, the following holds:*

1. *We have $\mathbb{P}_t\left[\mathcal{A}_{t+1}(\boldsymbol{K}_1,\boldsymbol{K}_2)\cap\{|y_{t+1}|\le L\}\right] \ge 1-2e^{\frac{-L/e}{(\sqrt{8}+\sigma)}}$. Moreover,*

$$\begin{aligned}
&\mathbb{E}_t\left[\left(\operatorname{sym}\left(\boldsymbol{K}_1^\top(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{N}_{t+1}\boldsymbol{G}_t\boldsymbol{K}_2\right)\right)^2\right]\\
&\preceq C\operatorname{tr}(\boldsymbol{K}_2\boldsymbol{K}_2^\top\boldsymbol{G}_t\boldsymbol{\Sigma}\boldsymbol{G}_t)\boldsymbol{K}_1^\top(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{\Sigma}(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{K}_1^\top\\
&+C\operatorname{tr}\left(\boldsymbol{K}_1\boldsymbol{K}_1^\top(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{\Sigma}(\boldsymbol{I}_d-\boldsymbol{G}_t)\right)\boldsymbol{K}_2^\top\boldsymbol{G}_t\boldsymbol{\Sigma}\boldsymbol{G}_t\boldsymbol{K}_2.
\end{aligned}$$

2. *We have $\mathbb{P}_t\left[\mathcal{B}_{t+1}(\boldsymbol{K}_1,\boldsymbol{K}_2)\cap\{|y_{t+1}|\le L\}\right] \ge 1-2e^{\frac{-L/e}{(\sqrt{8}+\sigma)}}$. Moreover,*

$$\begin{aligned}
\mathbb{E}_t\left[\left(\operatorname{sym}\left(\boldsymbol{K}_1^\top\boldsymbol{W}_t\boldsymbol{\mathcal{P}}_{t+1}\boldsymbol{W}_t^\top\boldsymbol{K}_2\right)\right)^2\right] &\preceq C\operatorname{tr}(\boldsymbol{\Sigma})^2(\operatorname{tr}(\boldsymbol{K}_2\boldsymbol{K}_2^\top\boldsymbol{G}_t\boldsymbol{\Sigma}\boldsymbol{G}_t)\boldsymbol{K}_1^\top\boldsymbol{G}_t\boldsymbol{\Sigma}\boldsymbol{G}_t\boldsymbol{K}_1\\
&+C\operatorname{tr}(\boldsymbol{\Sigma})^2\operatorname{tr}(\boldsymbol{K}_1\boldsymbol{K}_1^\top\boldsymbol{G}_t\boldsymbol{\Sigma}\boldsymbol{G}_t)\boldsymbol{K}_2^\top\boldsymbol{G}_t\boldsymbol{\Sigma}\boldsymbol{G}_t\boldsymbol{K}_2.
\end{aligned}$$

3. *We have $\mathbb{P}_t\left[\mathcal{C}_{t+1}(\boldsymbol{K}_1,\boldsymbol{K}_2)\cap\{|y_{t+1}|\le L\}\right] \ge 1-2e^{\frac{-L/e}{(\sqrt{8}+\sigma)}}$. Moreover,*

$$\begin{aligned}
&\mathbb{E}_t\left[\left(\operatorname{sym}\left(\boldsymbol{K}_1^\top\nabla_{St}\boldsymbol{L}_{t+1}\nabla_{St}\boldsymbol{L}_{t+1}^\top\boldsymbol{K}_2\right)\right)^2\right]\\
&\preceq C\operatorname{tr}(\boldsymbol{\Sigma})^2\operatorname{tr}\left(\boldsymbol{K}_2\boldsymbol{K}_2^\top(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{\Sigma}(\boldsymbol{I}_d-\boldsymbol{G}_t)\right)\boldsymbol{K}_1^\top(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{\Sigma}(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{K}_1\\
&+C\operatorname{tr}(\boldsymbol{\Sigma})^2\operatorname{tr}\left(\boldsymbol{K}_1\boldsymbol{K}_1^\top(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{\Sigma}(\boldsymbol{I}_d-\boldsymbol{G}_t)\right)\boldsymbol{K}_2^\top(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{\Sigma}(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{K}_2
\end{aligned}$$

4. *We have $\mathbb{P}_t\left[\mathcal{D}_{t+1}(\boldsymbol{K}_1,\boldsymbol{K}_2)\cap\{|y_{t+1}|\le L\}\right] \ge 1-2e^{\frac{-L/e}{(7/2+\sigma)}}$. Moreover,*

$$\begin{aligned}
&\mathbb{E}_t\left[\left(\operatorname{sym}\left(\boldsymbol{K}_1^\top\nabla_{St}\boldsymbol{L}_{t+1}\boldsymbol{\mathcal{P}}_{t+1}\boldsymbol{W}_t^\top\boldsymbol{K}_2\right)\right)^2\right] \preceq C\operatorname{tr}(\boldsymbol{\Sigma})^4\operatorname{tr}(\boldsymbol{K}_2\boldsymbol{K}_2^\top\boldsymbol{G}_t\boldsymbol{\Sigma}\boldsymbol{G}_t)\boldsymbol{K}_1^\top(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{\Sigma}(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{K}_1\\
&+C\operatorname{tr}(\boldsymbol{\Sigma})^4\operatorname{tr}\left(\boldsymbol{K}_1\boldsymbol{K}_1^\top(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{\Sigma}(\boldsymbol{I}_d-\boldsymbol{G}_t)\right)\boldsymbol{K}_2^\top\boldsymbol{G}_t\boldsymbol{\Sigma}\boldsymbol{G}_t\boldsymbol{K}_2.
\end{aligned}$$

5. We have $\mathbb{P}_t\left[\mathcal{F}_{t+1}\left(\boldsymbol{K}_1, \boldsymbol{K}_2\right) \cap \{|y_{t+1}| \leq L\}\right] \geq 1 - 2e^{\frac{-L/e}{(4\sqrt{2}+\sigma)}}$. *Moreover,*

$$\begin{aligned}
&\mathbb{E}_t\Big[\big(\operatorname{sym}(\boldsymbol{K}_1^\top \nabla_{S_t} \boldsymbol{L}_{t+1} \boldsymbol{\mathcal{P}}_{t+1} \nabla_{S_t} \boldsymbol{L}_{t+1}^\top \boldsymbol{K}_2)\big)^2\Big] \\
&\preceq C\operatorname{tr}(\boldsymbol{\Sigma})^6 \operatorname{tr}\big(\boldsymbol{K}_2\boldsymbol{K}_2^\top(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{\Sigma}(\boldsymbol{I}_d-\boldsymbol{G}_t)\big)\boldsymbol{K}_1^\top(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{\Sigma}(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{K}_1 \\
&+ C\operatorname{tr}(\boldsymbol{\Sigma})^6 \operatorname{tr}\big(\boldsymbol{K}_1\boldsymbol{K}_1^\top(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{\Sigma}(\boldsymbol{I}_d-\boldsymbol{G}_t)\big)\boldsymbol{K}_2^\top(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{\Sigma}(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{K}_2.
\end{aligned}$$

Proof First, we derive a concentration bound for $|y_{t+1}|$. We have

$$\mathbb{E}_t\left[|y_{t+1}|^p\right] \leq (\sqrt{2}+\sigma/2)^p p^p \text{ for } p \geq 2,$$

which implies $\mathbb{P}_t\left[|y_{t+1}| \geq u\right] \leq e^{\frac{-u/e}{(\sqrt{8}+\sigma)}}$ for $u \geq e(\sqrt{8}+\sigma)$. In the following, we prove each item separately.

First item. We define

$$\boldsymbol{K}_1^\top(\boldsymbol{I}_d - \boldsymbol{G}_t)\boldsymbol{N}_{t+1}\boldsymbol{G}_t\boldsymbol{K}_2 = \underbrace{\boldsymbol{K}_1^\top(\boldsymbol{I}_d - \boldsymbol{G}_t)y_{t+1}\boldsymbol{x}_{t+1}}_{:=\boldsymbol{u}_{t+1}}\underbrace{\boldsymbol{x}_{t+1}^\top\boldsymbol{G}_t\boldsymbol{K}_2}_{:=\boldsymbol{v}_{t+1}^\top}.$$

For $u, L > 0$

$$\begin{aligned}
&\mathbb{P}_t\Big[\Big\|\boldsymbol{u}_{t+1}\boldsymbol{v}_{t+1}^\top\Big\|_2 \geq uL\sqrt{\operatorname{tr}\left(\boldsymbol{K}_1\boldsymbol{K}_1^\top(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{\Sigma}(\boldsymbol{I}_d-\boldsymbol{G}_t)\right)\operatorname{tr}(\boldsymbol{K}_2\boldsymbol{K}_2^\top\boldsymbol{G}_t\boldsymbol{\Sigma}\boldsymbol{G}_t)} \text{ or } |y_{t+1}| \geq L\Big] \\
&\leq \mathbb{P}_t\Big[\Big\|\mathbb{1}\{|y_{t+1}| \leq L\}\boldsymbol{u}_{t+1}\boldsymbol{v}_{t+1}^\top\Big\|_2 \geq uL\sqrt{\operatorname{tr}\left(\boldsymbol{K}_1\boldsymbol{K}_1^\top(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{\Sigma}(\boldsymbol{I}_d-\boldsymbol{G}_t)\right)\operatorname{tr}(\boldsymbol{K}_2\boldsymbol{K}_2^\top\boldsymbol{G}_t\boldsymbol{\Sigma}\boldsymbol{G}_t)}\Big] \\
&+ \mathbb{P}_t\Big[|y_{t+1}| \geq L\Big].
\end{aligned}$$

We have for $p \geq 2$

$$\begin{aligned}
\mathbb{E}_t\left[\Big\|\mathbb{1}\{|y_{t+1}| \leq L\}\boldsymbol{u}_{t+1}\boldsymbol{v}_{t+1}^\top\Big\|_2^p\right] &\leq L^p\mathbb{E}_t\left[\|\boldsymbol{K}_1^\top(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{x}_{t+1}\|_2^p\right]\mathbb{E}_t\left[\|\boldsymbol{K}_2^\top\boldsymbol{G}_t\boldsymbol{x}_{t+1}\|_2^p\right] \\
&\overset{(a)}{\leq} L^p\left(\frac{p}{2}\right)^p\left(3\operatorname{tr}\left(\boldsymbol{K}_1\boldsymbol{K}_1^\top(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{\Sigma}(\boldsymbol{I}_d-\boldsymbol{G}_t)\right)\operatorname{tr}(\boldsymbol{K}_2\boldsymbol{K}_2^\top\boldsymbol{G}_t\boldsymbol{\Sigma}\boldsymbol{G}_t)\right)^{\frac{p}{2}}.
\end{aligned}$$

We have for $u \geq 2e$

$$\mathbb{P}_t\left[\Big\|\mathbb{1}\{|e_{t+1}| \leq t\}\boldsymbol{u}_{t+1}\boldsymbol{v}_{t+1}^\top\Big\|_2 \geq uL\sqrt{\operatorname{tr}\left(\boldsymbol{K}_1\boldsymbol{K}_1^\top(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{\Sigma}(\boldsymbol{I}_d-\boldsymbol{G}_t)\right)\operatorname{tr}(\boldsymbol{K}_2\boldsymbol{K}_2^\top\boldsymbol{G}_t\boldsymbol{\Sigma}\boldsymbol{G}_t)}\right] \leq e^{-\frac{u}{e}}.$$

By choosing $u = \frac{L}{2}$, we have the probability bound.

For the variance bound, we have

$$\begin{aligned}
\mathbb{E}_t\left[\operatorname{sym}\left(\boldsymbol{u}_{t+1}\boldsymbol{v}_{t+1}^\top\right)^2\right] &\preceq \boldsymbol{K}_1^\top(\boldsymbol{I}_d-\boldsymbol{G}_t)\mathbb{E}_t\left[y_{t+1}^2\|\boldsymbol{K}_2^\top\boldsymbol{G}_t\boldsymbol{x}_{t+1}\|_2^2\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}^\top\right](\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{K}_1^\top \\
&+ \boldsymbol{K}_2^\top\boldsymbol{G}_t\mathbb{E}_t\left[y_{t+1}^2\|\boldsymbol{K}_1(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{x}_{t+1}\|_2^2\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}^\top\right]\boldsymbol{G}_t\boldsymbol{K}_2^\top \\
&\overset{(b)}{\preceq} C\operatorname{tr}(\boldsymbol{K}_2\boldsymbol{K}_2^\top\boldsymbol{G}_t\boldsymbol{\Sigma}\boldsymbol{G}_t)\boldsymbol{K}_1^\top(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{\Sigma}(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{K}_1 \\
&+ C\operatorname{tr}\Big(\boldsymbol{K}_1\boldsymbol{K}_1^\top(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{\Sigma}(\boldsymbol{I}_d-\boldsymbol{G}_t)\Big)\boldsymbol{K}_2^\top\boldsymbol{G}_t\boldsymbol{\Sigma}\boldsymbol{G}_t\boldsymbol{K}_2,
\end{aligned}$$

where we used the Cauchy-Schwartz inequality in (b).

Second item. We define

$$\boldsymbol{K}_1^\top \boldsymbol{W}_t \boldsymbol{\mathcal{P}}_{t+1} \boldsymbol{W}_t^\top \boldsymbol{K}_2 = \underbrace{y_{t+1}^2 \|(\boldsymbol{I}_d - \boldsymbol{G}_t)\,\boldsymbol{x}_{t+1}\|_2^2 \boldsymbol{K}_1^\top \boldsymbol{G}_t \boldsymbol{x}_{t+1}}_{:=\boldsymbol{u}_{t+1}} \underbrace{\boldsymbol{x}_{t+1}^\top \boldsymbol{G}_t \boldsymbol{K}_2}_{:=\boldsymbol{v}_{t+1}^\top}.$$

We have for $p \geq 2$

$$\begin{aligned}
&\mathbb{E}_t\left[\left\|\mathbb{1}\{|y_{t+1}| \leq L\}\boldsymbol{u}_{t+1}\boldsymbol{v}_{t+1}^\top\right\|_2^p\right] \\
&\leq L^{2p}\mathbb{E}_t\left[\|(\boldsymbol{I}_d - \boldsymbol{G}_t)\,\boldsymbol{x}_{t+1}\|_2^{2p}\right]\mathbb{E}_t\left[\|\boldsymbol{K}_1^\top \boldsymbol{G}_t \boldsymbol{x}_{t+1}\|_2^{2p}\right]^{\frac{1}{2}}\mathbb{E}_t\left[\|\boldsymbol{K}_2^\top \boldsymbol{G}_t \boldsymbol{x}_{t+1}\|_2^{2p}\right]^{\frac{1}{2}} \\
&\overset{(c)}{\leq} L^{2p}p^{2p}\Big(3\operatorname{tr}(\boldsymbol{\Sigma})\sqrt{\operatorname{tr}(\boldsymbol{K}_1\boldsymbol{K}_1^\top \boldsymbol{G}_t\boldsymbol{\Sigma}\boldsymbol{G}_t)\operatorname{tr}(\boldsymbol{K}_2\boldsymbol{K}_2^\top \boldsymbol{G}_t\boldsymbol{\Sigma}\boldsymbol{G}_t)}\Big)^p.
\end{aligned}$$

We have for $u \geq (2e)^2$

$$\mathbb{P}_t\Big[\Big\|\mathbb{1}\{|y_{t+1}| \leq L\}\boldsymbol{u}_{t+1}\boldsymbol{v}_{t+1}^\top\Big\|_2 \geq uL^2\operatorname{tr}(\boldsymbol{\Sigma})\sqrt{\operatorname{tr}(\boldsymbol{K}_1\boldsymbol{K}_1^\top \boldsymbol{G}_t\boldsymbol{\Sigma}\boldsymbol{G}_t)\operatorname{tr}(\boldsymbol{K}_2\boldsymbol{K}_2^\top \boldsymbol{G}_t\boldsymbol{\Sigma}\boldsymbol{G}_t)}\Big] \leq e^{-\frac{u^{1/2}}{e}}.$$

By choosing $u = \frac{L^2}{6}$, we have the probability bound. For the variance bound, we have

$$\begin{aligned}
\mathbb{E}_t\left[\operatorname{sym}\left(\boldsymbol{u}_{t+1}\boldsymbol{v}_{t+1}^\top\right)^2\right] &\preceq \boldsymbol{K}_1^\top \boldsymbol{G}_t \mathbb{E}_t\left[y_{t+1}^4\|(\boldsymbol{I}_d - \boldsymbol{G}_t)\,\boldsymbol{x}_{t+1}\|_2^4\|\boldsymbol{K}_2^\top \boldsymbol{G}_t\boldsymbol{x}_{t+1}\|_2^2\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}^\top\right]\boldsymbol{G}_t\boldsymbol{K}_1 \\
&\quad + \boldsymbol{K}_2^\top \boldsymbol{G}_t \mathbb{E}_t\left[y_{t+1}^4\|(\boldsymbol{I}_d - \boldsymbol{G}_t)\,\boldsymbol{x}_{t+1}\|_2^4\|\boldsymbol{K}_1^\top \boldsymbol{G}_t\boldsymbol{x}_{t+1}\|_2^2\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}^\top\right]\boldsymbol{G}_t\boldsymbol{K}_2 \\
&\overset{(d)}{\preceq} C\operatorname{tr}(\boldsymbol{\Sigma})^2\left(\operatorname{tr}(\boldsymbol{K}_2\boldsymbol{K}_2^\top \boldsymbol{G}_t\boldsymbol{\Sigma}\boldsymbol{G}_t)\boldsymbol{K}_1^\top \boldsymbol{G}_t\boldsymbol{\Sigma}\boldsymbol{G}_t\boldsymbol{K}_1 + \operatorname{tr}(\boldsymbol{K}_1\boldsymbol{K}_1^\top \boldsymbol{G}_t\boldsymbol{\Sigma}\boldsymbol{G}_t)\boldsymbol{K}_2^\top \boldsymbol{G}_t\boldsymbol{\Sigma}\boldsymbol{G}_t\boldsymbol{K}_2\right),
\end{aligned}$$

where we use the Cauchy-Schwartz inequality in (d).

Third item. We define

$$\boldsymbol{K}_1^\top \nabla_{\mathrm{St}}\boldsymbol{L}_{t+1}\nabla_{\mathrm{St}}\boldsymbol{L}_{t+1}^\top \boldsymbol{K}_2 = \underbrace{y_{t+1}^2\|\boldsymbol{W}_t^\top \boldsymbol{x}_{t+1}\|_2^2\boldsymbol{K}_1^\top(\boldsymbol{I}_d - \boldsymbol{G}_t)\boldsymbol{x}_{t+1}}_{:=\boldsymbol{u}_{t+1}}\underbrace{\boldsymbol{x}_{t+1}^\top(\boldsymbol{I}_d - \boldsymbol{G}_t)\boldsymbol{K}_2}_{:=\boldsymbol{v}_{t+1}^\top}.$$

We have for $p \geq 2$

$$\begin{aligned}
&\mathbb{E}_t\left[\left\|\mathbb{1}\{|y_{t+1}| \leq L\}\boldsymbol{u}_{t+1}\boldsymbol{v}_{t+1}^\top\right\|_2^p\right] \\
&\leq L^{2p}\mathbb{E}_t\left[\|\boldsymbol{W}_t^\top \boldsymbol{x}_{t+1}\|_2^{2p}\right]\mathbb{E}_t\left[\|\boldsymbol{K}_1^\top(\boldsymbol{I}_d - \boldsymbol{G}_t)\boldsymbol{x}_{t+1}\|_2^{2p}\right]^{\frac{1}{2}}\mathbb{E}_t\left[\|\boldsymbol{K}_2^\top(\boldsymbol{I}_d - \boldsymbol{G}_t)\boldsymbol{x}_{t+1}\|_2^{2p}\right]^{\frac{1}{2}} \\
&\overset{(e)}{\leq} L^{2p}p^{2p}\left(3r_s\sqrt{\operatorname{tr}\left(\boldsymbol{K}_1\boldsymbol{K}_1^\top(\boldsymbol{I}_d - \boldsymbol{G}_t)\boldsymbol{\Sigma}(\boldsymbol{I}_d - \boldsymbol{G}_t)\right)\operatorname{tr}\left(\boldsymbol{K}_2\boldsymbol{K}_2^\top(\boldsymbol{I}_d - \boldsymbol{G}_t)\boldsymbol{\Sigma}(\boldsymbol{I}_d - \boldsymbol{G}_t)\right)}\right)^p,
\end{aligned}$$

We have for $u \geq (2e)^2$

$$\begin{aligned}
&\mathbb{P}_t\Big[\Big\|\mathbb{1}\{|y_{t+1}| \leq L\}\boldsymbol{u}_{t+1}\boldsymbol{v}_{t+1}^\top\Big\|_2 \\
&\qquad \geq uL^2 3r_s\sqrt{\operatorname{tr}\left(\boldsymbol{K}_1\boldsymbol{K}_1^\top(\boldsymbol{I}_d - \boldsymbol{G}_t)\boldsymbol{\Sigma}(\boldsymbol{I}_d - \boldsymbol{G}_t)\right)\operatorname{tr}\left(\boldsymbol{K}_2\boldsymbol{K}_2^\top(\boldsymbol{I}_d - \boldsymbol{G}_t)\boldsymbol{\Sigma}(\boldsymbol{I}_d - \boldsymbol{G}_t)\right)}\Big] \leq e^{-\frac{u^{1/2}}{e}}.
\end{aligned}$$

By choosing $u = \frac{L^2}{6}$, we have the probability bound. For the variance bound, we have

$$\begin{aligned}
&\mathbb{E}_t\left[\mathrm{sym}\left(\boldsymbol{u}_{t+1}\boldsymbol{v}_{t+1}^\top\right)^2\right] \\
&\preceq \boldsymbol{K}_1^\top(\boldsymbol{I}_d-\boldsymbol{G}_t)\,\mathbb{E}_t\left[y_{t+1}^4\|\boldsymbol{W}_t^\top\boldsymbol{x}_{t+1}\|_2^4\|\boldsymbol{K}_2(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{x}_{t+1}\|_2^2\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}^\top\right](\boldsymbol{I}_d-\boldsymbol{G}_t)\,\boldsymbol{K}_1 \\
&+ \boldsymbol{K}_2^\top(\boldsymbol{I}_d-\boldsymbol{G}_t)\,\mathbb{E}_t\left[y_{t+1}^4\|\boldsymbol{W}_t^\top\boldsymbol{x}_{t+1}\|_2^4\|\boldsymbol{K}_1(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{x}_{t+1}\|_2^2\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}^\top\right](\boldsymbol{I}_d-\boldsymbol{G}_t)\,\boldsymbol{K}_2 \\
&\overset{(f)}{\preceq} Cr_s^2\,\mathrm{tr}\left(\boldsymbol{K}_2\boldsymbol{K}_2^\top(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{\Sigma}(\boldsymbol{I}_d-\boldsymbol{G}_t)\right)\boldsymbol{K}_1^\top(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{\Sigma}(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{K}_1 \\
&+ Cr_s^2\,\mathrm{tr}\left(\boldsymbol{K}_1\boldsymbol{K}_1^\top(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{\Sigma}(\boldsymbol{I}_d-\boldsymbol{G}_t)\right)\boldsymbol{K}_2^\top(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{\Sigma}(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{K}_2,
\end{aligned}$$

where we used the Cauchy-Schwartz inequality in (f).

Fourth item. We define

$$\boldsymbol{K}_1^\top\nabla_{\mathrm{St}}\boldsymbol{L}_{t+1}\boldsymbol{\mathcal{P}}_{t+1}\boldsymbol{W}_t^\top\boldsymbol{K}_2 = \underbrace{y_{t+1}^3\|(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{x}_{t+1}\|_2^2\|\boldsymbol{W}_t^\top\boldsymbol{x}_{t+1}\|_2^2\boldsymbol{K}_1^\top(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{x}_{t+1}}_{:=\boldsymbol{u}_{t+1}}\underbrace{\boldsymbol{x}_{t+1}^\top\boldsymbol{G}_t\boldsymbol{K}_2}_{:=\boldsymbol{v}_{t+1}^\top}.$$

We have for $p \geq 2$

$$\begin{aligned}
&\mathbb{E}_t\left[\left\|\mathbb{1}\{|y_{t+1}|\leq L\}\boldsymbol{u}_{t+1}\boldsymbol{v}_{t+1}^\top\right\|_2^p\right] \\
&\leq L^{3p}\mathbb{E}_t\left[\|(\boldsymbol{I}_d-\boldsymbol{G}_t)\,\boldsymbol{x}_{t+1}\|_2^{2p}\|\boldsymbol{K}_1^\top(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{x}_{t+1}\|_2^p\right]\mathbb{E}_t\left[\|\boldsymbol{W}_t^\top\boldsymbol{x}_{t+1}\|_2^{2p}\|\boldsymbol{K}_2^\top\boldsymbol{G}_t\boldsymbol{x}_{t+1}\|_2^p\right] \\
&\leq L^{3p}(12\sqrt{3})^p p^{3p}\left(\mathrm{tr}(\boldsymbol{\Sigma})r_s\sqrt{\mathrm{tr}\left(\boldsymbol{K}_1\boldsymbol{K}_1^\top(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{\Sigma}(\boldsymbol{I}_d-\boldsymbol{G}_t)\right)\mathrm{tr}\left(\boldsymbol{K}_2\boldsymbol{K}_2^\top\boldsymbol{G}_t\boldsymbol{\Sigma}\boldsymbol{G}_t\right)}\right)^p.
\end{aligned}$$

We have for $u \geq (2e)^3$

$$\begin{aligned}
&\mathbb{P}_t\left[\left\|\mathbb{1}\{|t_{t+1}|\leq L\}\boldsymbol{u}_{t+1}\boldsymbol{v}_{t+1}^\top\right\|_2 \geq uL^3 12\sqrt{3}\,\mathrm{tr}(\boldsymbol{\Sigma})r_s\sqrt{\mathrm{tr}\left(\boldsymbol{K}_1\boldsymbol{K}_1^\top(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{\Sigma}(\boldsymbol{I}_d-\boldsymbol{G}_t)\right)\mathrm{tr}\left(\boldsymbol{K}_2\boldsymbol{K}_2^\top\boldsymbol{G}_t\boldsymbol{\Sigma}\boldsymbol{G}_t\right)}\right] \\
&\qquad\qquad \leq e^{-\frac{u^{1/3}}{e}}.
\end{aligned}$$

By choosing $u = \frac{L^3}{24\sqrt{3}}$, we have the probability bound. For the variance bound, we have

$$\begin{aligned}
&\mathbb{E}_t\left[\mathrm{sym}\left(\boldsymbol{u}_{t+1}\boldsymbol{v}_{t+1}^\top\right)^2\right] \\
&\preceq \boldsymbol{K}_1^\top(\boldsymbol{I}_d-\boldsymbol{G}_t)\mathbb{E}_t\left[y_{t+1}^6\|(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{x}_{t+1}\|_2^4\|\boldsymbol{W}_t^\top\boldsymbol{x}_{t+1}\|_2^4\|\boldsymbol{K}_2^\top\boldsymbol{G}_t\boldsymbol{x}_{t+1}\|_2^2\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}^\top\right](\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{K}_1 \\
&+ \boldsymbol{K}_2^\top\boldsymbol{G}_t\mathbb{E}_t\left[y_{t+1}^6\|(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{x}_{t+1}\|_2^4\|\boldsymbol{W}_t^\top\boldsymbol{x}_{t+1}\|_2^4\|\boldsymbol{K}_1^\top(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{x}_{t+1}\|_2^2\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}^\top\right]\boldsymbol{G}_t\boldsymbol{K}_2 \\
&\preceq C\,\mathrm{tr}(\boldsymbol{\Sigma})^2 r_s^2\,\mathrm{tr}(\boldsymbol{K}_2\boldsymbol{K}_2^\top\boldsymbol{G}_t\boldsymbol{\Sigma}\boldsymbol{G}_t)\boldsymbol{K}_1^\top(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{\Sigma}(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{K}_1 \\
&+ C\,\mathrm{tr}(\boldsymbol{\Sigma})^2 r_s^2\,\mathrm{tr}\left(\boldsymbol{K}_1\boldsymbol{K}_1^\top(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{\Sigma}(\boldsymbol{I}_d-\boldsymbol{G}_t)\right)\boldsymbol{K}_2^\top\boldsymbol{G}_t\boldsymbol{\Sigma}\boldsymbol{G}_t\boldsymbol{K}_2.
\end{aligned}$$

Fifth item. We define

$$\boldsymbol{K}_1^\top\nabla_{\mathrm{St}}\boldsymbol{L}_{t+1}\boldsymbol{\mathcal{P}}_{t+1}\nabla_{\mathrm{St}}\boldsymbol{L}_{t+1}^\top\boldsymbol{K}_2 = \underbrace{y_{t+1}^4\|(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{x}_{t+1}\|_2^2\|\boldsymbol{W}_t^\top\boldsymbol{x}_{t+1}\|_2^4\boldsymbol{K}_1^\top(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{x}_{t+1}}_{:=\boldsymbol{u}_{t+1}}\underbrace{\boldsymbol{x}_{t+1}^\top(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{K}_2}_{:=\boldsymbol{v}_{t+1}^\top}.$$

We have for $p \geq 2$

$$\begin{aligned}
&\mathbb{E}_t\left[\left\|\mathbb{1}\{|y_{t+1}| \leq L\}\boldsymbol{u}_{t+1}\boldsymbol{v}_{t+1}^\top\right\|_2^p\right] \\
&\leq L^{4p}\mathbb{E}_t\Big[\|(\boldsymbol{I}_d-\boldsymbol{G}_t)\,\boldsymbol{x}_{t+1}\|_2^{2p}\|\boldsymbol{W}_t^\top\boldsymbol{x}_{t+1}\|_2^{4p}\|\boldsymbol{K}_1^\top(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{x}_{t+1}\|_2^p\|\boldsymbol{K}_2^\top(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{x}_{t+1}\|_2^p\Big] \\
&\leq L^{4p}\mathbb{E}_t\Big[\|(\boldsymbol{I}_d-\boldsymbol{G}_t)\,\boldsymbol{x}_{t+1}\|_2^{4p}\Big]^{\frac{1}{2}}\mathbb{E}_t\Big[\|\boldsymbol{K}_1^\top(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{x}_{t+1}\|_2^{4p}\Big]^{\frac{1}{4}}\mathbb{E}_t\Big[\|\boldsymbol{K}_2^\top(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{x}_{t+1}\|_2^{4p}\Big]^{\frac{1}{4}}\mathbb{E}_t\Big[\|\boldsymbol{W}_t^\top\boldsymbol{x}_{t+1}\|_2^{4p}\Big] \\
&\leq L^{4p}(16p^4)^p\big(3\sqrt{3}r_s^2\operatorname{tr}(\boldsymbol{\Sigma})\big)^p\Big(3\operatorname{tr}\big(\boldsymbol{K}_1\boldsymbol{K}_1^\top(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{\Sigma}(\boldsymbol{I}_d-\boldsymbol{G}_t)\big)\operatorname{tr}\big(\boldsymbol{K}_2\boldsymbol{K}_2^\top(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{\Sigma}(\boldsymbol{I}_d-\boldsymbol{G}_t)\big)\Big)^{\frac{p}{2}} \\
&= L^{4p}(2\sqrt{3})^{4p}p^{4p}\left(\operatorname{tr}(\boldsymbol{\Sigma})r_s^2\sqrt{\operatorname{tr}\big(\boldsymbol{K}_1\boldsymbol{K}_1^\top(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{\Sigma}(\boldsymbol{I}_d-\boldsymbol{G}_t)\big)\operatorname{tr}\big(\boldsymbol{K}_2\boldsymbol{K}_2^\top(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{\Sigma}(\boldsymbol{I}_d-\boldsymbol{G}_t)\big)}\right)^p.
\end{aligned}$$

We have for $u \geq (2e)^4$

$$\begin{aligned}
&\mathbb{P}_t\Big[\Big\|\mathbb{1}\{|y_{t+1}|\leq L\}\boldsymbol{u}_{t+1}\boldsymbol{v}_{t+1}^\top\Big\|_2 \geq \\
&\qquad u(2\sqrt{3}L)^4\operatorname{tr}(\boldsymbol{\Sigma})r_s^2\sqrt{\operatorname{tr}\big(\boldsymbol{K}_1\boldsymbol{K}_1^\top(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{\Sigma}(\boldsymbol{I}_d-\boldsymbol{G}_t)\big)\operatorname{tr}\big(\boldsymbol{K}_2\boldsymbol{K}_2^\top(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{\Sigma}(\boldsymbol{I}_d-\boldsymbol{G}_t)\big)}\Big] \leq e^{-\frac{u^{1/4}}{e}}.
\end{aligned}$$

By choosing $u = \frac{L^4}{2(2\sqrt{3})^4}$, we have the probability bound. For the variance bound, we have

$$\begin{aligned}
&\mathbb{E}_t\left[\operatorname{sym}\left(\boldsymbol{u}_{t+1}\boldsymbol{v}_{t+1}^\top\right)^2\right] \\
&\preceq \boldsymbol{K}_1^\top(\boldsymbol{I}_d-\boldsymbol{G}_t)\mathbb{E}_t\left[y_{t+1}^8\|(\boldsymbol{I}_d-\boldsymbol{G}_t)\,\boldsymbol{x}_{t+1}\|_2^4\|\boldsymbol{W}_t^\top\boldsymbol{x}_{t+1}\|_2^8\|\boldsymbol{K}_2(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{x}_{t+1}\|_2^2\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}^\top\right](\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{K}_1 \\
&+ \boldsymbol{K}_2^\top(\boldsymbol{I}_d-\boldsymbol{G}_t)\mathbb{E}_t\left[y_{t+1}^8\|(\boldsymbol{I}_d-\boldsymbol{G}_t)\,\boldsymbol{x}_{t+1}\|_2^4\|\boldsymbol{W}_t^\top\boldsymbol{x}_{t+1}\|_2^8\|\boldsymbol{K}_1(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{x}_{t+1}\|_2^2\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}^\top\right](\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{K}_2 \\
&\preceq C\operatorname{tr}(\boldsymbol{\Sigma})^2r_s^4\operatorname{tr}\big(\boldsymbol{K}_2\boldsymbol{K}_2^\top(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{\Sigma}(\boldsymbol{I}_d-\boldsymbol{G}_t)\big)\boldsymbol{K}_1^\top(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{\Sigma}(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{K}_1 \\
&+ C\operatorname{tr}(\boldsymbol{\Sigma})^2r_s^4\operatorname{tr}\big(\boldsymbol{K}_1\boldsymbol{K}_1^\top(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{\Sigma}(\boldsymbol{I}_d-\boldsymbol{G}_t)\big)\boldsymbol{K}_2^\top(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{\Sigma}(\boldsymbol{I}_d-\boldsymbol{G}_t)\boldsymbol{K}_2.
\end{aligned}$$

■

D.3.1. APPLICATION OF NOISE BOUNDS

We will use $r_u = \log^{\frac{1}{\alpha+\beta}} d$. For $\eta \ll \frac{r_s}{d\operatorname{tr}(\boldsymbol{\Sigma})}$ and $r_s \gg \frac{d}{\operatorname{polylog} d}$, we can show that $\underline{\boldsymbol{M}}_t$ and $\bar{\boldsymbol{M}}_t$ have asymptotically same dynamics with

$$\mathsf{M}_{t+1} = \mathsf{M}_t + 2\eta(\boldsymbol{\Lambda}_1\boldsymbol{\Sigma}_{11}\mathsf{M}_t - \boldsymbol{\Lambda}_1\mathsf{M}_t^2), \;\; \text{with} \;\; \mathsf{M}_0 = \frac{r_s}{d}\boldsymbol{\Sigma}_{11}.$$

We start with the following proposition:

Proposition 36 *Under $r_u = \log^{\frac{1}{\alpha+\beta}} d$ and $r_s \gg \frac{d}{\operatorname{polylog} d}$, we have*

- $\mathsf{M}_t\boldsymbol{R}_t^{-1} = (1+o_d(1))\kappa_d$
- $(\eta\sum_{t=0}^{t-1}\mathsf{M}_s)\mathsf{M}_t^{-1} \preceq \min\{\eta t\boldsymbol{I}_{r_u}, \big(2\boldsymbol{\Lambda}_1(\boldsymbol{\Sigma}_{11}-\mathsf{M}_t)\big)^{-1}\}$

By recalling the definitions $\{\boldsymbol{R}_t\}_{t\in\mathbb{N}}$, $\boldsymbol{\Lambda}_1$, $\boldsymbol{\Sigma}_{11}$, we define the event:

$$\mathcal{A}_{t+1} \coloneqq \mathcal{A}_{t+1}\left(\boldsymbol{\Sigma}_{11}^{\frac{1}{2}}\boldsymbol{R}_t^{\frac{-1}{2}}, \boldsymbol{\Sigma}_{11}^{\frac{1}{2}}\boldsymbol{R}_t^{\frac{-1}{2}}\right) \cap \mathcal{A}_{t+1}\left(\boldsymbol{\Sigma}_{11}\boldsymbol{\Lambda}_1^{\frac{1}{2}}\boldsymbol{R}_t^{\frac{-1}{2}}, \boldsymbol{R}_t^{\frac{-1}{2}}\boldsymbol{\Lambda}_1^{\frac{1}{2}}\boldsymbol{\Sigma}_{11}\right) \cap \{|y_{t+1}| \le L\}$$

We define the events $\mathcal{B}_{t+1}$, $\mathcal{C}_{t+1}$, $\mathcal{D}_{t+1}$, and $\mathcal{F}_{t+1}$ in the same way. Based on these events, we define the clipped versions of the noise matrices:

$$\begin{aligned}
\mathsf{A}_{t+1} &\coloneqq \operatorname{sym}\Big((\boldsymbol{I}_d - \boldsymbol{G}_t)\Big(y_{t+1}\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}^\top \mathbb{1}\{\mathcal{A}_{t+1}\} - \mathbb{E}_t\big[y_{t+1}\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}^\top \mathbb{1}\{\mathcal{A}_{t+1}\}\big]\Big)\boldsymbol{G}_t\Big)\\
\mathsf{B}_{t+1} &\coloneqq \boldsymbol{W}_t\left(\frac{\boldsymbol{\mathcal{P}}_{t+1}\mathbb{1}\{\mathcal{B}_{t+1}\}}{1+c_{t+1}^2} - \mathbb{E}_t\left[\frac{\boldsymbol{\mathcal{P}}_{t+1}\mathbb{1}\{\mathcal{B}_{t+1}\}}{1+c_{t+1}^2}\right]\right)\boldsymbol{W}_t^\top\\
\mathsf{C}_{t+1} &\coloneqq \Big(\nabla_{\mathrm{St}}\boldsymbol{L}_{t+1}\nabla_{\mathrm{St}}\boldsymbol{L}_{t+1}^\top \mathbb{1}\{\mathcal{C}_{t+1}\} - \mathbb{E}_t\left[\nabla_{\mathrm{St}}\boldsymbol{L}_{t+1}\nabla_{\mathrm{St}}\boldsymbol{L}_{t+1}^\top \mathbb{1}\{\mathcal{C}_{t+1}\}\right]\Big)\\
\mathsf{D}_{t+1} &\coloneqq \operatorname{sym}\left(\left(\frac{\nabla_{\mathrm{St}}\boldsymbol{L}_{t+1}\boldsymbol{\mathcal{P}}_{t+1}\mathbb{1}\{\mathcal{D}_{t+1}\}}{1+c_{t+1}^2} - \mathbb{E}_t\left[\frac{\nabla_{\mathrm{St}}\boldsymbol{L}_{t+1}\boldsymbol{\mathcal{P}}_{t+1}\mathbb{1}\{\mathcal{D}_{t+1}\}}{1+c_{t+1}^2}\right]\right)\boldsymbol{W}_t^\top\right)\\
\mathsf{F}_{t+1} &\coloneqq \left(\frac{\nabla_{\mathrm{St}}\boldsymbol{L}_{t+1}\boldsymbol{\mathcal{P}}_{t+1}\nabla_{\mathrm{St}}\boldsymbol{L}_{t+1}^\top \mathbb{1}\{\mathcal{F}_{t+1}\}}{1+c_{t+1}^2} - \mathbb{E}_t\left[\frac{\nabla_{\mathrm{St}}\boldsymbol{L}_{t+1}\boldsymbol{\mathcal{P}}_{t+1}\nabla_{\mathrm{St}}\boldsymbol{L}_{t+1}^\top \mathbb{1}\{\mathcal{F}_{t+1}\}}{1+c_{t+1}^2}\right]\right).
\end{aligned}$$

Let $\mathsf{X} \in \left\{\frac{\eta}{r_s}\mathsf{A}, \frac{\eta^2}{4r_s^2}\mathsf{B}, \frac{\eta^2}{4r_s^2}\mathsf{C}, \frac{\eta^3}{4r_s^3}\mathsf{D}, \frac{\eta^4}{16r_s^4}\mathsf{F}\right\}$ and

$$\Gamma_1 \coloneqq \boldsymbol{I}_{r_u}, \qquad \Gamma_2 \coloneqq \boldsymbol{\Lambda}_1^{\frac{1}{2}}\boldsymbol{\Sigma}_{11}^{\frac{1}{2}}$$

For $\ell \in \{1,2\}$, we define:

$$\mathrm{Quad}_{k,t}^{(\ell)}(\mathsf{X}) \coloneqq \sum_{j=1}^{k}\mathbb{E}_{j-1}\left[\left(\Gamma_\ell\boldsymbol{\Sigma}_{11}^{\frac{1}{2}}\boldsymbol{R}_t^{\frac{-1}{2}}\mathsf{X}_j\boldsymbol{R}_t^{\frac{-1}{2}}\boldsymbol{\Sigma}_{11}^{\frac{1}{2}}\Gamma_\ell\right)^2\right].$$

We have the following corollary.

Corollary 37 *Let*

$$r_u = \log^{\frac{1}{\alpha+\beta}} d, \qquad \eta \ll \frac{r_s}{d\,\mathrm{tr}(\boldsymbol{\Sigma})}, \qquad r_s \gg \frac{d}{\mathrm{polylog}\, d}.$$

For $\eta t \le \frac{1}{2}\log d$, we have:

$$\|\mathit{Quad}_{t,t}^{(\ell)}(\mathsf{X})\|_2 \le \frac{C\log d}{\kappa_d^2}\frac{dr_u}{r_s}\begin{cases}\eta & \mathsf{X} = \frac{\eta}{r_s}\mathsf{A}\\ \eta^3\,\mathrm{tr}(\boldsymbol{\Sigma})^2, & \mathsf{X} = \frac{\eta^2}{4r_s^2}\mathsf{B}\\ \eta^3\frac{d\log d\,\mathrm{tr}(\boldsymbol{\Sigma})^2}{r_s}, & \mathsf{X} = \frac{\eta^2}{4r_s^2}\mathsf{C}\\ \eta^5\,\mathrm{tr}(\boldsymbol{\Sigma})^4, & \mathsf{X} = \frac{\eta^3}{16r_s^3}\mathsf{D}\\ \eta^7\frac{d\log \mathrm{tr}(\boldsymbol{\Sigma})^6}{r_s}, & \mathsf{X} = \frac{\eta^4}{16r_s^4}\mathsf{F}.\end{cases}$$

Proof By using the bounds in Proposition 35, the result can be proven. ■

By using matrix Bernstein inequality similar to [3, Propositions 21 and 22], we can derive the result.

D.4. Alignment step

As detailed in [3, Section E], by using the parameterization $\boldsymbol{M} = \boldsymbol{\Omega}\boldsymbol{\Omega}^\top$, we can reduce the alignment step to a linear regression problem over the space of $r_s \times r_s$ symmetric matrices, where the ℓ_2 regularization on $\boldsymbol{\Omega}$ translates to nuclear norm regularization on $\boldsymbol{M}$. The problem then reduces to finding the best sparse approximation of the labels via ℓ_1 regularization. Because $\boldsymbol{W}_t$ aligns with the top eigenvector of $\boldsymbol{\Sigma}$, we can show that with an appropriate regularization parameter, $T_{\text{align}} = \text{polylog}(d)$ samples are sufficient to drive the risk $\mathcal{R}(t)$ arbitrarily small (at a rate of $\mathcal{O}(1/\log d)$). Computationally, we can solve this Lasso objective up to a $\mathcal{O}(1/\log d)$ error tolerance with a cost of $\mathcal{O}(r_s^2 \,\text{polylog}\, d)$. This overhead is negligible compared to the first phase of Algorithm 1, which requires a matrix orthogonalization (e.g., QR decomposition) at every iteration.

Appendix E. Some moment bounds and concentration inequalities

Lemma 38 (Hypercontractivity) *Let $P_k : \boldsymbol{R}^d \to \boldsymbol{R}$ be a polynomial of degree-k and $\boldsymbol{x} \sim \mathcal{N}(0, \boldsymbol{I}_d)$. For $q \geq 2$, we have $\mathbb{E}\left[P_k(\boldsymbol{x})^q\right]^{1/q} \leq (q-1)^{k/2}\mathbb{E}\left[P_k(\boldsymbol{x})^2\right]^{1/2}$.*

Lemma 39 *Let $\boldsymbol{x} \sim \mathcal{N}(0, \boldsymbol{I}_d)$ and $\boldsymbol{S} \in \boldsymbol{R}^{d\times d}$ be a symmetric matrix. For $u > 0$,*

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

Proof We note that $\boldsymbol{x}^\top \boldsymbol{S}\boldsymbol{x} - \operatorname{tr}(\boldsymbol{S})$ has the same distribution with $\sum_{i=1}^d \lambda_i(\boldsymbol{S})(Z_i^2 - 1)$, where $Z_i \sim_{iid} \mathcal{N}(0,1)$. By using the Laurent-Massart lemma [14], we have the result. ■

Corollary 40 *Let $y = \boldsymbol{x}^\top \boldsymbol{S}\boldsymbol{x} - \operatorname{tr}(\boldsymbol{S})$ and $\boldsymbol{x} \sim \mathcal{N}(0, \boldsymbol{I}_d)$. For $p \geq 2$, we have $\mathbb{E}[|y|^p]^{\frac{1}{p}} \leq (p-1)\sqrt{2}\|\boldsymbol{S}\|_F$.*

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

Corollary 41 *For $\boldsymbol{A} \in \boldsymbol{R}^{d\times r}$, $p \geq 2$ and $\boldsymbol{x} \sim \mathcal{N}(0, \boldsymbol{I}_d)$, we have $\mathbb{E}[\|\boldsymbol{A}^\top \boldsymbol{x}\|_2^{2p}]^{\frac{1}{p}} \leq \sqrt{3}(p-1)\operatorname{tr}(\boldsymbol{A}^\top \boldsymbol{A})$.*

Proof By Lemma 38, we have $\mathbb{E}[\|\boldsymbol{A}^\top \boldsymbol{x}\|_2^{2p}]^{\frac{1}{p}} \leq (p-1)\mathbb{E}[\|\boldsymbol{A}^\top \boldsymbol{x}\|_2^4]^{\frac{1}{2}}$. For $\boldsymbol{S} = \boldsymbol{A}\boldsymbol{A}^\top$, we have

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

We have

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

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