ON THE DOUBLE DESCENT OF RANDOM FEATURES MODELS TRAINED WITH SGD

Anonymous authors
Paper under double-blind review

ABSTRACT

We study generalization properties of random features (RF) regression in high dimensions optimized by stochastic gradient descent (SGD). In this regime, we derive precise non-asymptotic error bounds of RF regression under both constant and adaptive step-size SGD setting, and observe the double descent phenomenon both theoretically and empirically. Our analysis shows how to cope with multiple randomness sources of initialization, label noise, and data sampling (as well as stochastic gradients) with no closed-form solution, and also goes beyond the commonly-used Gaussian/spherical data assumption. Our theoretical results demonstrate that, with SGD training, RF regression still generalizes well in the interpolation setting, and is able to characterize the double descent behavior by the unimodality of variance and monotonic decrease of bias. Besides, we also prove that the constant step-size SGD setting incurs no loss in convergence rate when compared to the exact minimal-norm interpolator, as a theoretical justification of using SGD in practice.

1 INTRODUCTION

Harmless interpolation or benign overfitting of over-parameterized neural network (NN) models has received significant attention in the literature (Zhang et al., 2016; Hastie et al., 2019; Bartlett et al., 2020). This important phenomenon is also inherently tied to the discovery of the *double descent* learning curve (Belkin et al., 2019) in deep learning.

Indeed, many key machine learning models, including but not limited to kernel regression (Wu & Xu, 2020; Mei & Montanari, 2019; Liu et al., 2021b) and neural networks (Nakkiran et al., 2019; Yang et al., 2020; Ju et al., 2020)) first decrease the test error with increasing number of model parameters in the under-parameterized regime. They then yield large error when they can first interpolate the data, which is called the interpolation threshold. Finally, the test error begins to decrease again in the over-parameterized regime in stark contrast to the conventional learning theory results.

Our work partakes in this research vein and studies the random features (RF) model (Rahimi & Recht, 2007) in the context of double descent phenomenon.[1] Briefly, RF model samples random features $\{\boldsymbol{\omega}_i\}_{i=1}^m$ from a specific distribution, corresponding to a kernel function. We then construct an explicit map: $\boldsymbol{x} \in \mathbb{R}^d \mapsto \sigma(\boldsymbol{W}\boldsymbol{x}) \in \mathbb{R}^m$, where $\boldsymbol{W} = [\boldsymbol{\omega}_1, \cdots, \boldsymbol{\omega}_m]^\top \in \mathbb{R}^{m \times d}$ is the random features matrix and $\sigma(\cdot)$ is the nonlinear (activation) function determined by the kernel. As a result, the RF model training can be viewed as training a two-layer neural network where the weights in the first layer are chosen randomly and then fixed (a.k.a. the random features) and only the output layer is optimized, striking a trade-off between practical performance and accessibility to analysis.

An RF model becomes an over-parameterized model if we take the number of random features m larger than that of training data n. The literature on RF under the over-parameterized regime can be split into various camps according to different assumptions on the formulation of target function, data distribution, and activation functions (Mei & Montanari, 2019; Ba et al., 2020; d'Ascoli et al., 2020b; Liao et al., 2020; Gerace et al., 2020; Lin & Dobriban, 2021). The existing theoretical results demonstrate that the excess risk curve exhibits double descent.

Nevertheless, the analysis framework of (most) previous work on RF regression assumes the data to be Gaussian or uniformly spread on a sphere, and largely relies on the least-squares closed-form solution, including *minimal-norm* interpolator and ridge regressor (see comparisons in Table 1 in Appendix A). Such specific data distribution and dependency on the closed-form solution in fact mismatch practical neural networks optimized by stochastic gradient descent (SGD) based algorithms on general data distribution.

Our work precisely bridges this gap: We provide a new analysis framework for the generalization properties of RF models trained with SGD, also accommodating adaptive step-size selection, and provide non-asymptotic results in

[1] Relevant work on random features model has received the Test-of-Time Award in NeurIPS2017 (Rahimi & Recht, 2007) and best paper finalist in ICML2019 (Li et al., 2019). Refer to a recent survey (Liu et al., 2021a) for details.

under-/over-parameterized regimes on (relatively) general data distribution and activation functions. We make the following findings and contributions:

- We characterize statistical properties of several covariance operators/matrices in RF, including $\Sigma_m := \frac{1}{m}\mathbb{E}_{\boldsymbol{x}}[\sigma(\boldsymbol{W}\boldsymbol{x}/\sqrt{d})\sigma(\boldsymbol{W}\boldsymbol{x}/\sqrt{d})^\top]$ and its expectation version $\widetilde{\Sigma}_m := \mathbb{E}_{\boldsymbol{W}}[\Sigma_m]$. We demonstrate that, under Gaussian initialization, $\mathrm{Tr}(\Sigma_m)$ is a sub-exponential random variable with $\mathcal{O}(1)$ sub-exponential norm; $\widetilde{\Sigma}_m$ has only two distinct eigenvalues at $\mathcal{O}(1)$ and $\mathcal{O}(1/m)$ order, respectively. Such analysis on the spectra of Σ_m and $\widetilde{\Sigma}_m$ (without spectral decay assumption) is helpful to obtain sharp error bounds for excess risk.
- Based on the bias-variance decomposition in stochastic approximation, we further take into account multiple randomness sources of initialization, label noise, and data sampling as well as stochastic gradients. We also derive non-asymptotic error bounds under the adaptive-size SGD setting: the error bounds for bias and variance as a function of the radio m/n are monotonic decreasing and unimodal, respectively. Importantly, our analysis holds for both constant and adaptive step-size SGD setting, and is valid under general assumptions on data distribution and activation functions.
- Our non-asymptotic results show that, RF regression with SGD still generalizes well for interpolation learning, and is able to capture the double descent behavior. In addition, we demonstrate that the constant step-size SGD setting incurs no loss on the convergence rate of excess risk when compared to the exact least-squares closed form solution. Our empirical evaluations support our theoretical results and findings.

Our analysis sheds light on the effect of SGD on high dimensional RF models in under-/over-parameterized regimes, and bridges the gap between the minimal-norm solution and numerical iteration solution in terms of optimization and generalization on double descent. Hence, looking forward to analysis of modern (deep) neural networks under the realistic setting where m, n, d are all large and comparable is important (and indeed difficult), e.g., (Hu et al., 2020; Ju et al., 2020) on two-layer neural networks. We expect that our analysis would be helpful for understanding large dimensional machine learning and neural network models more generally.

2 Related Work and Problem Setting

This section reviews relevant works while formally introducing our problem setting of RF regression with SGD.

2.1 Related works

A flurry of research papers are devoted to analysis of over-parameterized models on optimization (Kawaguchi & Huang, 2019; Allen-Zhu et al., 2019; Zou & Gu, 2019), generalization (or their combination) under neural tangent kernel (Jacot et al., 2018; Arora et al., 2019; Chizat et al., 2019) and mean-field analysis regime (Mei et al., 2019; Chizat & Bach, 2020). We take a unified perspective on optimization and generalization but work in the high-dimensional setting to fully capture the double descent behavior. By high-dimensional setting, we mean that m, n, and d increase proportionally, large and comparable (Mei & Montanari, 2019; Ba et al., 2020; Liao et al., 2020; d'Ascoli et al., 2020b).

Random features model and double descent: Characterizing the double descent of the RF model is first due to Belkin et al. (2020) under the one-dimensional setting. Other perspectives derive from random matrix theory (RMT) in high dimensional statistics (Hastie et al., 2019; Mei & Montanari, 2019; Ba et al., 2020; Liao et al., 2020; Li et al., 2021) and from the replica method (d'Ascoli et al., 2020b; Rocks & Mehta, 2020; Gerace et al., 2020). Under specific assumptions on data distribution, activation functions, target function, and initialization, these results show that the generalization error/excess risk increase when $m/n < 1$, diverge when $m/n \to 1$, and then decrease when $m/n > 1$.

Leveraging the bias-variance decomposition analysis in d'Ascoli et al. (2020b); Rocks & Mehta (2020), Adlam & Pennington (2020); Lin & Dobriban (2021) refine these results by focusing on the *analysis of variance* due to multiple randomness sources. (Ba et al., 2020) on RF optimized by gradient descent exhibits the double descent behavior under the Gaussian data assumption. We refer to comparisons in Table 1 in Appendix A for further details.

Technically speaking, since RF (least-squares) regression involves with inverse random matrices, these two classes of methods attempt to achieve a similar target: how to disentangle the nonlinear activation function by the Gaussian equivalence conjecture. RMT utilizes calculus of deterministic equivalents (or resolvents) for random matrices and replica methods focus on some specific scalar parameters that allows for circumventing the expectation computation. In fact, most of the above methods can be asymptotically equivalent to the Gaussian covariate model (Hu & Lu, 2020).

Non-asymptotic stochastic approximation: A series of papers on linear/kernel least-squares regression with constant/adaptive step-size SGD often work in the under-parameterized regime, where d is finite and much smaller than n. For linear least-squares regression (Bach & Moulines, 2013; Jain et al., 2018) and kernel regression (without explicit regularization) (Dieuleveut & Bach, 2016; Dieuleveut et al., 2017), averaged SGD offers a sub-linear rate on bias, while achieving minimax rates on variance, which leads to a certain $\mathcal{O}(1/n)$ convergence rate for excess risk.

Carratino et al. (2018) focus on regularized RF (least-squares) regression with SGD in the under-parameterized regime. Their analysis largely relies on the Tikhonov regularization in an approximation theory view, which provides a convergence rate of $\mathcal{O}(n^{-\frac{2r}{\alpha+2r}})$ under the regularity condition $r \in [0,1]$ and capacity condition $\alpha \in [0,1]$.

In the over-parameterized regime, the excess risk in (Chen et al., 2020b) on least squares in high dimensions with averaged constant step-size SGD can be independent of d, and is further improved to converge with n in (Zou et al., 2021). Berthier et al. (2020); Varre et al. (2021) also demonstrate this convergence result under $\min$ or last-iterate setting for noiseless least squares. Besides, the existence of multiple descent (Chen et al., 2020a; Liang et al., 2019) beyond double descent and SGD as implicit regularizer (Neyshabur et al., 2017; Smith et al., 2020) can be traced to the above two lines of work. Our work shares some similar technical tools with (Dieuleveut & Bach, 2016) and (Zou et al., 2021) but differs from them in several aspects. We detail the differences in Section 4.

2.2 The problem setting

We study the standard problem setting for RF least-squares regression and adopt the relevant terminologies from learning theory: *cf.*, (Cucker & Zhou, 2007; Dieuleveut & Bach, 2016; Carratino et al., 2018; Li et al., 2021) for details. Let $X \subseteq \mathbb{R}^d$ be a metric space and $Y \subseteq \mathbb{R}$. The training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ is assumed to be independently drawn from a non-degenerate unknown Borel probability measure ρ on $X \times Y$. The *target function* of ρ is defined by $f_\rho(\boldsymbol{x}) = \int_Y y \, \mathrm{d}\rho(y \mid \boldsymbol{x})$, where $\rho(\cdot \mid \boldsymbol{x})$ is the conditional distribution of ρ at $\boldsymbol{x} \in X$.

RF least squares regression: We study the RF regression problem with the squared loss as follows:

$$\min_{f \in \mathcal{H}} \mathcal{E}(f), \qquad \mathcal{E}(f) := \int (f(\boldsymbol{x}) - y)^2 \rho(\boldsymbol{x}, y) = \|f - f_\rho\|^2_{L^2_{\rho_X}}, \text{with } f(\boldsymbol{x}) = \langle \boldsymbol{\theta}, \varphi(\boldsymbol{x}) \rangle_{\mathbb{R}^m},$$

where the optimization vector $\boldsymbol{\theta} \in \mathbb{R}^m$ and the feature mapping $\varphi(\boldsymbol{x})$ is defined as

$$\varphi(\boldsymbol{x}) := \frac{1}{\sqrt{m}} \left[\sigma(\boldsymbol{\omega}_1^\top \boldsymbol{x}/\sqrt{d}), \cdots, \sigma(\boldsymbol{\omega}_m^\top \boldsymbol{x}/\sqrt{d}) \right]^\top := \frac{1}{\sqrt{m}} \sigma(\boldsymbol{W}\boldsymbol{x}/\sqrt{d}) \in \mathbb{R}^m, \tag{1}$$

where $\boldsymbol{W} = [\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \cdots, \boldsymbol{\omega}_m]^\top \in \mathbb{R}^{m \times d}$ with $W_{ij} \sim \mathcal{N}(0,1)$ corresponds to such two-layer neural network initialized with random Gaussian weights. Then, the corresponding hypothesis space $\mathcal{H}$ is a reproducing kernel Hilbert space

$$\mathcal{H} := \left\{ f \in L^2_{\rho_X} \middle| \; f(\boldsymbol{x}) = \frac{1}{\sqrt{m}} \langle \boldsymbol{\theta}, \sigma(\boldsymbol{W}\boldsymbol{x}/\sqrt{d}) \rangle, \boldsymbol{\theta} \in \mathbb{R}^m, \; W_{ij} \sim \mathcal{N}(0,1) \right\}, \tag{2}$$

with $\|f\|^2_{L^2_{\rho_X}} = \int_X |f(\boldsymbol{x})|^2 \mathrm{d}\rho_X(\boldsymbol{x}) = \langle f, \Sigma_m f \rangle_{\mathcal{H}}$ with the *covariance* operator $\Sigma_m : \mathbb{R}^m \to \mathbb{R}^m$

$$\Sigma_m = \int_X \varphi(\boldsymbol{x}) \otimes \varphi(\boldsymbol{x}) \mathrm{d}\rho_X(\boldsymbol{x}), \tag{3}$$

which is the usually (uncentered) covariance matrix in finite dimensions,[2] i.e., $\boldsymbol{\Sigma}_m = \mathbb{E}_{\boldsymbol{x}}[\varphi(\boldsymbol{x}) \otimes \varphi(\boldsymbol{x})]$. Clearly, Σ_m is random with respect to $\boldsymbol{W}$, and thus its deterministic version is defined as $\widetilde{\Sigma}_m = \mathbb{E}_{\boldsymbol{x},\boldsymbol{W}}[\varphi(\boldsymbol{x}) \otimes \varphi(\boldsymbol{x})]$. Define $J_m : \mathbb{R}^m \to L^2_{\rho_X}$ such that

$$(J_m \boldsymbol{v})(\cdot) = \langle \boldsymbol{v}, \varphi(\cdot) \rangle, \forall \boldsymbol{v} \in \mathbb{R}^m,$$

we have $\Sigma_m = J_m^* J_m$, where J_m^* denotes the adjoint operator of J_m.

SGD with averaging: Regarding the stochastic approximation, we consider the adaptive step-size SGD with iterate averaging (Dieuleveut & Bach, 2016; Zou et al., 2021; Nitanda & Suzuki, 2020): at each iteration t, after a training sample $(\boldsymbol{x}_t, y_t) \sim \rho$ is observed, we update the decision variable as

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} + \gamma_t [y_t - \langle \boldsymbol{\theta}_{t-1}, \varphi(\boldsymbol{x}_t) \rangle] \varphi(\boldsymbol{x}_t), \qquad t = 1, 2, \ldots \tag{4}$$

initialized at $\boldsymbol{\theta}_0$. Here the step-size is given by $\gamma_t := \gamma_0 t^{-\zeta}$ with $\zeta \in [0,1)$, which naturally holds for the constant step-size case by taking $\zeta = 0$. The final output is defined as the average of the iterates[3]:

$$\bar{\boldsymbol{\theta}}_n := \frac{1}{n} \sum_{t=0}^{n-1} \boldsymbol{\theta}_t.$$

[2] In this paper, we do not distinguish Σ_m and $\boldsymbol{\Sigma}_m$. This is also suitable to other operators/matrices, e.g., $\widetilde{\Sigma}_m$.

[3] We sum up $\{\theta_t\}_{t=0}^{n-1}$ with n terms for notational simplicity instead of summarizing $\{\theta_t\}_{t=0}^{n}$ with $n+1$ terms.

The optimality condition for Eq. 4 implies $\mathbb{E}_{(\boldsymbol{x},y)\sim\rho}[(y-\langle\boldsymbol{\theta}^*,\varphi(\boldsymbol{x})\rangle)\varphi(\boldsymbol{x})]=\mathbf{0}$, which corresponds to $f^*=J_m\boldsymbol{\theta}^*$ if we assume that $f^*=\arg\min_{f\in\mathcal{H}}\mathcal{E}(f)$ exists (see Assumption 1). Likewise, we have $f_t=J_m\boldsymbol{\theta}_t$ and $\bar{f}_n=J_m\bar{\boldsymbol{\theta}}_n$.

In this paper, we study the excess risk $\mathbb{E}\|\bar{f}_n-f^*\|^2_{L^2_{\rho_X}}$ instead of $\mathbb{E}\|\bar{f}_n-f_\rho\|^2_{L^2_{\rho_X}}$, that follows (Dieuleveut & Bach, 2016; Rudi & Rosasco, 2017; Carratino et al., 2018; Li et al., 2021), as f^* is the best possible solution in $\mathcal{H}$ and the mis-specification error $\|f^*-f_\rho\|^2_{L^2_{\rho_X}}$ pales into insignificance. Note that the expectation used here is considered with respect to the random features matrix $\boldsymbol{W}$, and the distribution of the training data $\{(\boldsymbol{x}_t,y_t)\}_{t=1}^n$ (note that $\|\bar{f}_n-f^*\|^2_{L^2_{\rho_X}}$ is itself a different expectation over ρ_X).

Notation: For two operators/matrices, $A\preccurlyeq B$ means $B-A$ is positive semi-definite (PSD). For any two positive sequences $\{a_t\}_{t=1}^s$ and $\{b_t\}_{t=1}^s$, the notation $a_t\lesssim b_t$ means that there exists a positive constant C independent of s such that $a_t\leq Cb_t$, and analogously for $\sim$, $\gtrsim$, and $\precsim$. For any $a,b\in\mathbb{R}$, $a\wedge b$ denotes the minimum of a and b.

3 Main results

In this section, we present our main theoretical results on the generalization properties employing error bounds for bias and variance of RF regression in high dimensions optimized by averaged SGD.

3.1 Assumptions

Before we present our result, we list the assumptions used in this paper.

Assumption 1. *(existence of f^*) There exists $f^*\in\mathcal{H}$ such that*

$$f^*=\operatorname*{arg\,min}_{f\in\mathcal{H}}\mathcal{E}(f)\,,$$

Remark: This is a standard assumption in learning theory, e.g., (Rudi & Rosasco, 2017; Carratino et al., 2018). Assuming the existence of $f^*\in\mathcal{H}$ implies that $\|f\|_{\mathcal{H}}$ is bounded in Eq. 2, which is in fact indispensable and standard.

Assumption 2. *(high dimensional assumption) We work in the high dimensional regime for some large d,n with $c\leqslant\{d/n,m/n\}\leqslant C$ for some constants $c,C>0$ such that m,n,d are large and comparable. The data point $\boldsymbol{x}\in\mathbb{R}^d$ is assumed to satisfy $\|\boldsymbol{x}\|_2^2\sim\mathcal{O}(d)$ and the covariance operator $\Sigma_d:=\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}\otimes\boldsymbol{x}]$ with bounded spectral norm $\|\Sigma_d\|_2$ (finite and independent of d).*

Remark: This is common and standard in high dimensional statistics (El Karoui, 2010; Hastie et al., 2019).

Assumption 3. *The activation function $\sigma(\cdot)$ is assumed to be Lipschitz continuous.*

Remark: This assumption is quite general to cover commonly-used activation functions used in random features and neural networks, e.g., ReLU, Sigmoid, $\sin$ / $\cos$. Under Assumption 2 and 3, $\mathbb{E}_{\boldsymbol{x}}\mathbb{V}[\sigma(z)]\sim\mathcal{O}(1)$ naturally holds as $\sigma(z)$ is sub-Gaussian with $\mathcal{O}(1)$ norm (Wainwright, 2019, Theorem 2.26) and its finite second moment, i.e., $\mathbb{V}[\sigma(z)]\sim\mathcal{O}(1)$.

Recall $\Sigma_m:=\mathbb{E}_{\boldsymbol{x}}[\varphi(\boldsymbol{x})\otimes\varphi(\boldsymbol{x})]$ in Eq. 3 and its expectation $\widetilde{\Sigma}_m:=\mathbb{E}_{\boldsymbol{W}}[\Sigma_m]$, we make the following fourth moment assumption.

Assumption 4 (Fourth moment condition)**.** *Assume there exists some positive constants $r',r\geqslant1$, such that for any PSD operator A, it holds that*

$$\mathbb{E}_{\boldsymbol{W}}[\Sigma_mA\Sigma_m]\preccurlyeq\mathbb{E}_{\boldsymbol{W}}\left(\mathbb{E}_{\boldsymbol{x}}\Big([\varphi(\boldsymbol{x})\otimes\varphi(\boldsymbol{x})]A[\varphi(\boldsymbol{x})\otimes\varphi(\boldsymbol{x})]\Big)\right)\preccurlyeq r'\mathbb{E}_{\boldsymbol{W}}[\mathrm{Tr}(\Sigma_mA)\Sigma_m]\preccurlyeq r\mathrm{Tr}(\widetilde{\Sigma}_mA)\widetilde{\Sigma}_m.$$

Remark: This assumption follows (Zou et al., 2021) that requires the data are drawn from some not-too-heavy-tailed distribution, e.g., $\Sigma_m^{-\frac{1}{2}}\boldsymbol{x}$ has sub-Gaussian, or sub-exponential tails. We make the following remarks:
1) The special case for $A:=I$ is proved by Lemma 4 (introduced in the next subsection) and thus this assumption is a natural extension. In fact, there is no need to require that this assumption holds for any PSD operator A (this is just for description simplicity). Validation on some specific PSD operators A is enough in our proof.

2) Assuming $\Sigma_m^{-\frac{1}{2}}\boldsymbol{x}$ to be sub-Gaussian/exponential is common in high dimensional statistics (Bartlett et al., 2020). This condition is much weaker than most previous work on double descent that requires the data to be Gaussian (Hastie et al., 2019; d'Ascoli et al., 2020b; Adlam & Pennington, 2020; Ba et al., 2020), or uniformly spread on a sphere (Mei & Montanari, 2019; Ghorbani et al., 2021), see comparisons in Table 1 in Appendix A. In stochastic approximation, the

boundeness of the fourth moment is also needed, see (Bach & Moulines, 2013; Dieuleveut & Bach, 2016; Jain et al., 2018; Berthier et al., 2020; Varre et al., 2021) for details.

Assumption 5 (Noise condition)**.** *There exists $\tau > 0$ such that*

$$\Xi := \mathbb{E}_{\boldsymbol{x}}[\varepsilon^2 \varphi(\boldsymbol{x}) \otimes \varphi(\boldsymbol{x})] \preccurlyeq \tau^2 \Sigma_m \,,$$

where the noise $\varepsilon := y - f^*(\boldsymbol{x})$.

Remark: This noise assumption is standard in (Dieuleveut & Bach, 2016; Zou et al., 2021) and holds for the standard noise model $y = f^*(\boldsymbol{x}) + \varepsilon$ with $\mathbb{E}[\varepsilon] = 0$ and $\mathbb{V}[\varepsilon] < \infty$ (Hastie et al., 2019). For proof simplicity, we consider the well-specified case $\mathbb{E}[\varepsilon|\boldsymbol{x}] = 0$ (can be extended to the model mis-specified case $\mathbb{E}[\varepsilon|\boldsymbol{x}] \neq 0$) and thus $\boldsymbol{y} - f^*(\boldsymbol{X})$ is independent of $\boldsymbol{X}$.

3.2 Properties of covariance operators

Before we present the main results, we study statistical properties of Σ_m and $\widetilde{\Sigma}_m$ by the following lemmas (with proof deferred to Appendix B), that will be needed for our main result.

Lemma 1. *Under Assumption 2,and 3, the covariance operator $\widetilde{\Sigma}_m := \mathbb{E}_{\boldsymbol{x},\boldsymbol{W}}[\varphi(\boldsymbol{x}) \otimes \varphi(\boldsymbol{x})]$ has the same diagonal elements*

$$(\widetilde{\Sigma}_m)_{ii} = \frac{1}{m} \mathbb{E}_{\boldsymbol{x}} \mathbb{E}_{z \sim \mathcal{N}(0, \|\boldsymbol{x}\|_2^2/d)} [\sigma(z)]^2 \sim \mathcal{O}(1/m), \;\; i = 1, 2, \ldots, m \,,$$

and the same non-diagonal elements

$$(\widetilde{\Sigma}_m)_{ij} = \frac{1}{m} \mathbb{E}_{\boldsymbol{x}} \left(\mathbb{E}_{z \sim \mathcal{N}(0, \|\boldsymbol{x}\|_2^2/d)} [\sigma(z)] \right)^2 \sim \mathcal{O}(1/m), \;\; i, j = 1, 2, \ldots, m \text{ with } i \neq j \,.$$

Accordingly, $\widetilde{\Sigma}_m$ has only two distinct eigenvalues

$$\widetilde{\lambda}_1 = (\widetilde{\Sigma}_m)_{ii} + (m-1)(\widetilde{\Sigma}_m)_{ij} \sim \mathcal{O}(1) \;\;, \;\; \widetilde{\lambda}_2 = \cdots = \widetilde{\lambda}_m = (\widetilde{\Sigma}_m)_{ii} - (\widetilde{\Sigma}_m)_{ij} = \frac{1}{m} \mathbb{E}_{\boldsymbol{x}} \mathbb{V}[\sigma(z)] \sim \mathcal{O}\left(\frac{1}{m}\right) \,.$$

Remark: Lemma 1 implies $\operatorname{tr}(\widetilde{\Sigma}_m) < \infty$. In fact, $\mathbb{E}_{\boldsymbol{x}} \mathbb{V}[\sigma(z)] > 0$ holds almost surely as $\sigma(\cdot)$ is not a constant, and thus $\widetilde{\Sigma}_m$ is positive definite. Our error bounds will largely depend on $\widetilde{\lambda}_2 = \frac{1}{m} \mathbb{E}_{\boldsymbol{x}} \mathbb{V}[\sigma(z)]$.

Here we take several examples by taking various activation functions $\sigma(\cdot)$ for demonstration.

1) If we choose $\sigma(x) = [\cos(x), \sin(x)]^\top$, RF actually approximates the Gaussian kernel with $\varphi(\boldsymbol{x}) \in \mathbb{R}^{2m}$ in Eq. 1. In this case, $(\widetilde{\Sigma}_m)_{ii} = 1/m$, and the non-diagonal element admits $(\widetilde{\Sigma}_m)_{ij} = \frac{1}{m} \mathbb{E}_{\boldsymbol{x}} \exp\left(-\frac{\|\boldsymbol{x}\|_2^2}{d}\right)$. 2) If we choose the ReLU activation $\sigma(x) = \max\{x, 0\}$, RF actually approximates the first-order arc-cosine kernel (Cho & Saul, 2009) with $\varphi(\boldsymbol{x}) \in \mathbb{R}^m$. We have $(\widetilde{\Sigma}_m)_{ii} = \frac{1}{2md} \operatorname{Tr}(\Sigma_d)$ and $(\widetilde{\Sigma}_m)_{ij} = \frac{1}{2md\pi} \operatorname{Tr}(\Sigma_d)$ (recall $\Sigma_d := \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}\boldsymbol{x}^\top]$). The calculation of the above two cases can be found in Appendix B.1.

Lemma 2. *Under Assumption 2,and 3, random variables $\|\Sigma_m\|_2$, $\|\Sigma_m - \widetilde{\Sigma}_m\|_2$, and $\operatorname{Tr}(\Sigma_m)$ are sub-exponential, and have sub-exponential norm at $\mathcal{O}(1)$ order.*

Lemma 3. *Under Assumption 2,and 3, we have $\left\| \widetilde{\Sigma}_m^{-2} \mathbb{E}_{\boldsymbol{W}}(\Sigma_m^2) \right\|_2 \sim \mathcal{O}(1)$.*

Lemma 4. *Under Assumption 2,and 3, there exists a constant $r > 0$ such that $\mathbb{E}_{\boldsymbol{W}}\left(\Sigma_m^2\right) \preccurlyeq \mathbb{E}_{\boldsymbol{x},\boldsymbol{W}}[\varphi(\boldsymbol{x}) \otimes \varphi(\boldsymbol{x}) \otimes \varphi(\boldsymbol{x}) \otimes \varphi(\boldsymbol{x})] \preccurlyeq r \operatorname{Tr}(\widetilde{\Sigma}_m) \widetilde{\Sigma}_m$.*

Remark: Lemma 4 is a special case of Assumption 4 if we take $A := I$ and $r := 1 + \mathcal{O}\left(\frac{1}{m}\right)$.

3.3 Results for error bounds

Recall the definition of the noise $\boldsymbol{\varepsilon} = [\varepsilon_1, \cdots, \varepsilon_n]^\top$ with $\varepsilon_t = y_t - f^*(\boldsymbol{x}_t)$, $t = 1, 2, \ldots, n$, the averaged excess risk can be expressed as

$$\mathbb{E}\|\bar{f}_n - f^*\|_{L^2_{\rho_X}}^2 := \mathbb{E}_{\boldsymbol{X},\boldsymbol{W},\boldsymbol{\varepsilon}} \|\bar{f}_n - f^*\|_{L^2_{\rho_X}}^2 = \mathbb{E}_{\boldsymbol{X},\boldsymbol{W},\boldsymbol{\varepsilon}} \langle \bar{f}_n - f^*, \Sigma_m(\bar{f}_n - f^*) \rangle = \mathbb{E}_{\boldsymbol{X},\boldsymbol{W},\boldsymbol{\varepsilon}} \langle \bar{\eta}_n, \Sigma_m \bar{\eta}_n \rangle \,,$$

where $\bar{\eta}_n := \frac{1}{n} \sum_{t=0}^{n-1} \eta_t$ with the centered SGD iterate $\eta_t := f_t - f^*$. Following the standard bias-variance decomposition in stochastic approximation (Dieuleveut & Bach, 2016; Jain et al., 2018; Zou et al., 2021), it admits

$$\eta_t = f_t - f^* = [I - \gamma_t \varphi(\boldsymbol{x}_t) \otimes \varphi(\boldsymbol{x}_t)](f_{t-1} - f^*) + \gamma_t \varepsilon_t \varphi(\boldsymbol{x}_t) \,,$$

where the first term corresponds to the bias by taking $y_t := f^*(\boldsymbol{x}_t)$

$$\eta_t^{\texttt{bias}} = [I - \gamma_t \varphi(\boldsymbol{x}_t) \otimes \varphi(\boldsymbol{x}_t)]\eta_{t-1}^{\texttt{bias}}, \quad \eta_0^{\texttt{bias}} = f^* , \tag{5}$$

and the second term corresponds to the variance

$$\eta_t^{\texttt{var}} = [I - \gamma_t \varphi(\boldsymbol{x}_t) \otimes \varphi(\boldsymbol{x}_t)]\eta_{t-1}^{\texttt{var}} + \gamma_t \varepsilon_t \varphi(\boldsymbol{x}_t), \quad \eta_0^{\texttt{var}} = 0 . \tag{6}$$

Accordingly, we have $f_t = \eta_t^{\texttt{bias}} + \eta_t^{\texttt{var}} + f^*$ due to $\mathbb{E}_{\boldsymbol{\varepsilon}} \bar{f}_n = \bar{\eta}_n^{\texttt{bias}} + f^*$ and $\|f\|^2_{L^2_{\rho_X}} = \langle f, \Sigma_m f \rangle$.

Proposition 1. *Based on the above setting, the averaged excess risk admits the following bias-variance decomposition*

$$\mathbb{E}_{\boldsymbol{X},\boldsymbol{W},\boldsymbol{\varepsilon}} \|\bar{f}_n - f^*\|^2_{L^2_{\rho_X}} = \mathbb{E}_{\boldsymbol{X},\boldsymbol{W},\boldsymbol{\varepsilon}} \|\bar{f}_n - \mathbb{E}_{\boldsymbol{\varepsilon}} \bar{f}_n + \mathbb{E}_{\boldsymbol{\varepsilon}} \bar{f}_n - f^*\|^2_{L^2_{\rho_X}} = \underbrace{\mathbb{E}_{\boldsymbol{X},\boldsymbol{W}} \langle \bar{\eta}_n^{\texttt{bias}}, \Sigma_m \bar{\eta}_n^{\texttt{bias}} \rangle}_{:=\texttt{Bias}} + \underbrace{\mathbb{E}_{\boldsymbol{X},\boldsymbol{W},\boldsymbol{\varepsilon}} \langle \bar{\eta}_n^{\texttt{var}}, \Sigma_m \bar{\eta}_n^{\texttt{var}} \rangle}_{:=\texttt{Variance}} .$$

By decoupling the multiple randomness sources of initialization, label noise, and data sampling (as well as stochastic gradients), we give precise non-asymptotic error bounds for bias and variance as below.

Theorem 1. *(Error bound for bias) Under Assumptions 1, 2, 3, 4 with $r' \geqslant 1$, if the step-size $\gamma_t := \gamma_0 t^{-\zeta}$ with $\zeta \in [0, 1)$ satisfies*

$$\gamma_0 < \min \left\{ \frac{1}{\operatorname{Tr}(\widetilde{\Sigma}_m)}, \frac{1}{r' \operatorname{Tr}(\Sigma_m)}, \frac{1}{2\operatorname{Tr}(\Sigma_m)} \right\} \sim \mathcal{O}(1) , \tag{7}$$

the `Bias` *defined in Proposition 1 holds*

$$\texttt{Bias} \lesssim \frac{\gamma_0 r' n^{\zeta - 1}}{\sqrt{\mathbb{E}[1 - \gamma_0 r' \operatorname{Tr}(\Sigma_m)]^4}} \|f^*\|^2 \sim \mathcal{O}\left(n^{\zeta-1}\right) . \tag{8}$$

Remark: In our paper, $I - \gamma_t \Sigma_m$ ($t = 1, 2, \ldots, n$) is required to be a contraction map by taking $\gamma_0 < 1/\operatorname{Tr}(\Sigma_m)$. Though $\operatorname{Tr}(\Sigma_m)$ is a random variable, the condition $\gamma_0 < 1/\operatorname{Tr}(\Sigma_m)$ can be equivalently substituted by $\gamma_0 < 1/[c\operatorname{Tr}(\widetilde{\Sigma}_m)]$ for some large c (independent of n, m, d) with exponentially high probability. This is because, $\operatorname{Tr}(\Sigma_m)$ is a sub-exponential random variable with $\mathcal{O}(1)$ norm in Lemma 2, which makes the constant c unnecessary to be quite large. For example, the probability with $\exp(-10) < 10^{-4}$ and $\exp(-100) < 10^{-43}$ by taking $c = 10$, or 100 is enough small in practice. Accordingly, the condition in Eq. 7 can be equivalently substituted by $\gamma_0 < \frac{1}{cr'\operatorname{Tr}(\widetilde{\Sigma}_m)}$ for some large c. This is also suitable for estimating `Variance`.

Theorem 2. *(Error bound for variance) Under Assumptions 2, 3, 4 with $r' \geqslant 1$, and Assumption 5 with $\tau > 0$, if the step-size $\gamma_t := \gamma_0 t^{-\zeta}$ with $\zeta \in [0, 1)$ satisfies Eq. 7, the* `Variance` *defined in Proposition 1 holds*

$$\texttt{Variance} \lesssim \frac{\gamma_0 r' \tau^2}{\sqrt{\mathbb{E}[1 - \gamma_0 r' \operatorname{Tr}(\Sigma_m)]^2}} \begin{cases} mn^{\zeta-1}, \text{ if } m \leqslant n \\ \gamma_0 \tau^2, \text{ if } m > n \end{cases} \sim \begin{cases} \mathcal{O}\left(mn^{\zeta-1}\right), \text{ if } m \leqslant n \\ \mathcal{O}\left(1\right), \text{ if } m > n . \end{cases} \tag{9}$$

Remark: The error bound for `Variance` is demonstrated to be unimodal, and converges to $\mathcal{O}(1)$ in the over-parameterized regimes, which matches recent results relying on closed-form solution on (refined) variance, e.g., (d'Ascoli et al., 2020b; Adlam & Pennington, 2020; Lin & Dobriban, 2021).

4 Proof Outline and Discussion

In this section, we first introduce the structure of the proofs with high level ideas, and then discuss our work with previous literature in terms of the used techniques and the obtained results.

4.1 Proof outline

We (partly) disentangle the multiple randomness sources on the data $\boldsymbol{X}$, the random features matrix $\boldsymbol{W}$, the noise $\boldsymbol{\varepsilon}$, make full use of statistical properties of covariance operators Σ_m and $\widetilde{\Sigma}_m$ in Section 3.2, and provide the respective (bias and variance) upper bounds in terms of multiple randomness sources, as shown in Figure 1.

Bias: To bound `Bias`, we need some auxiliary notations.

$$\eta_t^{\texttt{bX}} = (I - \gamma_t \Sigma_m)\eta_{t-1}^{\texttt{bX}}, \quad \eta_0^{\texttt{bX}} = f^*, \quad \text{with } \Sigma_m = \mathbb{E}_{\boldsymbol{x}}[\varphi(\boldsymbol{x}) \otimes \varphi(\boldsymbol{x})] . \tag{10}$$

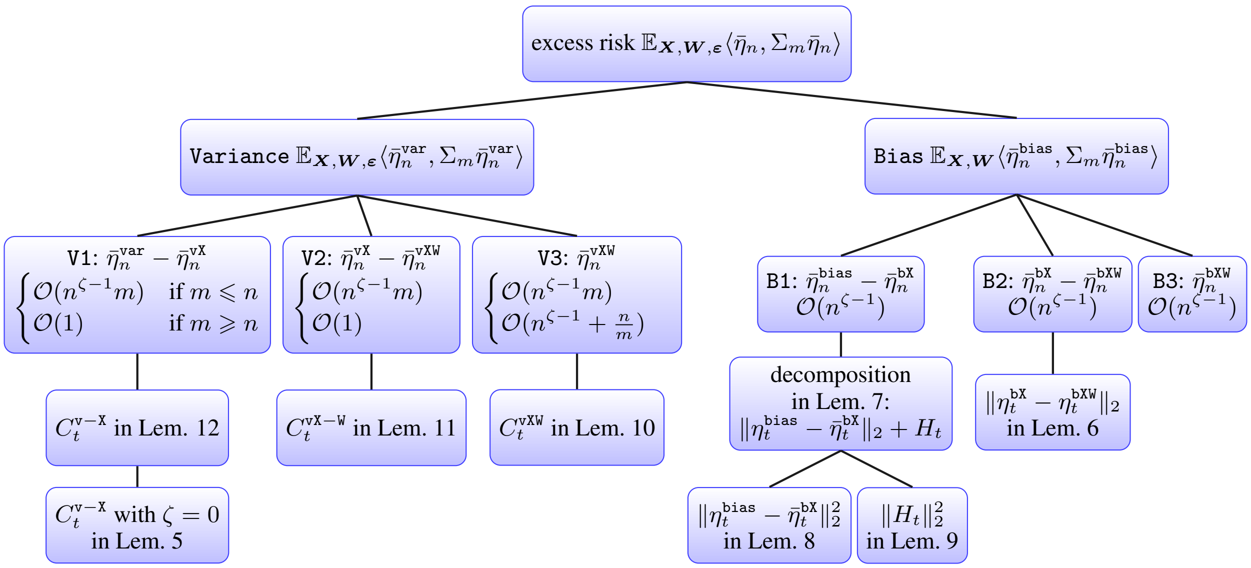


Figure 1: The roadmap of proofs.

$$\eta_t^{\mathtt{bXW}} = (I - \gamma_t \widetilde{\Sigma}_m)\eta_{t-1}^{\mathtt{bXW}}, \quad \eta_0^{\mathtt{bXW}} = f^*, \quad \text{with } \widetilde{\Sigma}_m = \mathbb{E}_{\boldsymbol{x},\boldsymbol{W}}[\varphi(\boldsymbol{x}) \otimes \varphi(\boldsymbol{x})], \tag{11}$$

with the average $\bar{\eta}_n^{\mathtt{bX}} := \frac{1}{n}\sum_{t=0}^{n-1} \bar{\eta}_t^{\mathtt{bX}}$ and $\bar{\eta}_n^{\mathtt{bXW}} := \frac{1}{n}\sum_{t=0}^{n-1} \bar{\eta}_t^{\mathtt{bXW}}$. Accordingly, $\eta_t^{\mathtt{bX}}$ can be regarded as a "deterministic" version of $\eta_t^{\mathtt{bias}}$: we omit the randomness on $\boldsymbol{X}$ (data sampling, stochastic gradients) by replacing $[\varphi(\boldsymbol{x})\varphi(\boldsymbol{x})^\top]$ with its expectation Σ_m. Likewise, $\eta_t^{\mathtt{bXW}}$ is a deterministic version of $\eta_t^{\mathtt{vX}}$ by replacing Σ_m with its expectation $\widetilde{\Sigma}_m$ (randomness on initialization).

By virtue of Minkowski inequality, the $\mathtt{Bias}$ can be decomposed as $\mathtt{Bias} \lesssim \mathtt{B1}+\mathtt{B2}+\mathtt{B3}$, where $\mathtt{B1} := \mathbb{E}_{\boldsymbol{X},\boldsymbol{W}}\big[\langle \bar{\eta}_n^{\mathtt{bias}} - \bar{\eta}_n^{\mathtt{bX}}, \Sigma_m(\bar{\eta}_n^{\mathtt{bias}} - \bar{\eta}_n^{\mathtt{bX}})\rangle\big]$ and $\mathtt{B2} := \mathbb{E}_{\boldsymbol{W}}\big[\langle \bar{\eta}_n^{\mathtt{bX}} - \bar{\eta}_n^{\mathtt{bXW}}, \Sigma_m(\bar{\eta}_n^{\mathtt{bX}} - \bar{\eta}_n^{\mathtt{bXW}})\rangle\big]$ and $\mathtt{B3} := \langle \bar{\eta}_n^{\mathtt{bXW}}, \widetilde{\Sigma}_m \bar{\eta}_n^{\mathtt{bXW}}\rangle$. Here $\mathtt{B3}$ is a deterministic quantity that is closely connected to model (intrinsic) bias without any randomness; while $\mathtt{B1}$ and $\mathtt{B2}$ evaluate the effect of randomness from $\boldsymbol{X}$ and $\boldsymbol{W}$ on the bias, respectively. The error bounds (convergence rates) for them can be directly found in Figure 1.

To bound $\mathtt{B3}$, we directly focus on its formulation by virtue of spectrum decomposition and integral estimation. To bound $\mathtt{B2}$, we need study $\|\eta_t^{\mathtt{bX}} - \eta_t^{\mathtt{bXW}}\|_2 \lesssim \|\Sigma_m\|_2 \|f^*\|$ in Lemma 6. To bound $\mathtt{B1}$, it can be further decomposed as (here we use inaccurate expression for description simplicity) $\mathtt{B1} \lesssim \sum_t \|\eta_t^{\mathtt{bX}} - \eta_t^{\mathtt{bXW}}\|_2^2 + \sum_t \mathbb{E}_{\boldsymbol{X}}\|H_t\|^2$ in Lemma 7, where $H_{t-1} := [\Sigma_m - \varphi(\boldsymbol{x}_t)\otimes\varphi(\boldsymbol{x}_t)]\eta_{t-1}^{\mathtt{bX}}$. The first term can be upper bounded by $\sum_t \|\eta_t^{\mathtt{bX}} - \eta_t^{\mathtt{bXW}}\|_2^2 \lesssim \mathrm{Tr}(\Sigma_m) n^{\zeta} \|f^*\|^2$ in Lemma 8, and the second term admits $\sum_t \mathbb{E}_{\boldsymbol{X}}\|H_t\|^2 \lesssim \mathrm{Tr}(\Sigma_m)\|f^*\|^2$ in Lemma 9.

Variance: To bound $\mathtt{Variance}$, we need some auxiliary notations.

$$\eta_t^{\mathtt{vX}} := (I - \gamma_t \Sigma_m)\eta_{t-1}^{\mathtt{vX}} + \gamma_t \varepsilon_t \varphi(\boldsymbol{x}_t), \quad \eta_0^{\mathtt{vX}} = 0, \quad \text{with } \Sigma_m = \mathbb{E}_{\boldsymbol{x}}[\varphi(\boldsymbol{x}) \otimes \varphi(\boldsymbol{x})]. \tag{12}$$

$$\eta_t^{\mathtt{vXW}} := (I - \gamma_t \widetilde{\Sigma}_m)\eta_{t-1}^{\mathtt{vXW}} + \gamma_t \varepsilon_t \varphi(\boldsymbol{x}_t), \quad \eta_0^{\mathtt{vXW}} = 0, \quad \text{with } \widetilde{\Sigma}_m = \mathbb{E}_{\boldsymbol{x},\boldsymbol{W}}[\varphi(\boldsymbol{x}) \otimes \varphi(\boldsymbol{x})], \tag{13}$$

with the averaged quantities $\bar{\eta}_n^{\mathtt{vX}} := \frac{1}{n}\sum_{t=0}^{n-1} \bar{\eta}_t^{\mathtt{vX}}$, $\bar{\eta}_n^{\mathtt{vXW}} := \frac{1}{n}\sum_{t=0}^{n-1} \bar{\eta}_t^{\mathtt{vXW}}$. Accordingly, $\eta_t^{\mathtt{vX}}$ can be regarded as a "semi-stochastic" version of $\eta_t^{\mathtt{var}}$: we keep the randomness due to the noise ε_t but omit the randomness on $\boldsymbol{X}$ (data sampling) by replacing $[\varphi(\boldsymbol{x})\varphi(\boldsymbol{x})^\top]$ with its expectation Σ_m. Likewise, $\eta_t^{\mathtt{vXW}}$ can be regarded as a "semi-stochastic" version of $\eta_t^{\mathtt{vX}}$ by replacing Σ_m with its expectation $\widetilde{\Sigma}_m$ (randomness on initialization).

By virtue of Minkowski inequality, the $\mathtt{Variance}$ can be decomposed as $\mathtt{Variance} \lesssim \mathtt{V1} + \mathtt{V2} + \mathtt{V3}$, where $\mathtt{V1} := \mathbb{E}_{\boldsymbol{X},\boldsymbol{W},\boldsymbol{\varepsilon}}\big[\langle \bar{\eta}_n^{\mathtt{var}} - \bar{\eta}_n^{\mathtt{vX}}, \Sigma_m(\bar{\eta}_n^{\mathtt{var}} - \bar{\eta}_n^{\mathtt{vX}})\rangle\big]$, $\mathtt{V2} := \mathbb{E}_{\boldsymbol{X},\boldsymbol{W},\boldsymbol{\varepsilon}}\big[\langle \bar{\eta}_n^{\mathtt{vX}} - \bar{\eta}_n^{\mathtt{vXW}}, \Sigma_m(\bar{\eta}_n^{\mathtt{vX}} - \bar{\eta}_n^{\mathtt{vXW}})\rangle\big]$, and $\mathtt{V3} := \mathbb{E}_{\boldsymbol{X},\boldsymbol{W},\boldsymbol{\varepsilon}}\langle \bar{\eta}_n^{\mathtt{vXW}}, \Sigma_m \bar{\eta}_n^{\mathtt{vXW}}\rangle$. Though $\mathtt{V1}$, $\mathtt{V2}$, $\mathtt{V3}$ still interact the multiple randomness, $\mathtt{V1}$ disentangles some randomness on data sampling, $\mathtt{V2}$ discards some randomness on initialization, and $\mathtt{V3}$ focuses on the "minimal" interaction between data sampling, label noise, and initialization. The error bounds for them can be found in Figure 1.

To bound $\mathtt{V3}$, we focus on the formulation of the covariance operator $C_t^{\mathtt{vXW}} := \mathbb{E}_{\boldsymbol{X},\boldsymbol{\varepsilon}}[\eta_t^{\mathtt{vXW}} \otimes \eta_t^{\mathtt{vXW}}]$ in Lemma 10 and the statistical properties of $\widetilde{\Sigma}_m$ and Σ_m. To bound $\mathtt{V2}$, we need study the covariance operator $C_t^{\mathtt{vX-W}} := \mathbb{E}_{\boldsymbol{X},\boldsymbol{\varepsilon}}[(\eta_t^{\mathtt{vX}} - \eta_t^{\mathtt{vXW}}) \otimes$

$(\eta_t^{\texttt{vX}} - \eta_t^{\texttt{vXW}})]$ admitting $\|C_t^{\texttt{vX-W}}\| \lesssim \|I + \widetilde{\Sigma}_m^{-2}\Sigma_m^2\|_2 \mathrm{Tr}(\Sigma_m)$ in Lemma 11. To bound V1, we need study the covariance operator $C_t^{\texttt{v-X}} := \mathbb{E}_{\boldsymbol{X},\boldsymbol{\varepsilon}}[(\eta_t^{\texttt{var}} - \eta_t^{\texttt{vX}}) \otimes (\eta_t^{\texttt{var}} - \eta_t^{\texttt{vX}})]$, as a function of $\zeta \in [0, 1)$, admitting $\mathrm{Tr}[C_t^{\texttt{v-X}}(\zeta)] \leqslant \mathrm{Tr}[C_t^{\texttt{v-X}}(0)]$ in Lemma 5, and further $C_t^{\texttt{v-X}} \precsim \mathrm{Tr}(\Sigma_m) I$ in Lemma 12.

4.2 Discussion with previous work

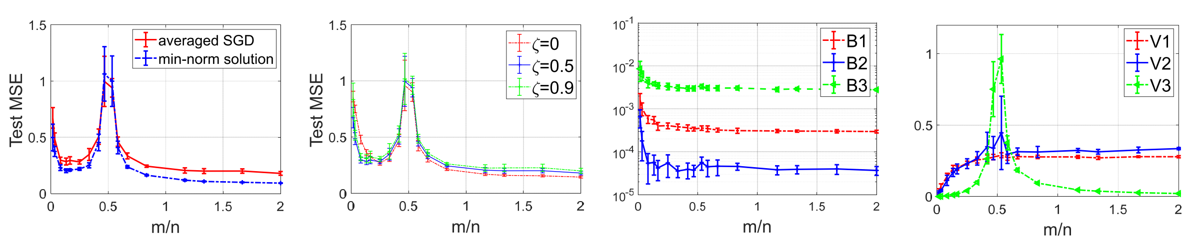
Difference in techniques: Our proof framework follows (Dieuleveut & Bach, 2016) that focuses on kernel regression with stochastic approximation in the under-parameterized regimes (d is regarded as finite and much smaller than n). Nevertheless, even in the under-parameterized regime, their results can not be directly extended to random features model due to the extra randomness on $\boldsymbol{W}$, coupling with other randomness sources on noise and data sampling, which makes their proof framework invalid on some points. To be specific, their results depend on (Bach & Moulines, 2013, Lemma 1) by taking conditional expectation to bridge the connection between $\mathbb{E}(\|\alpha_t\|_2)$ and $\mathbb{E}\langle \alpha_t, \Sigma_m \alpha_t \rangle$. This is valid for B1 but expires on other quantities. Besides, the results in (Carratino et al., 2018) on RF with SGD in the under-parameterized regimes depend on Tikhonov regularization in an approximation theory view, which appears invalid for our interpolation learning without any (explicit) regularization.

Some technical tools used in this paper follow (Zou et al., 2021) that focuses on linear regression with constant step-size SGD in over-parameterized regime, e.g., PSD operators and boundedness of $C_t^{\texttt{v-X}}$ when $\zeta = 0$ in Lemma 12. However, coupling with multiple randomness sources and adaptive step-size setting (no longer a homogeneous markov chain) make our analysis intractable. Besides, their results demonstrate that linear regression with SGD generalizes well (converges with n) but has few findings on double descent. Instead, our result depends on n and m (where d is implicitly included in m), and is able to explain double descent.

Comparison with previous work: Compared to (Ba et al., 2020) on RF optimized by gradient descent under the Gaussian data in an asymptotic view, our non-asymptotic result holds for more general data distribution under the SGD setting. In fact, our data assumption is weaker than most previous work assuming the data to be Gaussian, uniformly spread on a sphere, or isotropic/correlated features (with spectral decay assumption), except (Liao et al., 2020). Nevertheless, we extend their asymptotic results relying on the least-squares closed-form solution to non-asymptotic results under the SGD setting. Compared to (Li et al., 2021) relying on closed-form solution with correlated features, our result for bias achieves $\mathcal{O}(1/n)$ rate under the constant step-size SGD setting, which is better than their $\mathcal{O}(\sqrt{\log n/n})$ rate. Their result on variance requires $m \leqslant \mathcal{O}(nd)$ to generalize well while our result does not need this condition. Besides, our result coincides several findings with refined variance decomposition in (d'Ascoli et al., 2020b; Adlam & Pennington, 2020; Lin & Dobriban, 2021): the interaction effect can dominate the variance (between samples and initialization); the unimodality of variance is a prevalent phenomenon.

5 Numerical Validation

In this section, we provide some numerical experiments in Figure 2 to support our theoretical results and findings. Note that our results go beyond Gaussian data assumption and can be empirically validated on real-world datasets.



(a) SGD vs. min-norm solution (b) step-size (c) Bias (d) Variance

Figure 2: Test MSE (mean±std.) of RF regression as a function of the ratio m/n on MNIST data set (digit 3 vs. 7) across the Gaussian kernel, for $d = 784$ and $n = 600$ in (a) and (b). The interpolation threshold occurs at $2m = n$ due to $\sigma(\boldsymbol{W}\boldsymbol{x}) \in \mathbb{R}^{2m}$. Under this setting, the trends of Bias and Variance are empirically given in (c) and (d).

5.1 Behavior of RF for interpolation learning

Here we evaluate the test mean square error (MSE) of RF regression on the MNIST data set (Lecun et al., 1998) with minimal-norm solution and adaptive step-size SGD to study its generalization properties, see Figure 2(a) and 2(b).

Experimental settings: We take digit 3 vs. 7 as an example, and randomly select 300 training data in these two classes, resulting in $n = 600$ for training. Hence, our setting with $n = 600$ and $d = 784$ satisfies our realistic high dimensional assumption. The Gaussian kernel $k(\boldsymbol{x}, \boldsymbol{x}') = \exp(-\|\boldsymbol{x} - \boldsymbol{x}'\|_2^2/(2\sigma_0^2))$ is used, where the kernel width σ_0 is chosen as $\sigma_0^2 = d$ in high dimensional settings such that $\|\boldsymbol{x}\|_2^2/d \sim \mathcal{O}(1)$ in Assumption 2. In our experiment, each sample is normalized to zero-mean with deviations to 1. The initial step-size is set to $\gamma_0 = 1$ and the initial optimization parameter $\boldsymbol{\theta}_0$ is set to the min-norm solution[4] corrupted with zero-mean, unit-variance Gaussian noise, which is used to evaluate their difference on test error. The experiments are repeated 10 times and the test MSE (mean$\pm$std.) can be regarded as a function of the ratio m/n by tuning m.

SGD vs. minimal-norm solution: Figure 2(a) shows the test MSE of RF regression with averaged SGD (we take $\zeta = 0.5$ as an example; red line) and minimal-norm solution (blue line). First, we observe the double descent phenomenon: a phase transition on the two sides of the interpolation threshold at $2m = n$ when these two optimization algorithms are employed. Second, in terms of test error, RF with averaged SGD is slightly inferior to that with min-norm solution, but still generalizes well.

Different step-size: Figure 2(b) shows the test error of RF regression with averaged SGD under three different step-size settings, i.e., $\zeta = 0$ (red line), $\zeta = 0.5$ (blue line), and $\zeta = 0.9$ (green line). It is not surprising to observe the double descent phenomenon on these three settings. The constant step-size setting (i.e., $\zeta = 0$) achieves the best generalization performance in the over-parameterized regime, narrowly followed by the other two adaptive step-size settings, which experimentally validates the effectiveness of averaged SGD for interpolation learning.

5.2 Behavior of our error bounds

We have experimentally validate the phase transition and corresponding double descent in the previous section, and here we aim to semi-quantitatively assess our derived bounds for `Bias` and `Variance`, see Figure 2(c) and 2(d).

Experimental settings: Since the target function f^*, the covariance operators Σ_d, Σ_m, and the noise $\boldsymbol{\varepsilon}$ are unknown on the MNIST data set, we need some extra assumptions/settings to calculate `Bias` and `Variance` for our experimental evaluation. First, we assume the label noise $\varepsilon \sim \mathcal{N}(0, 1)$, which can in turn obtain $f^*(\boldsymbol{x})$ on both training and test data due to $f^*(\boldsymbol{x}) = y - \varepsilon$. Second, the covariance matrices Σ_d and Σ_m are estimated by the related sample covariance matrices. When using the Gaussian kernel, the covariance matrix $\widetilde{\Sigma}_m$ can be directly computed, i.e., $(\widetilde{\Sigma}_m)_{ii} = 1/m$ and $(\widetilde{\Sigma}_m)_{ij} = \frac{1}{m}\mathbb{E}_{\boldsymbol{x}} \exp\left(-\frac{\|\boldsymbol{x}\|_2^2}{d}\right)$, where the expectation is approximated by Monte Carlo sampling with n training samples. Accordingly, based on the above results, we are ready to calculate $\eta_t^{\texttt{bias}}$ in Eq. 5, $\eta_t^{\texttt{bX}}$ in Eq. 10, and $\eta_t^{\texttt{bXW}}$ in Eq. 11, respectively, which is further used to approximately compute $\texttt{B1} := \mathbb{E}_{\boldsymbol{X},\boldsymbol{W}}\left[\langle \bar{\eta}_n^{\texttt{bias}} - \bar{\eta}_n^{\texttt{bX}}, \Sigma_m(\bar{\eta}_n^{\texttt{bias}} - \bar{\eta}_n^{\texttt{bX}})\rangle\right]$ (red line) and $\texttt{B2} := \mathbb{E}_{\boldsymbol{W}}\left[\langle \bar{\eta}_n^{\texttt{bX}} - \bar{\eta}_n^{\texttt{bXW}}, \Sigma_m(\bar{\eta}_n^{\texttt{bX}} - \bar{\eta}_n^{\texttt{bXW}})\rangle\right]$ (blue line) and $\texttt{B3} := \langle \bar{\eta}_n^{\texttt{bXW}}, \widetilde{\Sigma}_m \bar{\eta}_n^{\texttt{bXW}}\rangle$ (green line). The (approximate) computation for `Variance` can be similar achieved by this process.

Error bounds for bias: Figure 2(c) shows the trends of (scaled) `B1`, `B2`, and `B3`. Recall our error bound: `B1`, `B2`, `B3` $\sim \mathcal{O}(n^{\zeta-1})$ with $\zeta = 0.5$ in our experiment. We find that, all of them monotonically decreases when m increases from the under-parameterized regime to the over-parameterized regime. These experimental results coincide with our error bound on them, i.e., converging with n at some certain rate (m and n are in the same order in our experiment).

Error bounds for variance: Figure 2(d) shows the trends of (scaled) `V1`, `V2`, and `V3`. Recall our error bound: in the under-parameterized regime, `V1`, `V2`, and `V3` increases with m at a certain $\mathcal{O}(n^{\zeta-1}m)$ rate; and in the over-parameterized regime, `V1` and `V2` are in $\mathcal{O}(1)$ order while `V3` decreases with m. Figure 2(d) shows that, when $2m < n$, `V1` and `V2` monotonically increases with m and then remain unchanged when $2m > n$. Besides, `V3` is observed to be unimodal: firstly increasing when $2m < n$, reaching to the peak at $2m = n$, and then decreasing when $2m > n$, which admits the phase transition at $2m = n$. Accordingly, these findings accord with our theoretical results, and also matches refined results in (d'Ascoli et al., 2020b; Adlam & Pennington, 2020; Lin & Dobriban, 2021): the unimodality of variance is a prevalent phenomenon.

6 Conclusion

We present the non-asymptotic results for RF regression under the averaged SGD setting for understanding benign overfitting. Our theoretical and empirical results demonstrate that, the error bounds for variance and bias can be unimodal and monotonically decreasing, respectively, which is able to recover the double descent phenomenon. Regarding to constant/adaptive step-size setting, there is no difference between the constant step-size case and the exact minimal-norm solution on the convergence rate; while the adaptive step-size case will slow down the learning rate, but does not change the error bound for variance in over-parameterized regime that remains $\mathcal{O}(1)$ order.

[4] In our numerical experiments, we take the regularization parameter fixed with 10^{-8} to avoid non-singular.

ETHICS STATEMENT

In this paper, we focus on generalization properties of random features models trained with SGD, working in a practical setting where n, m, d are large and comparable. The derived theoretical results in terms of optimization and generalization would have an important positive impact on over-parameterized models, e.g., deep neural networks. Our theoretical framework presents fair and non-offensive societal consequence.

REPRODUCIBILITY STATEMENT

We derived exact non-asymptotic error bounds for high dimensional RF regression trained with SGD in under/over-parameterized regimes. To support our theoretical results, in our main text, we have discussed that the assumptions used in this paper are fair and attainable; provided a proof roadmap of our proof framework; and detailed the experimental settings, e.g., data processing, training/test split, and parameter selection. In the appendix, we have provided the complete proof of our theoretical claims.

REFERENCES

Ben Adlam and Jeffrey Pennington. Understanding double descent requires a fine-grained bias-variance decomposition. In *Advances in neural information processing systems*, 2020.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pp. 242–252. PMLR, 2019.

Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pp. 322–332, 2019.

Jimmy Ba, Murat A. Erdogdu, Taiji Suzuki, Denny Wu, and Tianzong Zhang. Generalization of two-layer neural networks: an asymptotic viewpoint. In *International Conference on Learning Representations*, pp. 1–8, 2020.

Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $o(1/n)$. *Advances in Neural Information Processing Systems*, 26:773–781, 2013.

Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *the National Academy of Sciences*, 2020.

M. Belkin, D. Hsu, and J. Xu. Two models of double descent for weak features. *SIAM J. Math. Data Sci.*, 2(4): 1167–1180, 2020.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *the National Academy of Sciences*, 116(32):15849–15854, 2019.

Raphaël Berthier, Francis Bach, and Pierre Gaillard. Tight nonparametric convergence rates for stochastic gradient descent under the noiseless linear model. In *Advances in Neural Information Processing Systems*, volume 33, pp. 2576–2586, 2020.

Luigi Carratino, Alessandro Rudi, and Lorenzo Rosasco. Learning with SGD and random features. In *Advances in Neural Information Processing Systems*, pp. 10212–10223, 2018.

Lin Chen, Yifei Min, Mikhail Belkin, and Amin Karbasi. Multiple descent: Design your own generalization curve. *arXiv preprint arXiv:2008.01036*, 2020a.

Xi Chen, Qiang Liu, and Xin T Tong. Dimension independent generalization error by stochastic gradient descent. *arXiv preprint arXiv:2003.11196*, 2020b.

Lenaic Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Conference on Learning Theory*, pp. 1305–1338, 2020.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, pp. 2933–2943, 2019.

Youngmin Cho and Lawrence K Saul. Kernel methods for deep learning. In *Advances in Neural Information Processing Systems*, pp. 342–350, 2009.

Felipe Cucker and Dingxuan Zhou. *Learning theory: an approximation theory viewpoint*, volume 24. Cambridge University Press, 2007.

Oussama Dhifallah and Yue M Lu. A precise performance analysis of learning with random features. *arXiv preprint arXiv:2008.11904*, 2020.

Aymeric Dieuleveut and Francis Bach. Nonparametric stochastic approximation with large step-sizes. *Annals of Statistics*, 44(4):1363–1399, 2016.

Aymeric Dieuleveut, Nicolas Flammarion, and Francis Bach. Harder, better, faster, stronger convergence rates for least-squares regression. *Journal of Machine Learning Research*, 18(1):3520–3570, 2017.

S. d'Ascoli, M. Refinetti, G. Biroli, and F. Krzakala. Double trouble in double descent: Bias and variance (s) in the lazy regime. pp. 2280–2290, 2020a.

Stéphane d'Ascoli, Maria Refinetti, Giulio Biroli, and Florent Krzakala. Double trouble in double descent: Bias and variance (s) in the lazy regime. pp. 2280–2290, 2020b.

Noureddine El Karoui. The spectrum of kernel random matrices. *Annals of Statistics*, 38(1):1–50, 2010.

Federica Gerace, Bruno Loureiro, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. Generalisation error in learning with random features and the hidden manifold model. In *International Conference on Machine Learning*, pp. 3452–3462, 2020.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. *The Annals of Statistics*, 49(2):1029–1054, 2021.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

Alfred Horn. Doubly stochastic matrices and the diagonal of a rotation matrix. *American Journal of Mathematics*, 76 (3):620–630, 1954.

Hong Hu and Yue M Lu. Universality laws for high-dimensional learning with random features. *arXiv preprint arXiv:2009.07669*, 2020.

Wei Hu, Lechao Xiao, Ben Adlam, and Jeffrey Pennington. The surprising simplicity of the early-time learning dynamics of neural networks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 17116–17128, 2020.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, pp. 8571–8580, 2018.

Prateek Jain, Sham M Kakade, Rahul Kidambi, Praneeth Netrapalli, Venkata Krishna Pillutla, and Aaron Sidford. A markov chain theory approach to characterizing the minimax optimality of stochastic gradient descent (for least squares). *arXiv preprint arXiv:1710.09430*, 2017.

Prateek Jain, Sham Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. Parallelizing stochastic gradient descent for least squares regression: mini-batching, averaging, and model misspecification. *Journal of Machine Learning Research*, 18, 2018.

Peizhong Ju, Xiaojun Lin, and Ness B. Shroff. On the generalization power of overfitted two-layer neural tangent kernel models. In *International Conference on Machine Learning*, pp. 5137–5147. PMLR, 2020.

Kenji Kawaguchi and Jiaoyang Huang. Gradient descent finds global minima for generalizable deep neural networks of practical sizes. In *IEEE Conference on Communication, Control, and Computing*, pp. 92–99. IEEE, 2019.

Yann Lecun, Leon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Zhu Li, Jean-Francois Ton, Dino Oglic, and Dino Sejdinovic. Towards a unified analysis of random Fourier features. In *the 36th International Conference on Machine Learning*, pp. 3905–3914, 2019.

Zhu Li, Zhi-Hua Zhou, and Arthur Gretton. Towards an understanding of benign overfitting in neural networks. *arXiv preprint arXiv:2106.03212*, 2021.

Tengyuan Liang, Alexander Rakhlin, and Xiyu Zhai. On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels. In *Annual Conference on Learning Theory*, pp. 1–32, 2019.

Zhenyu Liao, Romain Couillet, and Michael Mahoney. A random matrix analysis of random fourier features: beyond the gaussian kernel, a precise phase transition, and the corresponding double descent. In *Neural Information Processing Systems*, 2020.

Licong Lin and Edgar Dobriban. What causes the test error? going beyond bias-variance via anova. *Journal of Machine Learning Research*, 22(155):1–82, 2021.

Fanghui Liu, Xiaolin Huang, Yudong Chen, and Johan A.K. Suykens. Random features for kernel approximation: A survey on algorithms, theory, and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–20, 2021a. doi: 10.1109/TPAMI.2021.3097011.

Fanghui Liu, Zhenyu Liao, and Johan A.K. Suykens. Kernel regression in high dimensions: Refined analysis beyond double descent. In *International Conference on Artificial Intelligence and Statistics*, pp. 649–657, 2021b.

Cosme Louart, Zhenyu Liao, and Romain Couillet. A random matrix approach to neural networks. *The Annals of Applied Probability*, 28(2):1190–1248, 2018.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*, 2019.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Conference on Learning Theory*, pp. 2388–2464. PMLR, 2019.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. In *International Conference on Learning Representations*, 2019.

Behnam Neyshabur, Ryota Tomioka, Ruslan Salakhutdinov, and Nathan Srebro. Geometry of optimization and implicit regularization in deep learning. *arXiv preprint arXiv:1705.03071*, 2017.

Atsushi Nitanda and Taiji Suzuki. Optimal rates for averaged stochastic gradient descent under neural tangent kernel regime. In *International Conference on Learning Representations*, 2020.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, pp. 1177–1184, 2007.

Jason W Rocks and Pankaj Mehta. Memorizing without overfitting: Bias, variance, and interpolation in over-parameterized models. *arXiv preprint arXiv:2010.13933*, 2020.

Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features. In *Advances in Neural Information Processing Systems*, pp. 3215–3225, 2017.

Samuel L. Smith, Benoit Dherin, David Barrett, and Soham De. On the origin of implicit regularization in stochastic gradient descent. In *International Conference on Learning Representations*, 2020.

Aditya Varre, Loucas Pillaud-Vivien, and Nicolas Flammarion. Last iterate convergence of sgd for least-squares in the interpolation regime. *arXiv preprint arXiv:2102.03183*, 2021.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Christopher KI Williams. Computation with infinite neural networks. *Neural Computation*, 10(5):1203–1216, 1998.

Denny Wu and Ji Xu. On the optimal weighted ℓ_2 regularization in overparameterized linear regression. In *Advances in Neural Information Processing Systems*, pp. 10112–10123, 2020.

Zitong Yang, Yaodong Yu, Chong You, Jacob Steinhardt, and Yi Ma. Rethinking bias-variance trade-off for generalization of neural networks. In *International Conference on Machine Learning*, 2020.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

Difan Zou and Quanquan Gu. An improved analysis of training over-parameterized deep neural networks. *Advances in Neural Information Processing Systems*, 32:2055–2064, 2019.

Difan Zou, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, and Sham M Kakade. Benign overfitting of constant-stepsize sgd for linear regression. In *Conference on Learning Theory*, 2021.

APPENDIX

The outline of the appendix is stated as follows.

- Section A summarizes representative results on random features regarding to double descent under various settings.
- Section B provides the proofs of lemmas in Section 3.2 on statistical properties of Σ_m and $\widetilde{\Sigma}_m$.
- Appendix C introduces preliminaries on PSD operators in stochastic approximation.
- Section D provides estimation for several typical integrals that are needed for our proof.
- Section E gives error bounds for `Bias`.
- In Section F, we provide the error bounds for `Variance`.

A COMPARISONS WITH PREVIOUS WORK

According to the used data assumption, the type of solution, and the derived results, we summarize various representative approaches in Table 1.

Table 1: Comparison of problem settings on analysis of high dimensional random features on double descent.

	data assumption	solution	result
(Hastie et al., 2019)	Gaussian	closed-form	variance $\nearrow\searrow$
(Ba et al., 2020)	Gaussian	GD	variance $\nearrow\searrow$
Mei & Montanari (2019)	i.i.d on sphere	closed-form	variance, bias $\nearrow\searrow$
(d'Ascoli et al., 2020b)	Gaussian	closed-form	refined [2]
(Gerace et al., 2020)	Gaussian	closed-form	$\nearrow\searrow$
(Adlam & Pennington, 2020)	Gaussian	closed-form	refined
(Dhifallah & Lu, 2020)	Gaussian	closed-form	$\nearrow\searrow$
Hu & Lu (2020)	Gaussian	closed-form	$\nearrow\searrow$
(Liao et al., 2020)	general	closed-form	$\nearrow\searrow$
(Lin & Dobriban, 2021)	isotropic features with finite moments	closed form	refined
(Li et al., 2021)	correlated features with polynomial decay on Σ_d	closed form	interpolation learning
Ours	(at least) sub-exponential data	SGD	variance $\nearrow\searrow$, bias $\searrow$

[1] A refined decomposition on variance is conducted by sources of randomness on data sampling, initialization, label noise to possess each term d'Ascoli et al. (2020a) or their full decomposition in Adlam & Pennington (2020); Lin & Dobriban (2021).

We mainly discuss the used assumption on data distribution here. It can be found that, most papers assume the data to be Gaussian or uniformly distributed on the sphere. The following papers admit weaker assumption on data. Given a correlated features model that is commonly used in high dimensional statistics (Hastie et al., 2019):

$$\boldsymbol{x} = \Sigma_d^{\frac{1}{2}} \boldsymbol{t}\,, \quad \mathbb{E}[t_i] = 0, \mathbb{V}[t_i] = 1, \quad \text{with } \Sigma_d := \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}\boldsymbol{x}^\top]\,, \tag{14}$$

where $\boldsymbol{t} \in \mathbb{R}^d$ has i.i.d entries t_i $(i = 1, 2, \ldots, d)$ with zero mean and unit variance. In (Li et al., 2021), they further require that each entry is i.i.d sub-Gaussian and Σ_d admits polynomial decay on eigenvalues. Lin & Dobriban (2021) consider isotropic features with finite moment, i.e., taking $\Sigma_d := I$ in Eq. 14 and $\mathbb{E}[t_i^{8+\eta}] < \infty$ for any arbitrary positive constant $\eta > 0$. Our model holds for sub-Gaussian and sub-exponential data (at least), and thus the used data assumption 4 is weaker than them. In (Liao et al., 2020), it makes no assumption on data distribution but requires that test data "behave" statistically like the training data by concentrated random vectors. Indeed, their data assumption is weaker than ours, but their analysis framework builds on the exact closed-form solution from random matrix theory. Instead, we focus on the SGD setting and thus take a unified perspective on optimization and generalization.

B RESULTS ON COVARIANCE OPERATORS

In this section, we present the proofs of Lemma 1, 2, 3, 4 on statistical properties of Σ_m and $\widetilde{\Sigma}_m$.

B.1 PROOF OF LEMMA 1

Here we present the proof of Lemma 1 and then give two examples by taking different activation functions.

Proof. Recall the definition of $\widetilde{\Sigma}_m$, we have

$$\widetilde{\Sigma}_m := \mathbb{E}_{\boldsymbol{x},\boldsymbol{W}}[\varphi(\boldsymbol{x}) \otimes \varphi(\boldsymbol{x})] = \frac{1}{m}\mathbb{E}_{\boldsymbol{x},W_{ij}\sim\mathcal{N}(0,1)}\left[\sigma\left(\frac{\boldsymbol{W}\boldsymbol{x}}{\sqrt{d}}\right)\sigma\left(\frac{\boldsymbol{W}\boldsymbol{x}}{\sqrt{d}}\right)^\top\right] \in \mathbb{R}^{m\times m}\,.$$

We consider the diagonal and non-diagonal elements of $\widetilde{\Sigma}_m$ separately. Here we assume $\sigma(\cdot):\mathbb{R}\to\mathbb{R}$ with single-output for description simplicity, and the results can be easily extended to multiple-output cases, e.g., $\sigma(x) = [\cos(x), \sin(x)]^\top$ corresponding to the Gaussian kernel.

Diagonal element: The diagonal entry $(\widetilde{\Sigma}_m)_{ii} = \frac{1}{m}\mathbb{E}_{\boldsymbol{x},\boldsymbol{\omega}_i}[\sigma(\frac{\boldsymbol{\omega}_i^\top \boldsymbol{x}}{\sqrt{d}})\sigma(\frac{\boldsymbol{\omega}_i^\top \boldsymbol{x}}{\sqrt{d}})^\top] = \frac{1}{m}\mathbb{E}_{\boldsymbol{x}}\mathbb{E}_{\boldsymbol{\omega}}[\sigma(\frac{\boldsymbol{\omega}^\top \boldsymbol{x}}{\sqrt{d}})]^2$ is the same. In fact, $\mathbb{E}_{\boldsymbol{\omega}}\left[\sigma\left(\frac{\boldsymbol{\omega}^\top \boldsymbol{x}}{\sqrt{d}}\right)\right]^2$ is actually a one-dimensional integration by considering the basis $(e_1, e_2, \cdots, e_d)$ with $\boldsymbol{e}_1 = \boldsymbol{x}/\|\boldsymbol{x}\|_2$, and $\boldsymbol{e}_2, \cdots, \boldsymbol{e}_d$ any completion of the basis. This technique is commonly used in (Williams, 1998; Louart et al., 2018). The random feature $\boldsymbol{\omega}$ admits the coordinate representation $\boldsymbol{\omega} = \bar{\omega}_1\boldsymbol{e}_1 + \bar{\omega}_2\boldsymbol{e}_2 + \cdots + \bar{\omega}_d\boldsymbol{e}_d$, and thus

$$\boldsymbol{\omega}^\top\boldsymbol{x} = (\bar{\omega}_1\boldsymbol{e}_1 + \bar{\omega}_2\boldsymbol{e}_2 + \cdots + \bar{\omega}_d\boldsymbol{e}_d)^\top(\|\boldsymbol{x}\|\boldsymbol{e}_1) = \|\boldsymbol{x}\|\bar{\omega}_1\,,$$

which implies

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\omega}}\left[\sigma\left(\frac{\boldsymbol{\omega}^\top\boldsymbol{x}}{\sqrt{d}}\right)\right]^2 &= (2\pi)^{-\frac{d}{2}}\int_{\mathbb{R}^d}\left[\sigma\left(\frac{\boldsymbol{\omega}^\top\boldsymbol{x}}{\sqrt{d}}\right)\right]^2\exp\left(-\frac{1}{2}\|\boldsymbol{\omega}\|_2^2\right)\mathrm{d}\boldsymbol{\omega} = \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}}\left[\sigma\left(\frac{z\|\boldsymbol{x}\|}{\sqrt{d}}\right)\right]^2\exp(-z^2)\mathrm{d}z\\ &= \mathbb{E}_{z\sim\mathcal{N}(0,\|\boldsymbol{x}\|_2^2/d)}[\sigma(z)]^2\,.\end{aligned}$$

That means, $(\widetilde{\Sigma}_m)_{ii} = \frac{1}{m}\mathbb{E}_{\boldsymbol{x}}\mathbb{E}_{z\sim\mathcal{N}(0,\|\boldsymbol{x}\|_2^2/d)}[\sigma(z)]^2$.

Non-diagonal element: The non-diagonal entry $(\widetilde{\Sigma}_m)_{ij} = \frac{1}{m}\mathbb{E}_{\boldsymbol{x},\boldsymbol{\omega}_i,\boldsymbol{\omega}_j}[\sigma(\frac{\boldsymbol{\omega}_i^\top \boldsymbol{x}}{\sqrt{d}})\sigma(\frac{\boldsymbol{\omega}_j^\top \boldsymbol{x}}{\sqrt{d}})^\top] = \frac{1}{m}\mathbb{E}_{\boldsymbol{x}}[\mathbb{E}_{\boldsymbol{\omega}}\sigma(\frac{\boldsymbol{\omega}^\top \boldsymbol{x}}{\sqrt{d}})]^2$ is the same due to the independence between $\boldsymbol{\omega}_i$ and $\boldsymbol{\omega}_j$. Likewise, it can be represented as a one-dimensional integration

$$(\widetilde{\Sigma}_m)_{ij} = \frac{1}{m}\mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}_{\boldsymbol{\omega}}\sigma\left(\frac{\boldsymbol{\omega}^\top\boldsymbol{x}}{\sqrt{d}}\right)\right]^2 = \frac{1}{m}\mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}_{z\sim\mathcal{N}(0,1)}\sigma\left(\frac{z\|\boldsymbol{x}\|}{\sqrt{d}}\right)\right]^2 = \frac{1}{m}\mathbb{E}_{\boldsymbol{x}}\left(\mathbb{E}_{z\sim\mathcal{N}(0,\|\boldsymbol{x}\|_2^2/d)}[\sigma(z)]\right)^2\,.$$

Accordingly, by denoting $a := (\widetilde{\Sigma}_m)_{ii}$ and $b := (\widetilde{\Sigma}_m)_{ij}$, the covariance operator $\widetilde{\Sigma}_m$ can be represented as

$$\widetilde{\Sigma}_m = (a-b)I_m + b\mathbf{1}\mathbf{1}^\top \in \mathbb{R}^{m\times m}\,, \tag{15}$$

with its determinant $\det(\widetilde{\Sigma}_m) = (1 + \frac{mb}{a-b})(a-b)^m$. Hence, the eigenvalues of $\widetilde{\Sigma}_m$ can be naturally obtained by the matrix determinant lemma: $\widetilde{\lambda}_1(\widetilde{\Sigma}_m) = a - b + bm$ and the remaining eigenvalues are $a - b$, which concludes the proof. $\square$

Here we give the calculation details for the Gaussian kernel that corresponds to the $\sin/\cos$ activation function $\sigma(x) = [\cos(x), \sin(x)]^\top$ and arc-cosine kernel that corresponds to the ReLU function $\sigma(x) = \max\{0, x\}$.

Regarding to the Gaussian kernel, by virtue of $\mathbb{E}[\cos(\boldsymbol{a}^\top\boldsymbol{z})] = \cos(\boldsymbol{\mu}^\top\boldsymbol{z})\exp(-\frac{1}{2}\boldsymbol{z}^\top\boldsymbol{A}\boldsymbol{z})$ for $\boldsymbol{a}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{A})$ and $\boldsymbol{\omega}_i - \boldsymbol{\omega}_j \sim \mathcal{N}(\boldsymbol{0}, 2\boldsymbol{I}_d)$, we have (for the non-diagonal element)

$$(\widetilde{\Sigma}_m)_{ij} = \frac{1}{m}\mathbb{E}_{\boldsymbol{x},\boldsymbol{\omega}_i,\boldsymbol{\omega}_j}\left[\sigma(\frac{\boldsymbol{\omega}_i^\top\boldsymbol{x}}{\sqrt{d}})\sigma(\frac{\boldsymbol{\omega}_j^\top\boldsymbol{x}}{\sqrt{d}})^\top\right] = \frac{1}{m}\mathbb{E}_{\boldsymbol{x}}\mathbb{E}_{\boldsymbol{\omega}_i,\boldsymbol{\omega}_j}\left[\cos(\frac{\boldsymbol{x}^\top(\boldsymbol{\omega}_i-\boldsymbol{\omega}_j)}{\sqrt{d}})\right] = \frac{1}{m}\mathbb{E}_{\boldsymbol{x}}\exp\left(-\frac{\|\boldsymbol{x}\|_2^2}{d}\right)\,.$$

The diagonal element admits $(\widetilde{\Sigma}_m)_{ii} = 1/m$.

Regarding to the first-order arc-cosine kernel, denote $\widetilde{z} := \max\{0, z\}$ with $z\sim\mathcal{N}(0, \|\boldsymbol{x}\|_2^2/d)$, it is subject to the Rectified Gaussian distribution admitting (Li et al. (2021) also present this)

$$\mathbb{E}[\widetilde{z}] = \frac{\|\boldsymbol{x}\|_2}{\sqrt{2d\pi}}\,,\quad \mathbb{E}[\widetilde{z}]^2 = \frac{\|\boldsymbol{x}\|_2^2}{2d}\,,\quad \mathbb{V}[\widetilde{z}] = \frac{\|\boldsymbol{x}\|_2^2}{2d}\left(1-\frac{1}{\pi}\right)\,.$$

Accordingly, the diagonal element is

$$(\widetilde{\Sigma}_m)_{ii} = \mathbb{E}_{z\sim\mathcal{N}(0,\|\boldsymbol{x}\|_2^2/d)}[\sigma(z)]^2 = \frac{1}{2md}\mathbb{E}_{\boldsymbol{x}}\|\boldsymbol{x}\|_2^2 = \frac{1}{2md}\mathrm{Tr}(\Sigma_d)\,,$$

and the non-diagonal element is

$$(\widetilde{\Sigma}_m)_{ij} = \frac{1}{m}\mathbb{E}_{\boldsymbol{x}}\left(\mathbb{E}_{z\sim\mathcal{N}(0,\|\boldsymbol{x}\|_2^2/d)}[\sigma(z)]\right)^2 = \frac{1}{2md\pi}\mathrm{Tr}(\Sigma_d)\,,$$

with the covariance operator $\Sigma_d := \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}\boldsymbol{x}^\top]$.

B.2 Proof of Lemma 2

Proof. According to (Wainwright, 2019, Theorem 2.26), by virtue of the Lipschitz function $\sigma(\cdot)$ of Gaussian variables, we have

$$\mathbb{P}\left[\left|\sigma\left(\frac{\boldsymbol{\omega}^\top\boldsymbol{x}}{\sqrt{d}}\right) - \mathbb{E}_{\boldsymbol{\omega}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I}_d)}\sigma\left(\frac{\boldsymbol{\omega}^\top\boldsymbol{x}}{\sqrt{d}}\right)\right| \geqslant t\right] \leqslant c\exp(-t^2),\ \ \forall t\geqslant 0\,,$$

which implies that $\sigma\big(\frac{\boldsymbol{\omega}^\top\boldsymbol{x}}{\sqrt{d}}\big)$ is a sub-Gaussian random variable due to its expectation in the $\mathcal{O}(1)$ order. Accordingly, $\|\Sigma_m - \widetilde{\Sigma}_m\|_2$ is a sub-exponential random variable with

$$\begin{aligned}
\|\Sigma_m - \widetilde{\Sigma}_m\|_2 \leqslant \|\Sigma_m\|_2 + \|\widetilde{\Sigma}_m\|_2 &= \frac{1}{m}\left\|\mathbb{E}_{\boldsymbol{x}}\Big[\sigma\left(\frac{\boldsymbol{W}\boldsymbol{x}}{\sqrt{d}}\right)\sigma\left(\frac{\boldsymbol{W}\boldsymbol{x}}{\sqrt{d}}\right)^\top\Big]\right\|_2 + \mathcal{O}(1)\\
&\leqslant \frac{1}{m}\mathbb{E}_{\boldsymbol{x}}\left\|\sigma\left(\frac{\boldsymbol{W}\boldsymbol{x}}{\sqrt{d}}\right)\right\|_2^2 + \mathcal{O}(1) \quad \text{[Jensen's inequality]}\\
&\lesssim \frac{1}{m}\left(\mathbb{E}_{\boldsymbol{x}}\|\sigma(\boldsymbol{0}_m)\|_2^2 + \mathbb{E}_{\boldsymbol{x}}\left\|\frac{\boldsymbol{W}\boldsymbol{x}}{\sqrt{d}}\right\|_2^2\right) + \mathcal{O}(1) \quad [\sigma\text{: Lipschitz continuous}]\\
&\lesssim \mathcal{O}(1) + \frac{1}{md}\sum_{i=1}^m \boldsymbol{\omega}_i^\top\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}\boldsymbol{x}^\top]\boldsymbol{\omega}_i \quad [\text{using } \|\Sigma_d\|_2 < \infty]\\
&\lesssim \frac{1}{d}\|\boldsymbol{\omega}\|_2^2 \quad [\text{here } \boldsymbol{\omega}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I}_d)]\,,
\end{aligned}$$

where $\|\boldsymbol{\omega}\|_2^2$ is a $\chi^2(d)$ random variable, and thus $\|\Sigma_m - \widetilde{\Sigma}_m\|_2$ has sub-exponential norm $\mathcal{O}(1)$. Accordingly, the high moment $\mathbb{E}\|\Sigma_m\|_2^p < \infty$ holds for finite p. Following the above derivation, we can also conclude that $\mathrm{Tr}(\Sigma_m)$ has the sub-exponential norm $\mathcal{O}(1)$, i.e.

$$\mathrm{Tr}(\Sigma_m) = \frac{1}{m}\mathbb{E}_{\boldsymbol{x}}\mathrm{Tr}\left[\sigma\left(\frac{\boldsymbol{W}\boldsymbol{x}}{\sqrt{d}}\right)\sigma\left(\frac{\boldsymbol{W}\boldsymbol{x}}{\sqrt{d}}\right)^\top\right] = \frac{1}{m}\mathbb{E}_{\boldsymbol{x}}\left\|\sigma\left(\frac{\boldsymbol{W}\boldsymbol{x}}{\sqrt{d}}\right)\right\|_2^2 \lesssim \frac{1}{d}\|\boldsymbol{\omega}\|_2^2\,.$$

Likewise, we can derive $\mathrm{Tr}(\Sigma_m^2) < \infty$ in the similar fashion. Besides, our work needs the error bound for the smallest eigenvalue of Σ_m, i.e., λ_m. By virtue of Schur–Horn theorem (Horn, 1954), λ_m admits

$$\lambda_m \leqslant \min_{i\in\{1,2,\dots,m\}}(\Sigma_m)_{ii} \sim \mathcal{O}\left(\frac{1}{m}\right)\,.$$

□

B.3 Proof of Lemma 3

Proof. The remark in Lemma 1 demonstrates that $\widetilde{\Sigma}_m$ is positive definite and thus $\widetilde{\Sigma}_m^{-1}$ exists. Recall the formulation of $\widetilde{\Sigma}_m$ in Eq. 15 with $a := (\widetilde{\Sigma}_m)_{ii}$ and $b := (\widetilde{\Sigma}_m)_{ij}$, by virtue of Sherman–Morrison formula, the inverse of $\widetilde{\Sigma}_m$ is

$$(\widetilde{\Sigma}_m)^{-1} = \frac{1}{a-b}\left(I_m - \frac{b}{a-b+bm}\boldsymbol{1}\boldsymbol{1}^\top\right) = \frac{1}{\widetilde{\lambda}_2}\left(I_m - \frac{b}{\widetilde{\lambda}_1}\boldsymbol{1}\boldsymbol{1}^\top\right)\,.$$

Accordingly, we have

$$
\begin{aligned}
\left\| \widetilde{\Sigma}_m^{-2} \mathbb{E}_{\boldsymbol{W}} (\Sigma_m^2) \right\|_2 &= \left\| \mathbb{E}_{\boldsymbol{W}} [\widetilde{\Sigma}_m^{-2} \Sigma_m^2] \right\|_2 \leqslant \mathbb{E}_{\boldsymbol{W}} \left\| \widetilde{\Sigma}_m^{-1} \Sigma_m \right\|_2^2 \\
&= \mathbb{E}_{\boldsymbol{W}} \left\| \frac{1}{m\widetilde{\lambda}_2} \left(I_m - \frac{b}{\widetilde{\lambda}_1} \mathbf{1}\mathbf{1}^\top \right) \mathbb{E}_{\boldsymbol{x}} \left[\sigma\left(\frac{\boldsymbol{W}\boldsymbol{x}}{\sqrt{d}} \right) \sigma\left(\frac{\boldsymbol{W}\boldsymbol{x}}{\sqrt{d}} \right)^\top \right] \right\|_2^2 \\
&\lesssim \left\| I_m - \frac{b}{\widetilde{\lambda}_1} \mathbf{1}\mathbf{1}^\top \right\|_2^2 \mathbb{E}_{\boldsymbol{W}} \left\| \mathbb{E}_{\boldsymbol{x}} \left[\sigma\left(\frac{\boldsymbol{W}\boldsymbol{x}}{\sqrt{d}} \right) \sigma\left(\frac{\boldsymbol{W}\boldsymbol{x}}{\sqrt{d}} \right)^\top \right] \right\|_2^2 \quad \text{[using } \widetilde{\lambda}_2 \sim \mathcal{O}(\tfrac{1}{m}) \text{ in Lemma 1]} \\
&\lesssim \left\| I_m - \frac{b}{\widetilde{\lambda}_1} \mathbf{1}\mathbf{1}^\top \right\|_2^2 \mathbb{E}_{\boldsymbol{W}} \left(\frac{\|\boldsymbol{\omega}\|_2^2}{d} \right)^2 \quad \text{[here } \boldsymbol{\omega} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d) \text{ in Lemma 2]} \\
&\leqslant \mathbb{E}_{\boldsymbol{W}} \left(\frac{\|\boldsymbol{\omega}\|_2^2}{d} \right)^2 \sim \mathcal{O}(1) \,,
\end{aligned}
$$

due to $b \sim \mathcal{O}(1/m)$, $\widetilde{\lambda}_1 \sim \mathcal{O}(1)$ in Lemma 1, and $\|\boldsymbol{\omega}\|_2^2 \sim \chi^2(d)$. □

B.4 Proof of Lemma 4

Proof. The first inequality naturally holds, and so we focus on the second inequality. Denote $\Phi := \mathbb{E}_{\boldsymbol{x},\boldsymbol{W}}[\varphi(\boldsymbol{x}) \otimes \varphi(\boldsymbol{x}) \otimes \varphi(\boldsymbol{x}) \otimes \varphi(\boldsymbol{x})]$, its diagonal elements are the same

$$
\Phi_{ii} = \frac{m-1}{m^2} \mathbb{E}_{\boldsymbol{x}} \left(\mathbb{E}_{z \sim \mathcal{N}(0, \|\boldsymbol{x}\|_2^2/d)} [\sigma(z)]^2 \right)^2 + \frac{1}{m^2} \mathbb{E}_{\boldsymbol{x}} \mathbb{E}_{z \sim \mathcal{N}(0, \|\boldsymbol{x}\|_2^2/d)} [\sigma(z)]^4 \sim \mathcal{O}\left(\frac{1}{m} \right) .
$$

Its non-diagonal elements Φ_{ij} with $i \neq j$ are the same

$$
\Phi_{ij} = \frac{m-3}{m^2} \mathbb{E}_{\boldsymbol{x}} \left[\left(\mathbb{E}_{z \sim \mathcal{N}(0, \|\boldsymbol{x}\|_2^2/d)} [\sigma(z)] \right)^2 \mathbb{E}_{z \sim \mathcal{N}(0, \|\boldsymbol{x}\|_2^2/d)} [\sigma(z)]^2 \right] + \frac{2}{m^2} \mathbb{E}_{\boldsymbol{x}} \left[\mathbb{E}_{z \sim \mathcal{N}(0, \|\boldsymbol{x}\|_2^2/d)} [\sigma(z)]^3 \mathbb{E}_{z \sim \mathcal{N}(0, \|\boldsymbol{x}\|_2^2/d)} [\sigma(z)] \right] ,
$$

where the first term is in $\mathcal{O}(\frac{1}{m})$ order and the second term is in $\mathcal{O}(\frac{1}{m^2})$ order. By denoting $a := (\widetilde{\Sigma}_m)_{ii}$, $b := (\widetilde{\Sigma}_m)_{ij}$ as given by Lemma 1, $A := \Phi_{ii}$, and $B := \Phi_{ij}$, the operator $r\mathrm{Tr}(\widetilde{\Sigma}_m)\widetilde{\Sigma}_m - \Phi$ can be represented as

$$
r\mathrm{Tr}(\widetilde{\Sigma}_m)\Sigma_m - \Phi = [rm(a-b) - A + B]\, I_m + (rmab - B)\mathbf{1}\mathbf{1}^\top ,
$$

of which the smallest eigenvalue is $rma(a-b) - A + B$. Accordingly, to ensure the positive definiteness of $r\mathrm{Tr}(\widetilde{\Sigma}_m)\widetilde{\Sigma}_m - \Phi$, which implies $\mathbb{E}_{\boldsymbol{W}} \Big(\mathbb{E}_{\boldsymbol{x}} \Big([\varphi(\boldsymbol{x}) \otimes \varphi(\boldsymbol{x})] A [\varphi(\boldsymbol{x}) \otimes \varphi(\boldsymbol{x})] \Big) \Big) \preccurlyeq r\mathrm{Tr}(\widetilde{\Sigma}_m)\widetilde{\Sigma}_m$, we require its smallest eigenvalue is non-negative, i.e., $rma(a-b) - A + B \geqslant 0$. That means, r should satisfies

$$
r \geqslant \frac{A-B}{ma(a-b)} = \frac{A-B}{\frac{1}{m} \mathbb{E}_{\boldsymbol{x}} \mathbb{E}_{z \sim \mathcal{N}(0, \|\boldsymbol{x}\|_2^2/d)} [\sigma(z)]^2 \mathbb{E}_{\boldsymbol{x}} \mathbb{V}[\sigma(z)]} . \tag{16}
$$

Since $A - B$ admits

$$
A - B \leqslant \frac{1}{m} \mathbb{E}_{\boldsymbol{x}} \mathbb{E}_z [\sigma(z)]^2 \mathbb{E}_{\boldsymbol{x}} \mathbb{V}[\sigma(z)] + \mathcal{O}\left(\frac{1}{m^2} \right) ,
$$

then by taking $r := 1 + \mathcal{O}\left(\frac{1}{m}\right)$, the condition in Eq. 16 satisfies, and thus $r\mathrm{Tr}(\widetilde{\Sigma}_m)\widetilde{\Sigma}_m - \Phi$ is positive definite, which concludes the proof. □

C Preliminaries on PSD operators

In this section, we first define some stochastic/deterministic PSD operators that follow (Jain et al., 2017; Zou et al., 2021) in stochastic approximation, and then present Lemma 5 that is based on PSD operators and is needed to estimate `B1` and `V1`. Note that, the PSD operators will make the notation in our proof simple and clarity but do not change the proof itself.

Following (Jain et al., 2017; Zou et al., 2021), we define several stochastic PSD operators as below. Given the random features matrix $\boldsymbol{W}$, define (for any PSD operator A)

$$S^{\mathtt{W}} := \mathbb{E}_{\boldsymbol{x}}[\varphi(\boldsymbol{x}) \otimes \varphi(\boldsymbol{x}) \otimes \varphi(\boldsymbol{x}) \otimes \varphi(\boldsymbol{x})], \quad \widetilde{S}^{\mathtt{W}} := \Sigma_m \otimes \Sigma_m \,,$$
$$S^{\mathtt{W}} \circ A := \mathbb{E}_{\boldsymbol{x}} \left[\varphi(\boldsymbol{x})^{\top} \varphi(\boldsymbol{x}) A \varphi(\boldsymbol{x}) \otimes \varphi(\boldsymbol{x})\right], \quad \widetilde{S}^{\mathtt{W}} \circ A := \Sigma_m A \Sigma_m \,.$$

Besides, for any γ_t ($t = 1, 2, \dots, n$), define the following operators

$$(I - \gamma_t T^{\mathtt{W}}) \circ A := \mathbb{E}_{\boldsymbol{x}}[I - \gamma_t \varphi(\boldsymbol{x}) \otimes \varphi(\boldsymbol{x})] A [I - \gamma_t \varphi(\boldsymbol{x}) \otimes \varphi(\boldsymbol{x})], \quad (I - \gamma_t \widetilde{T}^{\mathtt{W}}) \circ A := (I - \gamma_t \Sigma_m) A (I - \gamma_t \Sigma_m) \,,$$

associated with two corresponding operators (that depend on γ_t)

$$T^{\mathtt{W}} := \Sigma_m \otimes I + I \otimes \Sigma_m - \gamma_t S^{\mathtt{W}}, \quad \widetilde{T}^{\mathtt{W}} := \Sigma_m \otimes I + I \otimes \Sigma_m - \gamma_t \widetilde{S}^{\mathtt{W}} \,.$$

Clearly, the above operators $S^{\mathtt{W}}$, $\widetilde{S}^{\mathtt{W}}$, $(I - \gamma_t T^{\mathtt{W}})$, $(I - \gamma_t \widetilde{T}^{\mathtt{W}})$, $T^{\mathtt{W}}$, and $\widetilde{T}^{\mathtt{W}}$ are PSD, and $S^{\mathtt{W}} \succcurlyeq \widetilde{S}^{\mathtt{W}}$. The proof is similar to (Zou et al., 2021, Lemma B.1) and thus we omit it here.

Further, if $\gamma_0 < 1/\mathrm{Tr}(\Sigma_m)$, $I - \gamma_i \Sigma_m$ ($i = 1, 2, \dots, n$) is a contraction map, and thus for any PSD operator A and step-size γ_i, the following exists

$$\sum_{t=0}^{\infty} (I - \gamma_i \widetilde{T}^{\mathtt{W}})^t \circ A = \sum_{t=0}^{\infty} (I - \gamma_i \Sigma_m)^t A (I - \gamma_i \Sigma_m)^t \,.$$

Hence, $(\widetilde{T}^{\mathtt{W}})^{-1} := \gamma_i \sum_{t=0}^{\infty} (I - \gamma_i \widetilde{T}^{\mathtt{W}})^t$ exists and PSD.

Based on the above stochastic operators, we define several deterministic PSD ones by taking the expectation over $\boldsymbol{W}$ as below. For any given γ_i ($i = 1, 2, \dots, n$), we have the following PSD operators

$$S := \mathbb{E}_{\boldsymbol{W}}[\Sigma_m \otimes \Sigma_m], \quad \widetilde{S} := \widetilde{\Sigma}_m \otimes \widetilde{\Sigma}_m \,,$$
$$T := \widetilde{\Sigma}_m \otimes I + I \otimes \widetilde{\Sigma}_m - \gamma_i S, \quad \widetilde{T} := \widetilde{\Sigma}_m \otimes I + I \otimes \widetilde{\Sigma}_m - \gamma_i \widetilde{S} \,,$$
$$S \circ A := \mathbb{E}_{\boldsymbol{W}}[\Sigma_m A \Sigma_m], \quad \widetilde{S} \circ A := \widetilde{\Sigma}_m A \widetilde{\Sigma}_m \,,$$
$$(I - \gamma_i T) \circ A := \mathbb{E}_{\boldsymbol{W}}[(I - \gamma_i \Sigma_m) A (I - \gamma \Sigma_m)], \quad (I - \gamma_i \widetilde{T}) \circ A := (I - \gamma_i \widetilde{\Sigma}_m) A (I - \gamma_i \widetilde{\Sigma}_m) \,,$$

which implies $\widetilde{T} - T = \gamma_i (S - \widetilde{S})$.

Based on the above PSD operators, we present a lemma here that is used to estimate B1 and V1.[5]

Lemma 5. *Under Assumptions 1, 2, 3, 4 with $r' \geqslant 1$, denote*

$$D_t^{\mathtt{v-X}} := \sum_{s=1}^{t} \prod_{i=s+1}^{t} (I - \gamma_i T^{\mathtt{W}}) \circ \gamma_s^2 B \Sigma_m \,, \tag{17}$$

with a scalar B independent of k, if the step-size $\gamma_t := \gamma_0 t^{-\zeta}$ with $\zeta \in [0, 1)$ satisfies

$$\gamma_0 < \min \left\{ \frac{1}{r' \mathrm{Tr}(\Sigma_m)}, \frac{1}{2\mathrm{Tr}(\Sigma_m)} \right\} \,,$$

then $D_t^{\mathtt{v-X}}$ can be upper bounded by

$$D_t^{\mathtt{v-X}} \preccurlyeq \frac{\gamma_0 B}{1 - \gamma_0 r' \mathrm{Tr}(\Sigma_m)} I \,.$$

Remark: The PSD operator $I - \gamma_i T^{\mathtt{W}}$ cannot be guaranteed as a contraction map since we cannot directly choose $\gamma_0 < \frac{1}{\mathrm{Tr}[\varphi(\boldsymbol{x})\varphi(\boldsymbol{x})^{\top}]}$ for general data $\boldsymbol{x}$. However, its summation in Eq. 17 can be still bounded by our lemma. In our work, we set $B := r' \mathrm{Tr}(\Sigma_m)$ for estimate B1, and $B := \tau^2 r' \gamma_0 [\mathrm{Tr}(\Sigma_m) + \gamma_0 \mathrm{Tr}(\Sigma_m^2)]$ to bound V1, respectively.

Proof. Our proof can be divided into two parts: one part is to prove $\mathrm{Tr}[D_t^{\mathtt{v-X}}] \leqslant \mathrm{Tr}[D_t^{\mathtt{v-X}}(0)]$ for any $\zeta \in [0, 1)$; the other part is to provide the upper bound of $D_t^{\mathtt{v-X}}(0)$. We focus on the first part and the proof in the second part follows (Jain et al., 2017, Lemmas 3 and 5) and (Zou et al., 2021, Lemma B.4).

[5] Our proofs on the remaining quantities including V2, V3, B2, B3 do not use PSD operators.

Denote the constant step-size setting (special case) with $\zeta = 0$ for $D_t^{\texttt{v-X}}$ as

$$D_t^{\texttt{v-X}}(0) := \sum_{s=1}^{t} (I - \gamma_0 T^{\texttt{W}})^{t-s} \circ \gamma_0^2 B \Sigma_m \,.$$

The quantity $\mathrm{Tr}[D_t^{\texttt{v-X}}]$ admits the following representation by the definition of $I - \gamma_i T^{\texttt{W}}$

$$\begin{aligned}
\mathrm{Tr}[D_t^{\texttt{v-X}}] &= \sum_{s=1}^{t} \prod_{i=s+1}^{t} \mathrm{Tr}\left[(I - \gamma_i T^{\texttt{W}}) \circ \gamma_s^2 B \Sigma_m\right] \\
&= \sum_{s=1}^{t} B\gamma_s^2 \prod_{i=s+1}^{t} \mathrm{Tr}\Bigg(\mathbb{E}_{\boldsymbol{x}}[I - \gamma_i \varphi(\boldsymbol{x}) \otimes \varphi(\boldsymbol{x})] \Sigma_m [I - \gamma_i \varphi(\boldsymbol{x}) \otimes \varphi(\boldsymbol{x})] \Bigg) \\
&= B \sum_{s=1}^{t} \gamma_s^2 \prod_{i=s+1}^{t} \mathrm{Tr}\Bigg(\Sigma_m - 2\gamma_i \Sigma_m^2 + \gamma_i^2 \Sigma_m \mathbb{E}_{\boldsymbol{x}}\left[\varphi(\boldsymbol{x}) \otimes \varphi(\boldsymbol{x}) \otimes \varphi(\boldsymbol{x}) \otimes \varphi(\boldsymbol{x})\right] \Bigg) \,.
\end{aligned}$$

Based on the above results, we have

$$\begin{aligned}
\mathrm{Tr}[D_t^{\texttt{v-X}}(0)] - \mathrm{Tr}[D_t^{\texttt{v-X}}] &= B \sum_{s=1}^{t} \prod_{i=s+1}^{t} \mathrm{Tr}\Bigg(\Sigma_m \Big[(\gamma_0^2 - \gamma_s^2) I - 2(\gamma_0^3 - \gamma_s^2 \gamma_i) \Sigma_m \\
&\qquad + (\gamma_0^4 - \gamma_i^2 \gamma_s^2) \mathbb{E}_{\boldsymbol{x}}\left[\varphi(\boldsymbol{x}) \otimes \varphi(\boldsymbol{x}) \otimes \varphi(\boldsymbol{x}) \otimes \varphi(\boldsymbol{x})\right] \Big] \Bigg) \\
&\geqslant B \sum_{s=1}^{t} \prod_{i=s+1}^{t} \mathrm{Tr}\Bigg(\Sigma_m \Big[(\gamma_0^2 - \gamma_s^2) I - 2(\gamma_0^3 - \gamma_s^2 \gamma_i) \Sigma_m + (\gamma_0^4 - \gamma_i^2 \gamma_s^2) \Sigma_m^2 \Big] \Bigg) \\
&= B \sum_{s=1}^{t} \prod_{i=s+1}^{t} \sum_{j=1}^{m} \Bigg(\lambda_j \Big[(\gamma_0^2 - \gamma_s^2) - 2(\gamma_0^3 - \gamma_s^2 \gamma_i) \lambda_j + (\gamma_0^4 - \gamma_i^2 \gamma_s^2) \lambda_j^2 \Big] \Bigg) \\
&= B \sum_{s=1}^{t} \prod_{i=s+1}^{t} \sum_{j=1}^{m} \Bigg(\lambda_j \Big[(\gamma_0^4 - \gamma_i^2 \gamma_s^2) \left(\lambda_j - \frac{\gamma_0^3 - \gamma_s^2 \gamma_i}{\gamma_0^4 - \gamma_i^2 \gamma_s^2} \right)^2 - \frac{\gamma_0^2 \gamma_s^2 (\gamma_0 - \gamma_i)^2}{\gamma_0^4 - \gamma_i^2 \gamma_s^2} \Big] \Bigg) \,.
\end{aligned}$$

Accordingly, $\mathrm{Tr}[D_t^{\texttt{v-X}}(0)] - \mathrm{Tr}[D_t^{\texttt{v-X}}] \geqslant 0$ holds if $\lambda_i \leqslant \frac{\gamma_0^3 - \gamma_s^2 \gamma_i - \gamma_0^2 \gamma_s + \gamma_0 \gamma_s^2}{\gamma_0^4 - \gamma_s^2 \gamma_i^2}$. This condition can be satisfied by

$$\lambda_i \leqslant \mathrm{Tr}(\Sigma_m) \leqslant \frac{1}{2\gamma_0} \leqslant \frac{\gamma_0^3 - \gamma_s^2 \gamma_i - \gamma_0^2 \gamma_s + \gamma_0 \gamma_s^2}{\gamma_0^4 - \gamma_s^2 \gamma_i^2} \,,$$

where the second inequality holds by $\gamma_0 \leqslant \frac{1}{2\mathrm{Tr}(\Sigma_m)}$.

In the next, we give the upper bound for $D_t^{\texttt{v-X}}(0)$. The proof follows (Jain et al., 2017, Lemmas 3 and 5) and (Zou et al., 2021, Lemma B.4). We just present it here for completeness. We firstly demonstrate that $D_t^{\texttt{v-X}}(0)$ is increasing and bounded, which implies that the limit $D_\infty^{\texttt{v-X}}(0)$ exists, and then we seek for the upper bound of this limit. To be specific, $D_t^{\texttt{v-X}}(0)$ admits the following expression

$$D_t^{\texttt{v-X}}(0) := \sum_{k=1}^{t-1} (I - \gamma_0 T^{\texttt{W}})^k \circ \gamma_0^2 B \Sigma_m = D_{t-1}^{\texttt{v-X}}(0) + (I - \gamma_0 T^{\texttt{W}})^{t-1} \circ \gamma_0^2 B \Sigma_m \succcurlyeq D_{t-1}^{\texttt{v-X}}(0) \,,$$

which implies that $D_t^{\texttt{v-X}}(0)$ is increasing.

Let $A_t := (I - \gamma_0 T^{\texttt{W}})^{t-1} \circ B\Sigma_m$, and then $A_t = (I - \gamma_0 T^{\texttt{W}}) \circ A_{t-1}$. We have

$$\begin{aligned}
\mathrm{Tr}(A_t) &= \mathrm{Tr}[(I - \gamma_0 T^{\texttt{W}}) \circ A_{t-1}] = \mathrm{Tr}(A_{t-1}) - 2\gamma_0 \mathrm{Tr}(\Sigma_m A_{t-1}) + \gamma_0^2 \mathrm{Tr}(S^{\texttt{W}} \circ A_{t-1}) \\
&\leqslant \mathrm{Tr}(A_{t-1}) - 2\gamma_0 \mathrm{Tr}(\Sigma_m A_{t-1}) + \gamma_0^2 r' \mathrm{Tr}(\Sigma_m A_{t-1}) \mathrm{Tr}(\Sigma_m) \quad \text{[using Assumption 4]} \\
&\leqslant \mathrm{Tr}[(I - \gamma_0 \Sigma_m) A_{t-1}] \leqslant (1 - \gamma_0 \lambda_m) \mathrm{Tr}(A_{t-1}) \,, \quad \text{[using } \gamma_0 \leqslant \tfrac{1}{r' \mathrm{Tr}(\Sigma_m)}\text{]}
\end{aligned}$$

which implies

$$\mathrm{Tr}[D_t^{\texttt{v-X}}(0)] \leqslant \gamma_0^2 \sum_{t=0}^{\infty} \mathrm{Tr}\left((I - \gamma_0 T^{\texttt{W}})^t \circ B\Sigma_m \right) \leqslant \mathrm{Tr}(B\Sigma_m) \sum_{t=0}^{\infty} (1 - \gamma_0 \lambda_m)^t \leqslant \frac{\gamma_0 \mathrm{Tr}(B\Sigma_m)}{\lambda_m} < \infty \,.$$

Accordingly, the monotonicity and boundedness of $\{D_t^{\mathtt{v-X}}(0)\}_{t=0}^{\infty}$ implies that the limit exists, denoted as $D_{\infty}^{\mathtt{v-X}}(0)$ with

$$D_{\infty}^{\mathtt{v-X}}(0) = (I - \gamma_0 T^{\mathtt{W}}) \circ D_{\infty}^{\mathtt{v-X}}(0) + \gamma_0^2 B \Sigma_m \,,$$

which implies $D_{\infty}^{\mathtt{v-X}}(0) = \gamma_0 (T^{\mathtt{W}})^{-1} \circ B\Sigma_m$ Accordingly, we have

$$\begin{aligned}
\widetilde{T}^{\mathtt{W}} \circ D_{\infty}^{\mathtt{v-X}}(0) &= T^{\mathtt{W}} \circ D_{\infty}^{\mathtt{v-X}}(0) + \gamma_0 S^{\mathtt{W}} \circ D_{\infty}^{\mathtt{v-X}}(0) - \gamma_0 \widetilde{S}^{\mathtt{W}} \circ D_{\infty}^{\mathtt{v-X}}(0) \quad [\text{definition of } \widetilde{T}^{\mathtt{W}}] \\
&= \gamma_0 B\Sigma_m + \gamma_0 S^{\mathtt{W}} \circ D_{\infty}^{\mathtt{v-X}}(0) - \gamma_0 \widetilde{S}^{\mathtt{W}} \circ D_{\infty}^{\mathtt{v-X}}(0) \\
&\preccurlyeq \gamma_0 B\Sigma_m + \gamma_0 S^{\mathtt{W}} \circ D_{\infty}^{\mathtt{v-X}}(0) \,. \quad [\text{using } S^{\mathtt{W}} \succcurlyeq \widetilde{S}^{\mathtt{W}}]
\end{aligned} \tag{18}$$

Besides, $(\widetilde{T}^{\mathtt{W}})^{-1} \circ \Sigma_m$ can be bounded by

$$\begin{aligned}
(\widetilde{T}^{\mathtt{W}})^{-1} \circ \Sigma_m = \gamma_0 \sum_{t=0}^{\infty} (I - \gamma_0 \widetilde{T}^{\mathtt{W}}) \circ \Sigma_m &= \gamma_0 \sum_{t=0}^{\infty} (I - \gamma_0 \Sigma_m)^t \Sigma_m (I - \gamma_0 \Sigma_m)^t \\
&\preccurlyeq \gamma_0 \sum_{t=0}^{\infty} (I - \gamma_0 \Sigma_m)^t \Sigma_m = I \,. \quad [\text{using } \gamma_0 \leqslant 1/\mathrm{Tr}(\Sigma_m)]
\end{aligned} \tag{19}$$

Therefore, $D_{\infty}^{\mathtt{v-X}}(0)$ can be further upper bounded by

$$\begin{aligned}
D_{\infty}^{\mathtt{v-X}}(0) &\preccurlyeq \gamma_0 (\widetilde{T}^{\mathtt{W}})^{-1} \circ B\Sigma_m + \gamma_0 (\widetilde{T}^{\mathtt{W}})^{-1} \circ S^{\mathtt{W}} \circ D_{\infty}^{\mathtt{v-X}}(0) \quad [\text{using Eq. 18}] \\
&\preccurlyeq \gamma_0 B + \gamma_0 (\widetilde{T}^{\mathtt{W}})^{-1} \circ S^{\mathtt{W}} \circ D_{\infty}^{\mathtt{v-X}}(0) \quad [\text{using Eq. 19}] \\
&= \gamma_0 B \sum_{t=0}^{\infty} [\gamma_0 (\widetilde{T}^{\mathtt{W}})^{-1} \circ S^{\mathtt{W}}]^t \circ I \quad [\text{using telescopic sum}] \\
&\preccurlyeq \gamma_0 B \sum_{t=0}^{\infty} \left(\gamma_0 (\widetilde{T}^{\mathtt{W}})^{-1} \circ S^{\mathtt{W}} \right)^{t-1} \circ \gamma_0 (\widetilde{T}^{\mathtt{W}})^{-1} \circ S^{\mathtt{W}} \circ I \\
&\preccurlyeq \gamma_0 B \sum_{t=0}^{\infty} \left(\gamma_0 (\widetilde{T}^{\mathtt{W}})^{-1} \circ S^{\mathtt{W}} \right)^{t-1} \circ \gamma_0 (\widetilde{T}^{\mathtt{W}})^{-1} \circ \mathrm{Tr}(\Sigma_m) \Sigma_m \quad [\text{using Assumption 4}] \\
&\preccurlyeq \gamma_0 B \sum_{t=0}^{\infty} \left[\gamma_0 r' \mathrm{Tr}(\Sigma_m) \right]^t \circ I \quad [\text{using Eq. 19}] \\
&\preccurlyeq \frac{\gamma_0 B}{1 - \gamma_0 r' \mathrm{Tr}(\Sigma_m)} I \,. \quad [\text{using } \gamma_0 < \tfrac{1}{r' \mathrm{tr}(\Sigma_m)}]
\end{aligned} \tag{20}$$

Hence, based on the above results, $D_t^{\mathtt{v-X}}(0)$ can be further upper bounded by

$$\begin{aligned}
D_t^{\mathtt{v-X}}(0) &= (I - \gamma_0 T^{\mathtt{W}}) \circ D_{t-1}^{\mathtt{v-X}}(0) + \gamma_0^2 B \Sigma_m \\
&= (I - \gamma_0 \widetilde{T}^{\mathtt{W}}) \circ D_{t-1}^{\mathtt{v-X}}(0) + \gamma_0^2 (S^{\mathtt{W}} - \widetilde{S}^{\mathtt{W}}) \circ D_{t-1}^{\mathtt{v-X}} + \gamma_0^2 B \Sigma_m \\
&\preccurlyeq (I - \gamma_0 \widetilde{T}^{\mathtt{W}}) \circ D_{t-1}^{\mathtt{v-X}}(0) + \gamma_0^2 S^{\mathtt{W}} \circ D_{\infty}^{\mathtt{v-X}}(0) + \gamma_0^2 B \Sigma_m \\
&\preccurlyeq (I - \gamma_0 \widetilde{T}^{\mathtt{W}}) \circ D_{t-1}^{\mathtt{v-X}}(0) + \gamma_0^2 r' \mathrm{Tr}[D_{\infty}^{\mathtt{v-X}}(0)] \mathrm{Tr}(\Sigma_m) \Sigma_m + \gamma_0^2 B \Sigma_m \quad [\text{using Assumption 4}] \\
&\preccurlyeq (I - \gamma_0 \widetilde{T}^{\mathtt{W}}) \circ D_{t-1}^{\mathtt{v-X}}(0) + \gamma_0^2 B \Sigma_m \left(\frac{\mathrm{Tr}(\Sigma_m) r' \gamma_0}{1 - \gamma_0 r' \mathrm{Tr}(\Sigma_m)} + 1 \right) \\
&\preccurlyeq \gamma_0^2 B \left(\frac{\mathrm{Tr}(\Sigma_m) r' \gamma_0}{1 - \gamma_0 r' \mathrm{Tr}(\Sigma_m)} + 1 \right) \sum_{k=0}^{\infty} (I - \gamma_0 \Sigma_m)^k \Sigma_m \\
&\preccurlyeq \gamma_0 B \left(\frac{\mathrm{Tr}(\Sigma_m) r' \gamma_0}{1 - \gamma_0 r' \mathrm{Tr}(\Sigma_m)} + 1 \right) I \,,
\end{aligned} \tag{21}$$

which concludes the proof. □

D SOME USEFUL INTEGRALS ESTIMATION

In this section, we present the estimation for the following integrals that will be needed in our proof by denoting $\kappa := 1 - \zeta \in (0, 1]$.

Integral 1: the following integral admits

$$
\begin{aligned}
\int_1^t u^{-\zeta} \exp\left(-c\frac{u^{1-\zeta}-1}{1-\zeta}\right) \mathrm{d}u &= \frac{1}{c}\int_0^{[\frac{c}{\kappa}(t^\kappa-1)]^{\frac{1}{\kappa}}} u^{-\zeta} u^{1-\kappa} \kappa v^{\kappa-1} \exp(-v^\kappa) \mathrm{d}v \\
&\leqslant \frac{1}{c}\int_0^\infty \exp(-x)\mathrm{d}x = \left(\frac{1}{c} \wedge t\right) ,
\end{aligned}
\tag{22}
$$

where we directly obtain an exact estimation t. Here we change the integral variable $v^\kappa := c\frac{u^{1-\zeta}-1}{1-\zeta}$ and

$$
\frac{\mathrm{d}v}{\mathrm{d}u} = u^{1-\kappa}\left(\frac{\kappa}{c}\right)^{\frac{1}{\kappa}} (u^\kappa - 1)^{\frac{\kappa-1}{\kappa}} = \frac{1}{c} u^{1-\kappa} \kappa v^{\kappa-1} .
$$

Accordingly, if we take $\zeta = 0$ in Eq. 22, we have

$$
\int_1^t \exp\left(-c\frac{u^{1-\zeta}-1}{1-\zeta}\right) \mathrm{d}u \leqslant \left(\frac{1}{c}t^\zeta \wedge t\right) .
\tag{23}
$$

Similar to Eq. 23, we have

$$
\int_t^n \exp\left(-\widetilde{\lambda}_i \gamma_0 \frac{u^{1-\zeta}-t^{1-\zeta}}{1-\zeta}\right) \mathrm{d}u \leqslant (n-t) \wedge \frac{n^\zeta}{\widetilde{\lambda}_i \gamma_0} .
\tag{24}
$$

Integral 2: we consider the following integral

$$
\begin{aligned}
&\int_1^t u^{-\zeta} \exp\left(-c\frac{(t+1)^{1-\zeta}-(u+1)^{1-\zeta}}{1-\zeta}\right) \mathrm{d}u \\
&= \frac{(t+1)^{1-\kappa}}{c} \int_0^C [(t+1)(1-x)^{\frac{1}{\kappa}} - 1]^{\kappa-1} (1-x)^{\frac{1-\kappa}{\kappa}} \kappa v^{\kappa-1} \exp(-v^\kappa) \mathrm{d}v \quad \text{with } x := (\frac{v}{t+1})^\kappa \frac{\kappa}{c} \\
&\leqslant \frac{2^\zeta}{c} \int_0^\infty \kappa v^{\kappa-1} \exp(-v^\kappa) \mathrm{d}v \\
&= \left(\frac{2^\zeta}{c} \wedge t\right) ,
\end{aligned}
\tag{25}
$$

where we change the integral variable $v^\kappa := c\frac{(t+1)^{1-\zeta}-(u+1)^{1-\zeta}}{1-\zeta}$ with $\kappa := 1-\zeta$ such that

$$
\mathrm{d}u = -\frac{\kappa^{1/\kappa}}{c^{1/\kappa}} (\frac{u+1}{t+1})^{1-\kappa} \left[1 - (\frac{u+1}{t+1})^\kappa\right]^{1-\frac{1}{\kappa}} \mathrm{d}v = -\frac{\kappa}{c}\left[1 - (\frac{v}{t+1})^\kappa \frac{\kappa}{c}\right]^{\frac{1-\kappa}{\kappa}} \left(\frac{v}{t+1}\right)^{\kappa-1} \mathrm{d}v ,
$$

with $(\frac{u+1}{t+1})^\kappa = 1 - (v/(t+1))^\kappa \kappa / c$ and the upper limit of integral is $C := c^{1/\kappa}[(t+1)^\kappa - (u+1)^\kappa]^{1/\kappa}$. Due to $u = (t+1)(1-x)^{\frac{1}{\kappa}} - 1 \in [1, t]$, we have $(1-x)^{\frac{1}{\kappa}} \in [2/(t+1), 1]$ and accordingly

$$
g(x) := [(t+1)(1-x)^{\frac{1}{\kappa}} - 1]^{\kappa-1}(1-x)^{\frac{1-\kappa}{\kappa}} \leqslant 2^{1-\kappa}(t+1)^{\kappa-1} \quad \text{with } x \in \left[0, 1 - \left(\frac{2}{t+1}\right)^\kappa\right] ,
$$

as an increasing function of x.

Similar to Eq. 25, we have the following estimation

$$
\int_1^t \gamma_0^2 u^{-2\zeta} \exp\left(-2\widetilde{\lambda}_i \gamma_0 \frac{(t+1)^{1-\zeta}-(u+1)^{1-\zeta}}{1-\zeta}\right) \mathrm{d}u \lesssim \left(\frac{\gamma_0}{\widetilde{\lambda}_i} \wedge \gamma_0^2 t\right) .
\tag{26}
$$

E PROOFS FOR Bias

In this section, we present the error bound for Bias. By virtue of Minkowski inequality, we have

$$
\begin{aligned}
\left(\mathbb{E}_{\boldsymbol{X},\boldsymbol{W}}\left[\langle \bar{\eta}_n^{\texttt{bias}}, \Sigma_m \bar{\eta}_n^{\texttt{bias}} \rangle\right]\right)^{\frac{1}{2}} &\leqslant \Big(\underbrace{\mathbb{E}_{\boldsymbol{X},\boldsymbol{W}}\left[\langle \bar{\eta}_n^{\texttt{bias}} - \bar{\eta}_n^{\texttt{bX}}, \Sigma_m(\bar{\eta}_n^{\texttt{bias}} - \bar{\eta}_n^{\texttt{bX}}) \rangle\right]}_{\triangleq \texttt{B1}} \Big)^{\frac{1}{2}} + \left(\mathbb{E}_{\boldsymbol{W}}\left[\langle \bar{\eta}_n^{\texttt{bX}}, \Sigma_m \bar{\eta}_n^{\texttt{bX}} \rangle\right]\right)^{\frac{1}{2}} \\
&\leqslant (\texttt{B1})^{\frac{1}{2}} + \Big(\underbrace{\mathbb{E}_{\boldsymbol{W}}\left[\langle \bar{\eta}_n^{\texttt{bX}} - \bar{\eta}_n^{\texttt{bXW}}, \Sigma_m(\bar{\eta}_n^{\texttt{bX}} - \bar{\eta}_n^{\texttt{bXW}}) \rangle\right]}_{\triangleq \texttt{B2}} \Big)^{\frac{1}{2}} + [\underbrace{\langle \bar{\eta}_n^{\texttt{bXW}}, \widetilde{\Sigma}_m \bar{\eta}_n^{\texttt{bXW}} \rangle}_{\triangleq \texttt{B3}}]^{\frac{1}{2}} .
\end{aligned}
\tag{27}
$$

In the next, we give the error bounds for B3, B2, and B1, respectively.

E.1 Bound for B3

In this section, we aim to bound $\texttt{B3} := \langle \bar{\eta}_n^{\texttt{bXW}}, \widetilde{\Sigma}_m \bar{\eta}_n^{\texttt{bXW}} \rangle$.

Proposition 2. *Under Assumption 1, 2, 3, if the step-size $\gamma_t := \gamma_0 t^{-\zeta}$ with $\zeta \in [0,1)$ satisfies $\gamma_0 \leqslant \frac{1}{\operatorname{Tr}(\widetilde{\Sigma}_m)}$, then* B3 *can be bounded by*

$$\texttt{B3} \lesssim \frac{n^{\zeta-1}}{\gamma_0} \|f^*\|^2 \,.$$

Proof. Due to $\gamma_0 \leqslant \frac{1}{\operatorname{Tr}(\widetilde{\Sigma}_m)}$, the operator $I - \gamma_t \widetilde{\Sigma}_m$ is a contraction map for $t = 1, 2, \ldots, n$. Take spectral decomposition $\widetilde{\Sigma}_m = \widetilde{U}\widetilde{\Lambda}\widetilde{U}^\top$ where $\widetilde{U}$ is an orthogonal matrix and $\widetilde{\Lambda}$ is a diagonal matrix with $(\widetilde{\Lambda})_{11} = \widetilde{\lambda}_1$ and $(\widetilde{\Lambda})_{ii} = \widetilde{\lambda}_2$ $(i = 2, 3, \ldots, m)$ as $\widetilde{\Sigma}_m$ has only two distinct eigenvalues in Lemma 1. Accordingly, we have

$$\begin{aligned}
\langle \bar{\eta}_n^{\texttt{bXW}}, \widetilde{\Sigma}_m \bar{\eta}_n^{\texttt{bXW}} \rangle &= \frac{1}{n^2} \left\langle \sum_{t=0}^{n-1} \prod_{i=1}^{t} (I - \gamma_i \widetilde{\Sigma}_m) f^*, \widetilde{\Sigma}_m \sum_{t=0}^{n-1} \prod_{i=1}^{t} (I - \gamma_i \widetilde{\Sigma}_m) f^* \right\rangle \\
&= \frac{1}{n^2} \left\| \sum_{t=0}^{n-1} \prod_{i=1}^{t} (I - \gamma_i \widetilde{\Sigma}_m) \widetilde{\Sigma}_m^{\frac{1}{2}} f^* \right\|^2 \\
&\leqslant \frac{1}{n^2} \left\| \sum_{t=0}^{n-1} \prod_{i=1}^{t} (I - \gamma_i \widetilde{\Lambda})^t \widetilde{\Lambda}^{\frac{1}{2}} \right\|_2^2 \|f^*\|^2 \quad [\text{using } \widetilde{\Sigma}_m = \widetilde{U}\widetilde{\Lambda}\widetilde{U}^\top] \\
&\leqslant \frac{1}{n} \max_{k=1,2} \sum_{t=0}^{n-1} \prod_{i=1}^{t} (1 - \gamma_i \widetilde{\lambda}_k)^2 \widetilde{\lambda}_k \|f^*\|^2 \\
&\leqslant \frac{1}{n} \sum_{t=0}^{n-1} \prod_{i=1}^{t} (1 - \gamma_i \widetilde{\lambda}_1)^2 \widetilde{\lambda}_1 \|f^*\|^2 + \frac{1}{n} \sum_{t=0}^{n-1} \prod_{i=1}^{t} (1 - \gamma_i \widetilde{\lambda}_2)^2 \widetilde{\lambda}_2 \|f^*\|^2 \,.
\end{aligned} \tag{28}$$

Note that

$$\begin{aligned}
\sum_{t=0}^{n-1} \prod_{i=1}^{t} (1 - \gamma_i \widetilde{\lambda}_j)^2 &\leqslant \sum_{t=0}^{n-1} \exp\left(-2\gamma_0 \widetilde{\lambda}_j \sum_{i=1}^{t} i^{-\zeta} \right) \leqslant \sum_{t=0}^{n-1} \exp\left(-2\gamma_0 \widetilde{\lambda}_j \int_1^{t+1} \frac{1}{x^\zeta} \mathrm{d}x \right) \\
&= \sum_{t=0}^{n-1} \exp\left(-2\gamma_0 \widetilde{\lambda}_j \frac{(t+1)^{1-\zeta} - 1}{1-\zeta} \right) \\
&\leqslant 1 + \int_0^n \exp\left(-2\gamma_0 \widetilde{\lambda}_j \frac{(t+1)^{1-\zeta} - 1}{1-\zeta} \right) \mathrm{d}x \\
&\leqslant 1 + \left(\frac{n^\zeta}{2\gamma_0 \widetilde{\lambda}_j} \wedge n \right), \quad [\text{using Eq. 23}]
\end{aligned} \tag{29}$$

here according to Lemma 1, for $\widetilde{\lambda}_1$, the upper bound $\frac{n^\zeta}{2\gamma_0 \widetilde{\lambda}_1}$ is tighter than n due to $\widetilde{\lambda}_1 \sim \mathcal{O}(1)$; while this conclusion might not hold for $\widetilde{\lambda}_2$ due to $\widetilde{\lambda}_2 \sim \mathcal{O}(1/m)$. Then, taking Eq. 29 back to Eq. 28, we have

$$\begin{aligned}
\langle \bar{\eta}_n^{\texttt{bXW}}, \widetilde{\Sigma}_m \bar{\eta}_n^{\texttt{bXW}} \rangle &\lesssim \frac{n^{\zeta-1}}{\gamma_0} \|f^*\|^2 + \frac{\widetilde{\lambda}_2}{n} \left(\frac{n^\zeta}{\gamma_0 \widetilde{\lambda}_2} \wedge n \right) \|f^*\|^2 \\
&\lesssim \frac{n^{\zeta-1}}{\gamma_0} \|f^*\|^2 \sim \mathcal{O}(n^{\zeta-1}) \,,
\end{aligned} \tag{30}$$

which concludes the proof. □

E.2 Bound for B2

Here we aim to bound $\texttt{B2} := \mathbb{E}_{\boldsymbol{W}} \left[\langle \bar{\eta}_n^{\texttt{bX}} - \bar{\eta}_n^{\texttt{bXW}}, \Sigma_m (\bar{\eta}_n^{\texttt{bX}} - \bar{\eta}_n^{\texttt{bXW}}) \rangle \right] := \mathbb{E}_{\boldsymbol{W}} \left[\langle \bar{\alpha}_n^{\texttt{W}}, \Sigma_m \bar{\alpha}_n^{\texttt{W}} \rangle \right]$

$$\alpha_t^{\texttt{W}} := \eta_t^{\texttt{bX}} - \eta_t^{\texttt{bXW}} = (I - \gamma_t \Sigma_m)(\eta_{t-1}^{\texttt{bX}} - \eta_{t-1}^{\texttt{bXW}}) + \gamma_t (\widetilde{\Sigma}_m - \Sigma_m) \eta_{t-1}^{\texttt{bXW}} \,, \tag{31}$$

with $\alpha_0^{\mathtt{W}} = 0$. Accordingly, $\alpha_t^{\mathtt{W}}$ under the adaptive step-size setting can be formulated as

$$\alpha_t^{\mathtt{W}} = \sum_{k=1}^{t} \gamma_k \prod_{j=k+1}^{t} (I - \gamma_j \Sigma_m)(\widetilde{\Sigma}_m - \Sigma_m) \prod_{s=1}^{k-1} (I - \gamma_s \widetilde{\Sigma}_m) f^* \,, \tag{32}$$

where we use the recursion

$$A_t := (I - \gamma_t \Sigma_m) A_{t-1} + B_t = \sum_{s=1}^{t} \prod_{i=s+1}^{t} (I - \gamma_i \Sigma_m) B_s \,.$$

Note that $\mathbb{E}_{\boldsymbol{W}}[\alpha_t^{\mathtt{W}} | \alpha_{t-1}^{\mathtt{W}}] = (I - \gamma_t \widetilde{\Sigma}_m) \alpha_{t-1}^{\mathtt{W}}$, following (Zou et al., 2021, Lemma B.3), we can rewrite B2 as a double-sum formulation

$$\begin{aligned}
\mathtt{B2} &:= \mathbb{E}_{\boldsymbol{W}} \left[\langle \bar{\alpha}_n^{\mathtt{W}}, \Sigma_m \bar{\alpha}_n^{\mathtt{W}} \rangle \right] = \mathbb{E}_{\boldsymbol{W}} \left[\langle \Sigma_m, \bar{\alpha}_n^{\mathtt{W}} \otimes \bar{\alpha}_n^{\mathtt{W}} \rangle \right] = \operatorname{Tr} \left(\mathbb{E}_{\boldsymbol{W}} \left[\Sigma_m \bar{\alpha}_n^{\mathtt{W}} \otimes \bar{\alpha}_n^{\mathtt{W}} \right] \right) \\
&\leqslant \operatorname{Tr} \left(\widetilde{\Sigma}_m \mathbb{E}_{\boldsymbol{W}} \left[\bar{\alpha}_n^{\mathtt{W}} \otimes \bar{\alpha}_n^{\mathtt{W}} \right] \right) = \mathbb{E}_{\boldsymbol{W}} \left[\langle \widetilde{\Sigma}_m, \bar{\alpha}_n^{\mathtt{W}} \otimes \bar{\alpha}_n^{\mathtt{W}} \rangle \right] \\
&= \frac{1}{n^2} \mathbb{E}_{\boldsymbol{W}} \left(\left\langle \widetilde{\Sigma}_m, \sum_{0 \leqslant k \leqslant t \leqslant n-1} \mathbb{E}_{\boldsymbol{W}} [\alpha_t^{\mathtt{W}} \otimes \alpha_k^{\mathtt{W}}] + \sum_{0 \leqslant k < t \leqslant n-1} \mathbb{E}_{\boldsymbol{W}} [\alpha_t^{\mathtt{W}} \otimes \alpha_k^{\mathtt{W}}] \right\rangle \right) \\
&\leqslant \frac{1}{n^2} \mathbb{E}_{\boldsymbol{W}} \left(\left\langle \widetilde{\Sigma}_m, \sum_{0 \leqslant k \leqslant t \leqslant n-1} \mathbb{E}_{\boldsymbol{W}} [\alpha_t^{\mathtt{W}} \otimes \alpha_k^{\mathtt{W}}] + \sum_{0 \leqslant k \leqslant t \leqslant n-1} \mathbb{E}_{\boldsymbol{W}} [\alpha_t^{\mathtt{W}} \otimes \alpha_k^{\mathtt{W}}] \right\rangle \right) \\
&= \frac{2}{n^2} \sum_{t=0}^{n-1} \sum_{k=t}^{n-1} \mathbb{E}_{\boldsymbol{W}} \left\langle \prod_{j=t}^{k-1} (I - \gamma_j \widetilde{\Sigma}_m) \widetilde{\Sigma}_m, \underbrace{\mathbb{E}_{\boldsymbol{W}} [\alpha_t^{\mathtt{W}} \otimes \alpha_t^{\mathtt{W}}]}_{:= C_t^{\mathtt{bW}}} \right\rangle ,
\end{aligned} \tag{33}$$

and thus we have the following error bound for B2.

Proposition 3. *Under Assumption 1, 2, 3, if the step-size $\gamma_t := \gamma_0 t^{-\zeta}$ with $\zeta \in [0, 1)$ satisfies*

$$\gamma_0 \leqslant \min \left\{ \frac{1}{\operatorname{Tr}(\Sigma_m)}, \frac{1}{\operatorname{Tr}(\widetilde{\Sigma}_m)} \right\} ,$$

then B2 *can be bounded by*

$$\mathtt{B2} \lesssim \gamma_0 n^{\zeta - 1} \| f^* \|^2 \,.$$

To bound B2, we first show the error bound for $\|\alpha_t^{\mathtt{W}}\|_2$ for $\operatorname{Tr}[C_t^{\mathtt{bW}}] = \|\alpha_t^{\mathtt{W}}\|_2^2$ by the following lemma.

Lemma 6. *Based on the definition of $\alpha_t^{\mathtt{W}}$ in Eq. 32, under Assumption 1, 2, 3,, if the step-size $\gamma_t := \gamma_0 t^{-\zeta}$ with $\zeta \in [0, 1)$ satisfies*

$$\gamma_0 \leqslant \min \left\{ \frac{1}{\operatorname{Tr}(\Sigma_m)}, \frac{1}{\operatorname{Tr}(\widetilde{\Sigma}_m)} \right\} , \tag{34}$$

we have

$$\|\alpha_t^{\mathtt{W}}\|_2 \lesssim \gamma_0 \|\Sigma_m\|_2 \|f^*\| \,.$$

Proof. According to Eq. 32, we have

$$\begin{aligned}
\|\alpha_t^{\mathtt{W}}\|_2 &\leqslant \left\| \sum_{k=1}^{t} \gamma_k \prod_{j=k+1}^{t} (I - \gamma_j \Sigma_m) \widetilde{\Sigma}_m \prod_{s=1}^{k-1} (I - \gamma_s \widetilde{\Sigma}_m) f^* \right\| + \left\| \sum_{k=1}^{t} \gamma_k \prod_{j=k+1}^{t} (I - \gamma_j \Sigma_m) \Sigma_m \prod_{s=1}^{k-1} (I - \gamma_s \widetilde{\Sigma}_m) f^* \right\| \\
&\leqslant \left\| \sum_{k=1}^{t} \gamma_k \prod_{s=1}^{k-1} (I - \gamma_s \widetilde{\Sigma}_m) \widetilde{\Sigma}_m \right\| \|f^*\| + \left\| \sum_{k=1}^{t} \gamma_k \prod_{j=k+1}^{t} (I - \gamma_j \Sigma_m) \Sigma_m \right\| \|f^*\| \\
&\leqslant \sum_{k=1}^{t} \gamma_k \left\| \prod_{s=1}^{k-1} (I - \gamma_s \widetilde{\Sigma}_m) \widetilde{\Sigma}_m \right\|_2 \|f^*\| + \sum_{k=1}^{t} \gamma_k \left\| \prod_{j=k+1}^{t} (I - \gamma_j \Sigma_m) \Sigma_m \right\|_2 \|f^*\| ,
\end{aligned} \tag{35}$$

where $I-\gamma_i\Sigma_m$ and $I-\gamma_i\widetilde{\Sigma}_m$ are contraction maps for $i=1,2,\dots,n$ under our condition in Eq. 34.

For the first term $\sum_{k=1}^t \gamma_k \left\|\prod_{s=1}^{k-1}(I-\gamma_s\widetilde{\Sigma}_m)\widetilde{\Sigma}_m\right\|_2 = \max_{i=1,2}\left(\sum_{k=1}^t \gamma_k \prod_{j=1}^{k-1}(1-\gamma_j\widetilde{\lambda}_i)\widetilde{\lambda}_i\right)$, we have

$$\begin{aligned}
I_i &:= \sum_{k=1}^t \gamma_k \prod_{j=1}^{k-1}(1-\gamma_j\widetilde{\lambda}_i)\widetilde{\lambda}_i \leqslant \widetilde{\lambda}_i \sum_{k=1}^t \gamma_k \exp\Big(-\sum_{j=1}^{k-1}\gamma_j\widetilde{\lambda}_i\Big) \\
&\leqslant \widetilde{\lambda}_i\gamma_0 \int_1^{t+1} u^{-\zeta}\exp\Big(-\widetilde{\lambda}_i\gamma_0\frac{u^{1-\zeta}-1}{1-\zeta}\Big)\mathrm{d}u \\
&\leqslant \int_0^\infty \exp(-x)\mathrm{d}x = 1\,. \quad \text{[using Eq. 22]}
\end{aligned}$$

Similarly, for the second term $\sum_{k=1}^t \gamma_k\left\| \prod_{j=k+1}^t (I-\gamma_j\Sigma_m)\Sigma_m\right\|_2$, we have

$$\begin{aligned}
I_i &:= \sum_{k=1}^t \gamma_k \prod_{j=k+1}^{t}(1-\gamma_j\lambda_i)\lambda_i \leqslant \lambda_i \sum_{k=1}^t \gamma_k \exp\Big(-\sum_{j=k+1}^{t}\gamma_j\lambda_i\Big) \\
&\leqslant \lambda_i \sum_{k=1}^t \gamma_0 k^{-\zeta}\exp\Big(-\lambda_i\gamma_0\frac{(t+1)^{1-\zeta}-(k+1)^{1-\zeta}}{1-\zeta}\Big) \\
&\leqslant \lambda_i\gamma_0 t^{-\zeta} + \lambda_i\gamma_0\int_1^t u^{-\zeta}\exp\Big(-\lambda_i\gamma_0\frac{(t+1)^{1-\zeta}-(u+1)^{1-\zeta}}{1-\zeta}\Big)\mathrm{d}u\,.
\end{aligned} \tag{36}$$

Due to $\|\Sigma_m\|_2 \sim \mathcal{O}(1)$ in Lemma 2, we have

$$\begin{aligned}
\max_{i\in\{1,2,\dots,m\}} I_i &\leqslant \lambda_1\gamma_0 t^{-\zeta} + 2^{1-\kappa}\int_0^\infty \kappa v^{\kappa-1}\exp(-v^\kappa)\mathrm{d}v \quad \text{[using Eq. 25]} \\
&\leqslant \gamma_0\|\Sigma_m\|_2 + 2^{1-\kappa}\,.
\end{aligned}$$

Accordingly, we have

$$\begin{aligned}
\|\alpha_t^{\mathtt{W}}\|_2 &\leqslant \|f^*\| + \sum_{k=1}^t \gamma_k \left\|\prod_{j=k+1}^t (I-\gamma_j\Sigma_m)\Sigma_m\right\|_2 \|f^*\| = \|f^*\| + \max_{i=1,2,\dots m}\left(\sum_{k=1}^t\gamma_k \prod_{j=k+1}^t (1-\gamma_j\lambda_i)\lambda_i\right)\|f^*\| \\
&\lesssim \gamma_0\|\Sigma_m\|_2\|f^*\|\,. \quad \text{[using } \|\Sigma_m\|_2 \sim \mathcal{O}(1) \text{ by Lemma 2]}
\end{aligned}$$

□

Proof of Proposition 3. Based on the above results, B2 can be bounded by

$$
\begin{aligned}
\texttt{B2} &:= \mathbb{E}_{\boldsymbol{W}}\left[\langle \bar{\alpha}_n^{\texttt{W}}, \Sigma_m \bar{\alpha}_n^{\texttt{W}} \rangle\right] \leqslant \operatorname{Tr}\left(\mathbb{E}_{\boldsymbol{W}}\left[\Sigma_m \bar{\alpha}_n^{\texttt{W}} \otimes \bar{\alpha}_n^{\texttt{W}}\right]\right) \\
&\leqslant \frac{2}{n^2} \sum_{t=0}^{n-1} \sum_{k=t}^{n-1} \mathbb{E}_{\boldsymbol{W}} \left\langle \prod_{j=t}^{k-1} (I - \gamma_j \widetilde{\Sigma}_m) \widetilde{\Sigma}_m, \underbrace{\mathbb{E}_{\boldsymbol{W}}[\alpha_t^{\texttt{W}} \otimes \alpha_t^{\texttt{W}}]}_{:=C_t^{\texttt{bW}}} \right\rangle \quad \text{[using Eq. 33]} \\
&\lesssim \frac{\gamma_0^2}{n^2} \mathbb{E}_{\boldsymbol{W}}[\|\Sigma_m\|_2^2] \|f^*\|^2 \sum_{t=0}^{n-1} \sum_{k=t}^{n-1} \left\| \prod_{j=t}^{k-1} (I - \gamma_j \widetilde{\Sigma}_m) \widetilde{\Sigma}_m \right\|_2 \quad \text{[using Lemma 6]} \\
&\lesssim \frac{\gamma_0^2}{n^2} \|f^*\|^2 \sum_{t=0}^{n-1} \sum_{k=t}^{n-1} \max_{i \in \{1,2,\ldots,m\}} \lambda_i \exp\left(-\lambda_i \sum_{j=t}^{k-1} \gamma_j\right) \quad \text{[using Lemma 2]} \\
&\leqslant \frac{\gamma_0^2}{n^2} \|f^*\|^2 \max_{i \in \{1,2,\ldots,m\}} \sum_{t=0}^{n-1} \sum_{k=t}^{n-1} \lambda_i \exp\left(-\lambda_i \gamma_0 \frac{k^{1-\zeta} - t^{1-\zeta}}{1-\zeta}\right) \quad \text{[using } \textstyle\sum_{j=t}^{k-1} \gamma_j \leqslant \gamma_0 \int_t^k \frac{1}{x^\zeta} \mathrm{d}x] \\
&\leqslant \frac{\gamma_0^2}{n^2} \|f^*\|^2 \max_{i \in \{1,2,\ldots,m\}} \sum_{t=0}^{n-1} \lambda_i \int_t^n \exp\left(-\lambda_i \gamma_0 \frac{u^{1-\zeta} - t^{1-\zeta}}{1-\zeta}\right) \mathrm{d}u \\
&\leqslant \frac{\gamma_0^2}{n^2} \|f^*\|^2 \max_{i \in \{1,2,\ldots,m\}} \sum_{t=0}^{n-1} \lambda_i \left[\frac{n^\zeta}{\lambda_i \gamma_0} \wedge (n-t)\right] \quad \text{[using Eq. 24]} \\
&\leqslant \gamma_0 n^{\zeta-1} \|f^*\|^2 \,,
\end{aligned}
$$

where in the last inequality we choose $\frac{n^\zeta}{\lambda_i \gamma_0}$ instead of $n-t$ for a tight error bound. Finally, we conclude our proof. □

E.3 BOUND FOR B1

Here we aim to bound $\texttt{B1} := \mathbb{E}_{\boldsymbol{X},\boldsymbol{W}}\left[\langle \bar{\eta}_n^{\texttt{bias}} - \bar{\eta}_n^{\texttt{bX}}, \Sigma_m(\bar{\eta}_n^{\texttt{bias}} - \bar{\eta}_n^{\texttt{bX}}) \rangle\right]$. Define $\alpha_t^{\texttt{X}} := \eta_t^{\texttt{bias}} - \eta_t^{\texttt{bX}}$, we have

$$
\alpha_t^{\texttt{X}} = [I - \gamma_t \varphi(\boldsymbol{x}_t) \otimes \varphi(\boldsymbol{x}_t)] \alpha_{t-1}^{\texttt{X}} + \gamma_t [\Sigma_m - \varphi(\boldsymbol{x}_t) \otimes \varphi(\boldsymbol{x}_t)] \eta_{t-1}^{\texttt{bX}} \,, \tag{37}
$$

with $\alpha_0^{\texttt{X}} = 0$ and $\eta_{t-1}^{\texttt{bX}} = \prod_{j=1}^{t-1}(I - \gamma_j \Sigma_m) f^*$. Accordingly, we have

$$
\texttt{B1} := \mathbb{E}_{\boldsymbol{X},\boldsymbol{W}}\left[\langle \bar{\eta}_n^{\texttt{bias}} - \bar{\eta}_n^{\texttt{bX}}, \Sigma_m(\bar{\eta}_n^{\texttt{bias}} - \bar{\eta}_n^{\texttt{bX}}) \rangle\right] = \mathbb{E}_{\boldsymbol{W}}\left(\mathbb{E}_{\boldsymbol{X}}[\langle \bar{\alpha}_n^{\texttt{X}}, \Sigma_m \bar{\alpha}_n^{\texttt{X}} \rangle]\right) \,.
$$

Proposition 4. *Under Assumption 1, 2, 3, 4 with $r' \geqslant 1$, if the step-size $\gamma_t := \gamma_0 t^{-\zeta}$ with $\zeta \in [0,1)$ satisfies*

$$
\gamma_0 < \min\left\{\frac{1}{r' \operatorname{Tr}(\Sigma_m)}, \frac{1}{2\operatorname{Tr}(\Sigma_m)}\right\} \,,
$$

then B1 *can be bounded by*

$$
\texttt{B1} \lesssim \frac{\gamma_0 r' n^{\zeta-1}}{\sqrt{\mathbb{E}[1 - \gamma_0 r' \operatorname{Tr}(\Sigma_m)]^4}} \|f^*\|^2 \sim \mathcal{O}\left(n^{\zeta-1}\right) \,.
$$

To prove Proposition 4, we need a lemma on stochastic recursions based on $\mathbb{E}[\alpha_t^{\texttt{X}} | \alpha_{t-1}^{\texttt{X}}] = (I - \gamma_t \Sigma_m) \alpha_{t-1}^{\texttt{X}}$, that shares the similar proof fashion with (Bach & Moulines, 2013, Lemma 1) and (Dieuleveut & Bach, 2016, Lemma 11).

Lemma 7. *Under Assumption 1, 2, 3, 4 with $r' \geqslant 1$, denoting $H_{t-1} := [\Sigma_m - \varphi(\boldsymbol{x}_t) \otimes \varphi(\boldsymbol{x}_t)] \eta_{t-1}^{\texttt{bX}}$, if the step-size $\gamma_t := \gamma_0 t^{-\zeta}$ with $\zeta \in [0,1)$ satisfies*

$$
\gamma_0 < \frac{1}{r' \operatorname{Tr}(\Sigma_m)} \,,
$$

we have

$$
\mathbb{E}_{\boldsymbol{X}}[\langle \bar{\alpha}_n^{\texttt{X}}, \Sigma_m \bar{\alpha}_n^{\texttt{X}} \rangle] \leqslant \frac{1}{2n[1 - \gamma_0 r' \operatorname{Tr}(\Sigma_m)]} \left(\sum_{k=1}^{n-1} \mathbb{E}\|\alpha_k^{\texttt{X}}\|^2 \left(\frac{1}{\gamma_{k+1}} - \frac{1}{\gamma_k}\right) + 2\sum_{t=0}^{n-1} \gamma_{t+1} \mathbb{E}_{\boldsymbol{X}} \|H_t\|^2\right) \,.
$$

Remark: We require $\|\Sigma_m\|_2 \neq \frac{1}{r'\gamma_0}$ to avoid the denominator to be zero, which naturally holds as the probability measure of the continuous random variable $\|\Sigma_m\|_2$ at a point is zero.

Proof. According to the definition of $\alpha_t^{\mathtt{X}}$ in Eq. 37, define $H_{t-1} := [\Sigma_m - \varphi(\boldsymbol{x}_t) \otimes \varphi(\boldsymbol{x}_t)]\eta_{t-1}^{\mathtt{bX}}$, we have

$$\begin{aligned}\|\alpha_t^{\mathtt{X}}\|^2 &= \|\alpha_{t-1}^{\mathtt{X}} - \gamma_t([\varphi(\boldsymbol{x}_t) \otimes \varphi(\boldsymbol{x}_t)]\alpha_{t-1}^{\mathtt{W}} - H_{t-1})\|^2 \\ &= \|\alpha_{t-1}^{\mathtt{X}}\|^2 + \gamma_t^2 \|H_{t-1} - [\varphi(\boldsymbol{x}_t) \otimes \varphi(\boldsymbol{x}_t)]\alpha_{t-1}^{\mathtt{X}}\|^2 + 2\gamma_t \langle \alpha_{t-1}^{\mathtt{W}}, H_{t-1} - [\varphi(\boldsymbol{x}_t) \otimes \varphi(\boldsymbol{x}_t)]\alpha_{t-1}^{\mathtt{X}} \rangle \\ &\leqslant \|\alpha_{t-1}^{\mathtt{X}}\|^2 + 2\gamma_t^2 \left(\|H_{t-1}\|^2 + \|[\varphi(\boldsymbol{x}_t) \otimes \varphi(\boldsymbol{x}_t)]\alpha_{t-1}^{\mathtt{X}}\|^2\right) + 2\gamma_t \langle \alpha_{t-1}^{\mathtt{X}}, H_{t-1} - [\varphi(\boldsymbol{x}_t) \otimes \varphi(\boldsymbol{x}_t)]\alpha_{t-1}^{\mathtt{X}} \rangle \,,\end{aligned}$$

which implies (by taking the conditional expectation)

$$\begin{aligned}\mathbb{E}_{\boldsymbol{X}}[\|\alpha_t^{\mathtt{W}}\|^2 | \alpha_{t-1}^{\mathtt{W}}] &\leqslant \|\alpha_{t-1}^{\mathtt{X}}\|^2 + 2\gamma_t^2 \|H_{t-1}\|^2 + 2\gamma_t^2 \langle \alpha_{t-1}^{\mathtt{X}}, \mathbb{E}_{\boldsymbol{X}}[\varphi(\boldsymbol{x}_t) \otimes \varphi(\boldsymbol{x}_t) \otimes \varphi(\boldsymbol{x}_t) \otimes \varphi(\boldsymbol{x}_t)]\alpha_{t-1}^{\mathtt{X}} \rangle \\ &\quad - 2\gamma_t \langle \alpha_{t-1}^{\mathtt{X}}, \Sigma_m \alpha_{t-1}^{\mathtt{X}} \rangle \\ &\leqslant \|\alpha_{t-1}^{\mathtt{X}}\|^2 + 2\gamma_t^2 \|H_{t-1}\|^2 + 2\gamma_t^2 r' \mathrm{Tr}(\Sigma_m) \langle \alpha_{t-1}^{\mathtt{X}}, \Sigma_m \alpha_{t-1}^{\mathtt{X}} \rangle - 2\gamma_t \langle \alpha_{t-1}^{\mathtt{X}}, \Sigma_m \alpha_{t-1}^{\mathtt{X}} \rangle \\ &= \|\alpha_{t-1}^{\mathtt{X}}\|^2 + 2\gamma_t^2 \|H_{t-1}\|^2 - 2\gamma_t [1 - \gamma_t r' \mathrm{Tr}(\Sigma_m)] \langle \alpha_{t-1}^{\mathtt{X}}, \Sigma_m \alpha_{t-1}^{\mathtt{X}} \rangle \,.\end{aligned} \tag{38}$$

where the first inequality holds by $\mathbb{E}_{\boldsymbol{X}}[H_{t-1}] = 0$, and the second inequality satisfies by Assumption 4.

By taking the expectation of Eq. 38, we have

$$\mathbb{E}_{\boldsymbol{X}}[\|\alpha_t^{\mathtt{X}}\|^2] \leqslant \mathbb{E}_{\boldsymbol{X}}[\|\alpha_{t-1}^{\mathtt{X}}\|^2] + 2\gamma_t^2 \mathbb{E}_{\boldsymbol{X}}[\|H_{t-1}\|^2] - 2\gamma_t [1 - \gamma_t r' \mathrm{Tr}(\Sigma_m)] \mathbb{E}_{\boldsymbol{X}} \langle \alpha_{t-1}^{\mathtt{X}}, \Sigma_m \alpha_{t-1}^{\mathtt{X}} \rangle \,,$$

which indicates that

$$\begin{aligned}\mathbb{E}_{\boldsymbol{X}}\left[\langle \bar{\alpha}_n^{\mathtt{X}}, \Sigma_m \bar{\alpha}_n^{\mathtt{X}} \rangle]\right\rangle &\leqslant \frac{1}{n} \sum_{t=0}^{n-1} \mathbb{E}_{\boldsymbol{X}} \langle \alpha_t^{\mathtt{W}}, \Sigma_m \alpha_t^{\mathtt{W}} \rangle \leqslant \frac{1}{2n[1 - \gamma_0 r' \mathrm{Tr}(\Sigma_m)]} \left(\sum_{k=1}^{n-1} \mathbb{E}_{\boldsymbol{X}} \|\alpha_k^{\mathtt{X}}\|^2 (\frac{1}{\gamma_{k+1}} - \frac{1}{\gamma_k}) \right. \\ &\quad \left. + \frac{1}{2\gamma_1} \mathbb{E}_{\boldsymbol{X}} \|\alpha_0^{\mathtt{X}}\|^2 - \frac{1}{2\gamma_t} \mathbb{E}_{\boldsymbol{X}} \|\alpha_t^{\mathtt{X}}\|^2 + \sum_{t=0}^{n-1} \gamma_{t+1} \mathbb{E}_{\boldsymbol{X}} \|H_t\|^2 \right) \\ &\leqslant \frac{1}{2n[1 - \gamma_0 r' \mathrm{Tr}(\Sigma_m)]} \left(\sum_{k=1}^{n-1} \mathbb{E}_{\boldsymbol{X}} \|\alpha_k^{\mathtt{X}}\|^2 (\frac{1}{\gamma_{k+1}} - \frac{1}{\gamma_k}) + 2 \sum_{t=0}^{n-1} \gamma_{t+1} \mathbb{E}_{\boldsymbol{X}} \|H_t\|^2 \right) \,,\end{aligned}$$

due to $\alpha_0^{\mathtt{W}} = 0$. □

In the next, we present the error bounds for two respective terms in Lemma 7.

Lemma 8. *Based on the definition of $\alpha_t^{\mathtt{X}}$ in Eq. 39, under Assumption 1, 2, 3, 4 with $r' \geqslant 1$, if the step-size $\gamma_t := \gamma_0 t^{-\zeta}$ with $\zeta \in [0, 1)$ satisfies*

$$\gamma_0 < \min \left\{ \frac{1}{r' \mathrm{Tr}(\Sigma_m)}, \frac{1}{2\mathrm{Tr}(\Sigma_m)} \right\} \,,$$

we have

$$\sum_{k=1}^{n-1} \mathbb{E} \|\alpha_k^{\mathtt{X}}\|^2 (\frac{1}{\gamma_{k+1}} - \frac{1}{\gamma_k}) \lesssim \frac{\gamma_0 r' \mathrm{Tr}(\Sigma_m)}{1 - \gamma_0 r' \mathrm{Tr}(\Sigma_m)} (n^\zeta - 1) \|f^*\|^2 \,.$$

Proof. Based on the definition of $\alpha_t^{\mathtt{X}}$ in Eq. 37, it can be reformulated as

$$\begin{aligned}\alpha_t^{\mathtt{X}} &= [I - \gamma_t \varphi(\boldsymbol{x}_t) \otimes \varphi(\boldsymbol{x}_t)]\alpha_{t-1}^{\mathtt{X}} + \gamma_t [\Sigma_m - \varphi(\boldsymbol{x}_t) \otimes \varphi(\boldsymbol{x}_t)] \prod_{s=1}^{k-1} (I - \gamma_s \Sigma_m) f^* \\ &= \sum_{k=1}^{t} \gamma_k \prod_{j=k+1}^{t} [I - \gamma_j \varphi(\boldsymbol{x}_t) \otimes \varphi(\boldsymbol{x}_t)][\Sigma_m - \varphi(\boldsymbol{x}_t) \otimes \varphi(\boldsymbol{x}_t)] \prod_{s=1}^{k-1} (I - \gamma_s \Sigma_m) f^* \,.\end{aligned} \tag{39}$$

and accordingly

$$
\begin{aligned}
C_t^{\mathtt{b-X}} &:= \mathbb{E}_{\boldsymbol{X}}[\alpha_t^{\mathtt{X}} \otimes \alpha_t^{\mathtt{X}}] = (I - \gamma_t T^{\mathtt{W}}) \circ C_{t-1}^{\mathtt{b-X}} + \gamma_t^2 (S^{\mathtt{W}} - \widetilde{S}^{\mathtt{W}}) \circ [\eta_{t-1}^{\mathtt{bX}} \otimes \eta_{t-1}^{\mathtt{bX}}] \\
&\preccurlyeq (I - \gamma_t T^{\mathtt{W}}) \circ C_{t-1}^{\mathtt{b-X}} + \gamma_t^2 S^{\mathtt{W}} \circ [\eta_{t-1}^{\mathtt{bX}} \otimes \eta_{t-1}^{\mathtt{bX}}] \\
&\preccurlyeq (I - \gamma_t T^{\mathtt{W}}) \circ C_{t-1}^{\mathtt{b-X}} + \gamma_t^2 r' \mathrm{Tr}\left[\prod_{s=1}^{t-1} (I - \gamma_s \Sigma_m)^2 \Sigma_m\right] \Sigma_m (f^* \otimes f^*) \quad \text{[using Assumption 4]} \\
&\preccurlyeq (I - \gamma_t T^{\mathtt{W}}) \circ C_{t-1}^{\mathtt{b-X}} + \gamma_t^2 r' \mathrm{Tr}(\Sigma_m) \Sigma_m (f^* \otimes f^*) \quad \text{[following Eq. 36: } \exp(-2\lambda_i \gamma_0 \tfrac{t^{1-\zeta}-1}{1-\zeta}) \leqslant 1] \qquad (40) \\
&= r' \mathrm{Tr}(\Sigma_m) \sum_{s=1}^{t} \prod_{i=s+1}^{t} \left(I - \gamma_i T^{\mathtt{W}}\right) \circ \gamma_s^2 \Sigma_m (f^* \otimes f^*) \\
&\preccurlyeq \frac{\gamma_0 r' \mathrm{Tr}(\Sigma_m)}{1 - \gamma_0 r' \mathrm{Tr}(\Sigma_m)} (f^* \otimes f^*) \,. \quad \text{[using Lemma 5]}
\end{aligned}
$$

Accordingly, we have

$$
\begin{aligned}
\sum_{t=1}^{n-1} \mathbb{E}_{\boldsymbol{X}} \|\alpha_t^{\mathtt{X}}\|^2 \left(\frac{1}{\gamma_{t+1}} - \frac{1}{\gamma_t}\right) &= \sum_{t=1}^{n-1} \|C_t^{\mathtt{b-X}}\|_2 \left(\frac{1}{\gamma_{t+1}} - \frac{1}{\gamma_t}\right) \quad \text{[using Eq. 40]} \\
&\leqslant \sum_{t=1}^{n-1} \frac{\gamma_0 r' \mathrm{Tr}(\Sigma_m)}{1 - \gamma_0 r' \mathrm{Tr}(\Sigma_m)} [(t+1)^{\zeta} - t^{\zeta}] \|f^*\|^2 \\
&\lesssim \frac{\gamma_0 r' \mathrm{Tr}(\Sigma_m)}{1 - \gamma_0 r' \mathrm{Tr}(\Sigma_m)} (n^{\zeta} - 1) \|f^*\|^2 \,,
\end{aligned}
$$

which concludes the proof. □

Lemma 9. *Denote $H_{t-1} := [\Sigma_m - \varphi(\boldsymbol{x}_t) \otimes \varphi(\boldsymbol{x}_t)] \eta_{t-1}^{\mathtt{bX}}$, Assumption 1, 2, 3, 4 with $r' \geqslant 1$, if the step-size $\gamma_t := \gamma_0 t^{-\zeta}$ with $\zeta \in [0, 1)$ satisfies*

$$
\gamma_0 \leqslant \frac{1}{\mathrm{Tr}(\Sigma_m)} \,,
$$

we have

$$
\sum_{t=0}^{n-1} \gamma_{t+1} \mathbb{E}_{\boldsymbol{X}} \|H_t\|^2 \leqslant \frac{1}{2} \|f^*\|^2 r' \mathrm{Tr}(\Sigma_m) \,.
$$

Proof.

$$
\begin{aligned}
\sum_{t=0}^{n-1} \gamma_{t+1} \mathbb{E}_{\boldsymbol{X}} \|H_t\|^2 &= \sum_{t=0}^{n-1} \gamma_{t+1} \left\langle f^*, \prod_{j=1}^{t-1} (I - \gamma_j \Sigma_m) \mathbb{E}_{\boldsymbol{X}} [\Sigma_m - \varphi(\boldsymbol{x}_t) \otimes \varphi(\boldsymbol{x}_t)]^2 \prod_{j=1}^{t-1} (I - \gamma_j \Sigma_m) \right\rangle \\
&\leqslant \sum_{t=0}^{n-1} \gamma_{t+1} \left\langle f^*, r' \mathrm{Tr}(\Sigma_m) \Big[\prod_{j=1}^{t-1} (I - \gamma_j \Sigma_m)\Big]^2 \Sigma_m \right\rangle \quad \text{[using Assumption 4]} \\
&\leqslant \|f^*\|^2 r' \mathrm{Tr}(\Sigma_m) \left\| \sum_{t=0}^{n-1} \gamma_{t+1} \Big[\prod_{j=1}^{t-1} (I - \gamma_j \Sigma_m)\Big]^2 \Sigma_m \right\| \\
&\leqslant \|f^*\|^2 r' \mathrm{Tr}(\Sigma_m) \max_{i \in \{1,2,\ldots,m\}} \sum_{t=0}^{n-1} \gamma_{t+1} \prod_{j=1}^{t-1} (1 - \gamma_j \lambda_i)^2 \lambda_i \\
&\leqslant \|f^*\|^2 r' \mathrm{Tr}(\Sigma_m) \max_{i \in \{1,2,\ldots,m\}} \gamma_0 \lambda_i \int_0^n u^{-\zeta} \exp\left(-2\gamma_0 \lambda_i \frac{u^{1-\zeta} - 1}{1 - \zeta}\right) \mathrm{d}u \\
&\leqslant \frac{1}{2} \|f^*\|^2 r' \mathrm{Tr}(\Sigma_m) \quad \text{[using Eq. 22]}
\end{aligned}
$$

which concludes the proof. □

Based on the above results, we are ready to prove Proposition 4.

Proof. According to Lemma 8, we have

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{W}} \frac{\sum_{k=1}^{n-1} \mathbb{E}\|\alpha_k^{\mathtt{X}}\|^2 (\frac{1}{\gamma_{k+1}} - \frac{1}{\gamma_k})}{2n[1-\gamma_0 r' \mathrm{Tr}(\Sigma_m)]} &\lesssim \mathbb{E}_{\boldsymbol{W}} \frac{\gamma_0 r' \mathrm{Tr}(\Sigma_m)}{2n[1-\gamma_0 r' \mathrm{Tr}(\Sigma_m)]^2} (n^{\zeta} - 1) \|f^*\|^2 \\
&\leqslant \gamma_0 r' n^{\zeta-1} \sqrt{\mathbb{E}[\mathrm{Tr}(\Sigma_m)]^2} \frac{1}{\sqrt{\mathbb{E}[1-\gamma_0 r' \mathrm{Tr}(\Sigma_m)]^4}} \|f^*\|^2 \\
&\lesssim \frac{\gamma_0 r' n^{\zeta-1}}{\sqrt{\mathbb{E}[1-\gamma_0 r' \mathrm{Tr}(\Sigma_m)]^4}} \|f^*\|^2 \\
&\sim \mathcal{O}(n^{\zeta-1}) \,,
\end{aligned}
$$

where the second inequality holds by Cauchy–Schwarz inequality and the last inequality holds by Lemma 2.
According to Lemma 9, we have

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{W}} \frac{2\sum_{t=0}^{n-1} \gamma_{t+1} \mathbb{E}_{\boldsymbol{X}} \|H_t\|^2}{2n[1-\gamma_0 r' \mathrm{Tr}(\Sigma_m)]} &\leqslant \mathbb{E}_{\boldsymbol{W}} \frac{r' \mathrm{Tr}(\Sigma_m)}{2n[1-\gamma_0 r' \mathrm{Tr}(\Sigma_m)]} \|f^*\|^2 \\
&\lesssim \frac{r'}{n} \sqrt{\mathbb{E}[\mathrm{Tr}(\Sigma_m)]^2} \frac{1}{\sqrt{\mathbb{E}[1-\gamma_0 r' \mathrm{Tr}(\Sigma_m)]^2}} \|f^*\|^2 \\
&\lesssim \frac{r'}{n\sqrt{\mathbb{E}[1-\gamma_0 r' \mathrm{Tr}(\Sigma_m)]^2}} \|f^*\|^2 \quad \text{[using Lemma 2]} \\
&\sim \mathcal{O}\left(\frac{1}{n}\right) \,.
\end{aligned}
$$

Accordingly, combining the above two equations, we have

$$
\begin{aligned}
\mathtt{B1} := \mathbb{E}_{\boldsymbol{W}} \mathbb{E}_{\boldsymbol{X}} [\langle \bar{\alpha}_n^{\mathtt{X}}, \Sigma_m \bar{\alpha}_n^{\mathtt{X}} \rangle] &\leqslant \frac{1}{2n[1-\gamma_0 r' \mathrm{Tr}(\Sigma_m)]} \mathbb{E}_{\boldsymbol{W}} \left(\sum_{k=1}^{n-1} \mathbb{E}\|\alpha_k^{\mathtt{X}}\|^2 (\frac{1}{\gamma_{k+1}} - \frac{1}{\gamma_k}) + 2\sum_{t=0}^{n-1} \gamma_{t+1} \mathbb{E}_{\boldsymbol{X}} \|H_t\|^2 \right) \\
&\lesssim \frac{\gamma_0 r' n^{\zeta-1}}{\sqrt{\mathbb{E}[1-\gamma_0 r' \mathrm{Tr}(\Sigma_m)]^4}} \|f^*\|^2 \,,
\end{aligned}
$$

which concludes the proof. □

E.4 Proof of Theorem 1

Proof. Combining the above results for three terms $\mathtt{B1}$, $\mathtt{B2}$, $\mathtt{B3}$, the $\mathtt{Bias}$ can be upper bounded by

$$
\begin{aligned}
\mathtt{Bias} &\leqslant \left(\sqrt{\mathtt{B1}} + \sqrt{\mathtt{B2}} + \sqrt{\mathtt{B3}}\right)^2 \leqslant \sqrt{3}(\mathtt{B1} + \mathtt{B2} + \mathtt{B3}) \\
&\lesssim \frac{\gamma_0 r' n^{\zeta-1}}{\sqrt{\mathbb{E}[1-\gamma_0 r' \mathrm{Tr}(\Sigma_m)]^4}} \|f^*\|^2 \,.
\end{aligned}
$$

□

F Proof for Variance

In this section, we present the error bound for $\mathtt{Variance}$. Recall the definition of $\eta_t^{\mathtt{vX}}$ in Eq. 12 and $\eta_t^{\mathtt{vXW}}$ in Eq. 13, and

$$
\bar{\eta}_n^{\mathtt{vX}} := \frac{1}{n} \sum_{t=0}^{n-1} \bar{\eta}_t^{\mathtt{vX}}, \qquad \bar{\eta}_n^{\mathtt{vXW}} := \frac{1}{n} \sum_{t=0}^{n-1} \bar{\eta}_t^{\mathtt{vXW}} \,,
$$

by virtue of Minkowski inequality, `Variance` can be further decomposed as

$$\Big(\mathbb{E}_{\boldsymbol{X},\boldsymbol{W},\boldsymbol{\varepsilon}}\big[\langle\bar{\eta}_n^{\mathtt{var}},\Sigma_m\bar{\eta}_n^{\mathtt{var}}\rangle\big]\Big)^{\frac{1}{2}}\leqslant\Big(\underbrace{\mathbb{E}_{\boldsymbol{X},\boldsymbol{W},\boldsymbol{\varepsilon}}\big[\langle\bar{\eta}_n^{\mathtt{var}}-\bar{\eta}_n^{\mathtt{vX}},\Sigma_m(\bar{\eta}_n^{\mathtt{var}}-\bar{\eta}_n^{\mathtt{vX}})\rangle\big]}_{\triangleq\mathtt{V1}}\Big)^{\frac{1}{2}}+\Big(\mathbb{E}_{\boldsymbol{X},\boldsymbol{W},\boldsymbol{\varepsilon}}\big[\langle\bar{\eta}_n^{\mathtt{vX}},\Sigma_m\bar{\eta}_n^{\mathtt{vX}}\rangle\big]\Big)^{\frac{1}{2}}$$
$$\leqslant(\mathtt{V1})^{\frac{1}{2}}+\Big(\underbrace{\mathbb{E}_{\boldsymbol{X},\boldsymbol{W},\boldsymbol{\varepsilon}}\big[\langle\bar{\eta}_n^{\mathtt{vX}}-\bar{\eta}_n^{\mathtt{vXW}},\Sigma_m(\bar{\eta}_n^{\mathtt{vX}}-\bar{\eta}_n^{\mathtt{vXW}})\rangle\big]}_{\triangleq\mathtt{V2}}\Big)^{\frac{1}{2}}+\big[\underbrace{\mathbb{E}_{\boldsymbol{X},\boldsymbol{W},\boldsymbol{\varepsilon}}\langle\bar{\eta}_n^{\mathtt{vXW}},\Sigma_m\bar{\eta}_n^{\mathtt{vXW}}\rangle}_{\triangleq\mathtt{V3}}\big]^{\frac{1}{2}}\,. \tag{41}$$

F.1 Bound for V3

In this section, we aim to bound $\mathtt{V3}:=\mathbb{E}_{\boldsymbol{X},\boldsymbol{W},\boldsymbol{\varepsilon}}\langle\bar{\eta}_n^{\mathtt{vXW}},\Sigma_m\bar{\eta}_n^{\mathtt{vXW}}\rangle$. Note that $\mathbb{E}_{\boldsymbol{X},\boldsymbol{\varepsilon}}[\eta_t^{\mathtt{vXW}}|\eta_{t-1}^{\mathtt{vXW}}]=(I-\gamma_t\widetilde{\Sigma}_m)\eta_{t-1}^{\mathtt{vXW}}$, similar to Eq. 33 for B2, we have the following expression for V3

$$\begin{aligned}\mathtt{V3}&:=\mathbb{E}_{\boldsymbol{X},\boldsymbol{W},\boldsymbol{\varepsilon}}\langle\bar{\eta}_n^{\mathtt{vXW}},\Sigma_m\bar{\eta}_n^{\mathtt{vXW}}\rangle=\mathbb{E}_{\boldsymbol{W}}[\mathbb{E}_{\boldsymbol{X},\boldsymbol{\varepsilon}}\langle\Sigma_m,\bar{\eta}_n^{\mathtt{vXW}}\otimes\bar{\eta}_n^{\mathtt{vXW}}\rangle]\\&=\frac{1}{n^2}\mathbb{E}_{\boldsymbol{W}}\left(\left\langle\Sigma_m,\sum_{0\leqslant k\leqslant t\leqslant n-1}\mathbb{E}_{\boldsymbol{X},\boldsymbol{\varepsilon}}[\eta_t^{\mathtt{vXW}}\otimes\eta_k^{\mathtt{vXW}}]+\sum_{0\leqslant k<t\leqslant n-1}\mathbb{E}_{\boldsymbol{X},\boldsymbol{\varepsilon}}[\eta_t^{\mathtt{vXW}}\otimes\eta_k^{\mathtt{vXW}}]\right\rangle\right)\\&\leqslant\frac{1}{n^2}\mathbb{E}_{\boldsymbol{W}}\left(\left\langle\Sigma_m,\sum_{0\leqslant k\leqslant t\leqslant n-1}\mathbb{E}_{\boldsymbol{X},\boldsymbol{\varepsilon}}[\eta_t^{\mathtt{vXW}}\otimes\eta_k^{\mathtt{vXW}}]+\sum_{0\leqslant k\leqslant t\leqslant n-1}\mathbb{E}_{\boldsymbol{X},\boldsymbol{\varepsilon}}[\eta_t^{\mathtt{vXW}}\otimes\eta_k^{\mathtt{vXW}}]\right\rangle\right)\\&=\frac{2}{n^2}\sum_{t=0}^{n-1}\sum_{k=t}^{n-1}\mathbb{E}_{\boldsymbol{W}}\left\langle\prod_{j=t}^{k-1}(I-\gamma_j\widetilde{\Sigma}_m)\Sigma_m,\underbrace{\mathbb{E}_{\boldsymbol{X},\boldsymbol{\varepsilon}}[\eta_t^{\mathtt{vXW}}\otimes\eta_t^{\mathtt{vXW}}]}_{:=C_t^{\mathtt{vXW}}}\right\rangle\,,\end{aligned}\tag{42}$$

and thus we have the following error bound for V3.

Proposition 5. *Under Assumption 2, 3, 5 with $\tau>0$, if the step-size $\gamma_t:=\gamma_0t^{-\zeta}$ with $\zeta\in[0,1)$ satisfies $\gamma_0\leqslant\frac{1}{\mathrm{Tr}(\widetilde{\Sigma}_m)}$, then V3 can be bounded by*

$$\mathtt{V3}\lesssim\begin{cases}\gamma_0\tau^2\dfrac{m}{n^{1-\zeta}}, & \text{if } m\leqslant n\\ \gamma_0\tau^2\left(n^{-1+\zeta}+\dfrac{n}{m}\right), & \text{if } m>n\,.\end{cases}$$

To prove Proposition 5, we need the following lemma.

Lemma 10. *Denote $C_t^{\mathtt{vXW}}:=\mathbb{E}_{\boldsymbol{X},\boldsymbol{\varepsilon}}[\eta_t^{\mathtt{vXW}}\otimes\eta_t^{\mathtt{vXW}}]$, under Assumptions 2, 3, 5 with $\tau>0$, if $\gamma_0\leqslant 1/\mathrm{Tr}(\widetilde{\Sigma}_m)$, we have*

$$C_t^{\mathtt{vXW}}\preccurlyeq\tau^2\sum_{k=1}^{t}\gamma_k^2\prod_{j=k+1}^{t}[I-\gamma_j\widetilde{\Sigma}_m]\Sigma_m\,.$$

Proof. Recall the definition of $\eta_t^{\mathtt{vXW}}$ in Eq. 13, it can be further represented as

$$\eta_t^{\mathtt{vXW}}=(I-\gamma_t\widetilde{\Sigma}_m)\eta_{t-1}^{\mathtt{vXW}}+\gamma_t\varepsilon_k\varphi(\boldsymbol{x}_k)=\sum_{k=1}^{t}\prod_{j=k+1}^{t}(I-\gamma_j\widetilde{\Sigma}_m)\gamma_k\varepsilon_k\varphi(\boldsymbol{x}_k)\quad\text{with } \eta_0^{\mathtt{vXW}}=0\,.$$

Accordingly, $C_t^{\mathtt{vXW}}$ admits (with $C_0^{\mathtt{vXW}}=0$)

$$C_t^{\mathtt{vXW}}=\sum_{k=1}^{t}\prod_{j=k+1}^{t}(I-\gamma_j\widetilde{\Sigma}_m)^2\gamma_k^2\Xi\preccurlyeq\tau^2\sum_{k=1}^{t}\gamma_k^2\prod_{j=k+1}^{t}(I-\gamma_j\widetilde{\Sigma}_m)^2\Sigma_m\quad[\text{using Assumption 5}]$$

where we use $\mathbb{E}[\varepsilon_i\varepsilon_j]=0$ for $i\neq j$. □

In the next, we are ready to bound V3 in Proposition 5.

Proof of Proposition 5. Note that $\widetilde{\lambda}_1 \sim \mathcal{O}(1)$ and $\widetilde{\lambda}_2 \sim \mathcal{O}(1/m)$ in Lemma 2, we take the upper bound of the integral in Eq. 24 to $\frac{n^\zeta}{\widetilde{\lambda}_1\gamma_0}$ for $\widetilde{\lambda}_1$. However, according to the order of $\widetilde{\lambda}_2$, if $\widetilde{\lambda}_2 \lesssim 1/n$, the exact upper bound is tight. Based on this, we first consider that $m \leqslant n$ case such that $\widetilde{\lambda}_2 \gtrsim 1/n$, and then focus on the $m \geqslant n$ case. Taking $\frac{n^\zeta}{\widetilde{\lambda}_i\gamma_0}$ in Eq. 24 and $\frac{\gamma_0}{\widetilde{\lambda}_i}$ in Eq. 26, we have

$$
\begin{aligned}
\texttt{V3} &:= \mathbb{E}_{\boldsymbol{X},\boldsymbol{W},\boldsymbol{\varepsilon}}\langle \bar{\eta}_n^{\texttt{vXW}}, \Sigma_m \bar{\eta}_n^{\texttt{vXW}} \rangle = \mathbb{E}_{\boldsymbol{X},\boldsymbol{W},\boldsymbol{\varepsilon}}\langle \Sigma_m, \bar{\eta}_n^{\texttt{vXW}} \otimes \bar{\eta}_n^{\texttt{vXW}} \rangle \\
&\leqslant \frac{2}{n^2}\sum_{t=0}^{n-1}\sum_{k=t}^{n-1}\mathbb{E}_{\boldsymbol{W}}\left\langle \prod_{j=t}^{k-1}(I-\gamma_j\widetilde{\Sigma}_m)\Sigma_m, \underbrace{\mathbb{E}_{\boldsymbol{X},\boldsymbol{\varepsilon}}[\bar{\eta}_t^{\texttt{vXW}} \otimes \bar{\eta}_t^{\texttt{vXW}}]}_{:=C_t^{\texttt{vXW}}} \right\rangle \quad \text{[using Eq. 42]} \\
&\leqslant \frac{2\tau^2}{n^2}\sum_{t=0}^{n-1}\sum_{k=t}^{n-1}\mathbb{E}_{\boldsymbol{W}}\left\langle \prod_{j=t}^{k-1}(I-\gamma_j\widetilde{\Sigma}_m)\Sigma_m, \sum_{s=1}^{t}\gamma_s^2\prod_{j=s+1}^{t}(I-\gamma_j\widetilde{\Sigma}_m)^2\Sigma_m \right\rangle \quad \text{[using Lemma 10]} \\
&\leqslant \frac{2\tau^2}{n^2}\sum_{t=0}^{n-1}\sum_{k=t}^{n-1}\operatorname{Tr}\left[\prod_{j=t}^{k-1}(I-\gamma_j\widetilde{\Sigma}_m)\widetilde{\Sigma}_m\sum_{s=1}^{t}\gamma_s^2\prod_{j=s+1}^{t}(I-\gamma_j\widetilde{\Sigma}_m)^2\widetilde{\Sigma}_m\right]\left\|\mathbb{E}_{\boldsymbol{W}}[\Sigma_m^2\widetilde{\Sigma}_m^{-2}]\right\|_2 \\
&\lesssim \frac{2\tau^2}{n^2}\sum_{t=0}^{n-1}\sum_{k=t}^{n-1}\sum_{i=1}^{m}\left[\prod_{j=t}^{k-1}(1-\gamma_j\widetilde{\lambda}_i)\widetilde{\lambda}_i\sum_{s=1}^{t}\gamma_s^2\prod_{j=s+1}^{t}(1-\gamma_j\widetilde{\lambda}_i)^2\widetilde{\lambda}_i\right] \quad \text{[using Lemma 3]} \\
&\leqslant \frac{2\tau^2}{n^2}\sum_{t=0}^{n-1}\sum_{k=t}^{n-1}\sum_{i=1}^{m}\left[\widetilde{\lambda}_i^2\exp\left(-\widetilde{\lambda}_i\gamma_0\frac{k^{1-\zeta}-t^{1-\zeta}}{1-\zeta}\right)\sum_{s=1}^{t}\gamma_s^2\exp\left(-2\widetilde{\lambda}_i\gamma_0\frac{(t+1)^{1-\zeta}-(s+1)^{1-\zeta}}{1-\zeta}\right)\right] \\
&\lesssim \frac{\tau^2}{n^2}\sum_{t=0}^{n-1}\sum_{i=1}^{m}\left[\widetilde{\lambda}_i^2\frac{n^\zeta}{\widetilde{\lambda}_i\gamma_0}\left(\frac{\gamma_0}{\widetilde{\lambda}_i}+\gamma_t^2\right)\right] \quad \text{[using Eqs. 24, 26]} \\
&\leqslant \frac{\tau^2}{n^2}\left[n^{1+\zeta}m + n^\zeta\operatorname{Tr}(\widetilde{\Sigma}_m)\gamma_0\int_0^n t^{-2\zeta}\mathrm{d}t\right] \\
&\lesssim \gamma_0\tau^2\frac{m}{n^{1-\zeta}}\,, \quad \text{[using Lemma 2]}
\end{aligned}
\tag{43}
$$

where the last equality holds that $\int_0^n t^{-2\zeta}\mathrm{d}t \leqslant n$ for any $\zeta \in [0,1)$.

If $\widetilde{\lambda}_2 \lesssim 1/n$, that means, $m > n$ in the over-parameterized regime, we have

$$
\begin{aligned}
\texttt{V3} &\lesssim \frac{2\tau^2}{n^2}\sum_{t=0}^{n-1}\left[\widetilde{\lambda}_1^2\frac{n^\zeta}{\widetilde{\lambda}_1\gamma_0}\left(\frac{\gamma_0}{\widetilde{\lambda}_1}+\gamma_t^2\right)+(m-1)\widetilde{\lambda}_2^2(n-t)t\right] \\
&\lesssim \frac{2\tau^2}{n^2}\Big(\gamma_0 n^{1+\zeta}+(m-1)\gamma_0^2\widetilde{\lambda}_2^2\frac{n(n-1)(n+1)}{6}\Big) \quad \text{[since } \lambda_1 \sim \mathcal{O}(1)\text{]} \\
&\lesssim \gamma_0\tau^2\left(n^{-1+\zeta}+\frac{n}{m}\right)\,,
\end{aligned}
$$

which concludes the proof.

□

F.2 Bound for V2

Here we aim to bound V2

$$
\texttt{V2} := \mathbb{E}_{\boldsymbol{X},\boldsymbol{W},\boldsymbol{\varepsilon}}\left[\langle \bar{\eta}_n^{\texttt{vX}} - \bar{\eta}_n^{\texttt{vXW}}, \Sigma_m(\bar{\eta}_n^{\texttt{vX}} - \bar{\eta}_n^{\texttt{vXW}})\rangle\right].
$$

Recall the definition of $\eta_t^{\texttt{vX}}$ and $\eta_t^{\texttt{vXW}}$ in Eqs. 12 and 13, we have

$$
\eta_t^{\texttt{vXW}} = (I-\gamma_t\widetilde{\Sigma}_m)\eta_{t-1}^{\texttt{vXW}} + \gamma_t\varepsilon_k\varphi(\boldsymbol{x}_k) = \sum_{k=1}^{t}\prod_{j=k+1}^{t}(I-\gamma_j\widetilde{\Sigma}_m)\gamma_k\varepsilon_k\varphi(\boldsymbol{x}_k) \quad \text{with } \eta_0^{\texttt{vXW}} = 0\,,
$$

and accordingly, we define

$$
\begin{aligned}
\alpha_t^{\mathtt{vX-W}} &:= \eta_t^{\mathtt{vX}} - \eta_t^{\mathtt{vXW}} = (I - \gamma_t \Sigma_m)\alpha_{t-1}^{\mathtt{vX-W}} + \gamma_t(\widetilde{\Sigma}_m - \Sigma_m)\eta_{t-1}^{\mathtt{vXW}}\,, \quad \text{with } \alpha_0^{\mathtt{vX-W}} = 0 \\
&= \sum_{s=1}^{t} \prod_{i=s+1}^{t} (I - \gamma_i \Sigma_m)\gamma_s(\widetilde{\Sigma}_m - \Sigma_m) \sum_{k=1}^{s-1} \prod_{j=k+1}^{s-1} (I - \gamma_j \widetilde{\Sigma}_m)\gamma_k \varepsilon_k \varphi(\boldsymbol{x}_k)\,.
\end{aligned}
$$

Proposition 6. *Under Assumptions 2, 3, 5 with $\tau > 0$, if the step-size $\gamma_t := \gamma_0 t^{-\zeta}$ with $\zeta \in [0, 1)$ satisfies*

$$
\gamma_0 \leqslant \frac{1}{\mathrm{Tr}(\Sigma_m)}\,, \tag{44}
$$

then $\mathtt{V2}$ *can be bounded by*

$$
\mathtt{V2} \lesssim \begin{cases} \gamma_0 \tau^2 \dfrac{m}{n^{1-\zeta}}, & \text{if } m \leqslant n \\ \gamma_0 \tau^2, & \text{if } m > n\,. \end{cases}
$$

To prove Proposition 6, we need the following lemma.

Lemma 11. *Denote $C_t^{\mathtt{vX-W}} := \mathbb{E}_{\boldsymbol{X},\boldsymbol{\varepsilon}}[\alpha_t^{\mathtt{vX-W}} \otimes \alpha_t^{\mathtt{vX-W}}]$, under Assumptions 2, 3, 5 with $\tau > 0$, if the step-size $\gamma_t := \gamma_0 t^{-\zeta}$ with $\zeta \in [0, 1)$ satisfies*

$$
\gamma_0 \leqslant \min\left\{ \frac{1}{\mathrm{Tr}(\Sigma_m)}, \frac{1}{\mathrm{Tr}(\widetilde{\Sigma}_m)} \right\}\,,
$$

we have

$$
\|C_t^{\mathtt{vX-W}}\|_2 \lesssim \tau^2 \left\| I + \widetilde{\Sigma}_m^{-2}\Sigma_m^2 \right\|_2 \mathrm{Tr}(\widetilde{\Sigma}_m)\gamma_0^2 \left[\mathrm{Tr}(\Sigma_m)\gamma_0 + 1\right]\,.
$$

Proof. According to the definition of $C_t^{\mathtt{vX-W}}$, it admits the following expression

$$
\begin{aligned}
C_t^{\mathtt{vX-W}} &= \sum_{s=1}^{t} \prod_{i=s+1}^{t} (I - \gamma_i \Sigma_m)\gamma_s^2(\widetilde{\Sigma}_m - \Sigma_m) \sum_{k=1}^{s-1} \prod_{j=k+1}^{s-1} (I - \gamma_j \widetilde{\Sigma}_m)^2 \gamma_k^2 \Xi (\widetilde{\Sigma}_m - \Sigma_m)(I - \gamma_i \Sigma_m) \\
&\preccurlyeq \sum_{s=1}^{t} \prod_{i=s+1}^{t} (I - \gamma_i \Sigma_m)\gamma_s^2(\widetilde{\Sigma}_m - \Sigma_m) \sum_{k=1}^{s-1} \prod_{j=k+1}^{s-1} (I - \gamma_j \widetilde{\Sigma}_m)^2 \gamma_k^2 \Xi (\widetilde{\Sigma}_m - \Sigma_m)(I - \gamma_i \Sigma_m) \\
&\preccurlyeq \tau^2 \sum_{s=1}^{t} \prod_{i=s+1}^{t} (I - \gamma_i \Sigma_m)\gamma_s^2(\widetilde{\Sigma}_m - \Sigma_m) \sum_{k=1}^{s-1} \prod_{j=k+1}^{s-1} (I - \gamma_j \widetilde{\Sigma}_m)^2 \gamma_k^2 \Sigma_m (\widetilde{\Sigma}_m - \Sigma_m)(I - \gamma_i \Sigma_m)\,,
\end{aligned}
$$

where the first equality holds by $\mathbb{E}[\varepsilon_i \varepsilon_j] = 0$ for $i \neq j$ and the second inequality holds by Assumption 5.

Accordingly, $\|C_t^{\mathtt{vX-W}}\|_2$ can be upper bounded by

$$
\begin{aligned}
\|C_t^{\mathtt{vX-W}}\|_2 &\leqslant \tau^2 \sum_{s=1}^{t} \gamma_s^2 \left\| \prod_{i=s+1}^{t} (I - \gamma_i \Sigma_m)^2 \Sigma_m (\widetilde{\Sigma}_m - \Sigma_m)^2 \sum_{k=1}^{s-1} \gamma_k^2 \prod_{j=k+1}^{s-1} (I - \gamma_j \widetilde{\Sigma}_m)^2 \right\|_2 \\
&\leqslant \tau^2 \sum_{s=1}^{t} \gamma_s^2 \left\| \prod_{i=s+1}^{t} (I - \gamma_i \Sigma_m)^2 \Sigma_m \right\|_2 \left\| \sum_{k=1}^{s-1} \gamma_k^2 \prod_{j=k+1}^{s-1} (I - \gamma_j \widetilde{\Sigma}_m)^2 \widetilde{\Sigma}_m^2 \right\|_2 \left\| I + \widetilde{\Sigma}_m^{-2}\Sigma_m^2 \right\|_2 \\
&\leqslant \tau^2 \sum_{s=1}^{t} \max_{q \in \{1,2,\dots,m\}} \gamma_s^2 \exp\left(-2\lambda_q \sum_{i=s+1}^{t} \gamma_i \right) \lambda_q \sum_{k=1}^{s-1} \gamma_k^2 \max_{p \in \{1,2\}} \exp\left(-2\widetilde{\lambda}_p \sum_{j=k+1}^{s-1} \gamma_j \right) \widetilde{\lambda}_p^2 \left\| I + \widetilde{\Sigma}_m^{-2}\Sigma_m^2 \right\|_2\,,
\end{aligned}
$$

where the second inequality holds by $\widetilde{\Sigma}_m \succ 0$ and $\Sigma_m \succcurlyeq 0$ such that

$$
(\widetilde{\Sigma}_m - \Sigma_m)^2 = \widetilde{\Sigma}_m^2 (I - \widetilde{\Sigma}_m^{-1}\Sigma_m)^2 \preccurlyeq \widetilde{\Sigma}_m^2 (I + \widetilde{\Sigma}_m^{-2}\Sigma_m^2)\,,
$$

and $\mathrm{Tr}(AB) \leqslant \|A\|_2 \mathrm{Tr}(B)$ for any two PSD operators A and B.

Similar to Eq. 25, we have the following estimation

$$\begin{aligned}\sum_{k=1}^{s-1}\gamma_k^2\prod_{j=k+1}^{s-1}(1-\gamma_j\widetilde{\lambda}_p)^2 &\leqslant \sum_{k=1}^{s-1}\gamma_k^2\exp\left(-2\widetilde{\lambda}_p\sum_{j=k+1}^{s-1}\gamma_j\right)\\ &\leqslant \gamma_{s-1}^2+\gamma_0^2\int_1^{s-1}u^{-2\zeta}\exp\left(-2\widetilde{\lambda}_p\gamma_0\frac{s^{1-\zeta}-(u+1)^{1-\zeta}}{1-\zeta}\right)\mathrm{d}u\\ &\leqslant \gamma_0^2+\left(\frac{\gamma_0}{\widetilde{\lambda}_p}\wedge\gamma_0^2 s\right),\end{aligned}$$

which implies

$$\max_{p=1,2}\ \widetilde{\lambda}_p^2\sum_{k=1}^{s-1}\gamma_k^2\prod_{j=k+1}^{s-1}(1-\gamma_j\widetilde{\lambda}_p)^2\leqslant\gamma_0^2\widetilde{\lambda}_1^2+\gamma_0\widetilde{\lambda}_1\leqslant 2\gamma_0\widetilde{\lambda}_1=2\gamma_0\|\widetilde{\Sigma}_m\|_2\,,\tag{45}$$

where we use our condition on the step-size $\gamma_0\leqslant\frac{1}{\mathrm{Tr}(\widetilde{\Sigma}_m)}$.

Similar to Eq. 25, we have the following estimation

$$\begin{aligned}\sum_{s=1}^{t}\gamma_s^2\exp\left(-2\lambda_q\sum_{i=s+1}^{t}\gamma_i\right) &\leqslant \sum_{s=1}^{t}\gamma_s^2\exp\left(-2\lambda_q\gamma_0\frac{(t+1)^{1-\zeta}-(s+1)^{1-\zeta}}{1-\zeta}\right)\\ &\leqslant \gamma_t^2+\gamma_0^2\int_1^{t}u^{-2\zeta}\exp\left(-2\lambda_q\gamma_0\frac{(t+1)^{1-\zeta}-(u+1)^{1-\zeta}}{1-\zeta}\right)\mathrm{d}u\\ &\leqslant \gamma_0^2+\left(\frac{\gamma_0}{\lambda_q}\wedge\gamma_0^2 t\right),\end{aligned}$$

which implies

$$\max_{q\in\{1,2,\dots,m\}}\sum_{s=1}^{t}\gamma_s^2\lambda_q\exp\left(-2\lambda_q\sum_{i=s+1}^{t}\gamma_i\right)=\gamma_0^2\|\Sigma_m\|_2+\gamma_0\,.\tag{46}$$

Take the above two equations 45 and 46 back to Eq. 45

$$\begin{aligned}\|C_t^{\mathtt{vX-W}}\|_2 &\lesssim \tau^2\left\|I+\widetilde{\Sigma}_m^{-2}\Sigma_m^2\right\|_2\left(\gamma_0^2\|\Sigma_m\|_2+\gamma_0\right)\left(\gamma_0\|\widetilde{\Sigma}_m\|_2\right)\\ &=\tau^2\gamma_0^2\left\|I+\widetilde{\Sigma}_m^{-2}\Sigma_m^2\right\|_2\left(\gamma_0\|\Sigma_m\|_2+1\right)\|\widetilde{\Sigma}_m\|_2\,.\end{aligned}$$

□

Proof of Proposition 6. By virtue of $\mathbb{E}_{\boldsymbol{X},\boldsymbol{\varepsilon}}[\alpha_t^{\mathtt{vX-W}}|\alpha_{t-1}^{\mathtt{vX-W}}]=(I-\gamma_t\Sigma_m)\alpha_{t-1}^{\mathtt{vX-W}}$ and Lemma 11, V2 can be bounded by

$$\begin{aligned}\mathtt{V2}&=\mathbb{E}_{\boldsymbol{X},\boldsymbol{W},\boldsymbol{\varepsilon}}\left[\langle\bar{\eta}_n^{\mathtt{vX}}-\bar{\eta}_n^{\mathtt{vXW}},\Sigma_m(\bar{\eta}_n^{\mathtt{vX}}-\bar{\eta}_n^{\mathtt{vXW}})\rangle\right]=\mathbb{E}_{\boldsymbol{W}}\langle\Sigma_m,\mathbb{E}_{\boldsymbol{X},\boldsymbol{\varepsilon}}[\bar{\alpha}_n^{\mathtt{vX-W}}\otimes\bar{\alpha}_n^{\mathtt{vX-W}}]\rangle\\ &\leqslant\frac{2}{n^2}\sum_{t=0}^{n-1}\sum_{k=t}^{n-1}\mathbb{E}_{\boldsymbol{W}}\left\langle\prod_{j=t}^{k-1}(I-\gamma_j\Sigma_m)\Sigma_m,\underbrace{\mathbb{E}_{\boldsymbol{X},\boldsymbol{\varepsilon}}[\bar{\eta}_t^{\mathtt{vX-W}}\otimes\bar{\eta}_t^{\mathtt{vX-W}}]}_{:=C_t^{\mathtt{vX-W}}}\right\rangle\\ &\lesssim\frac{\tau^2\gamma_0^2}{n^2}\|\widetilde{\Sigma}_m\|_2\mathbb{E}_{\boldsymbol{W}}\left(\left\|I+\widetilde{\Sigma}_m^{-2}\Sigma_m^2\right\|_2[\|\Sigma_m\|_2\gamma_0+1]\,\mathrm{Tr}\left[\sum_{t=0}^{n-1}\sum_{k=t}^{n-1}\prod_{j=t}^{k-1}(I-\gamma_j\Sigma_m)\Sigma_m\right]\right)\\ &\lesssim\frac{\tau^2\gamma_0^2}{n^2}\|\widetilde{\Sigma}_m\|_2\mathbb{E}_{\boldsymbol{W}}\left[\|\Sigma_m\|_2\left\|I+\widetilde{\Sigma}_m^{-2}\Sigma_m^2\right\|_2\sum_{i=1}^{m}\sum_{t=0}^{n-1}\lambda_i\left(\frac{n^\zeta}{\lambda_i\gamma_0}\wedge(n-t)\right)\right].\quad\text{[using Eq. 24]}\end{aligned}$$

In the $m\leqslant n$ case, we choose $n^\zeta/(\lambda_i\gamma_0)$, and thus

$$\begin{aligned}\mathtt{V2}&\lesssim\frac{\tau^2 m\gamma_0^2}{n^2}\|\widetilde{\Sigma}_m\|_2\mathbb{E}_{\boldsymbol{W}}\left[\|\Sigma_m\|_2\left\|I+\widetilde{\Sigma}_m^{-2}\Sigma_m^2\right\|_2\right]\frac{n^{1+\zeta}}{\gamma_0}\\ &\leqslant\tau^2\gamma_0\frac{m\|\widetilde{\Sigma}_m\|_2}{n^{1-\zeta}}\sqrt{\mathbb{E}_{\boldsymbol{W}}\|\Sigma_m\|_2^2}\sqrt{\mathbb{E}_{\boldsymbol{W}}\left\|I+\widetilde{\Sigma}_m^{-2}\Sigma_m^2\right\|_2^2}\quad\text{[using Cauchy–Schwarz inequality]}\\ &\lesssim\tau^2\gamma_0\frac{m}{n^{1-\zeta}}\,.\quad\text{[using Lemma 2 and 3]}\end{aligned}$$

If $m > n$, we have

$$\begin{aligned}
\texttt{V2} &\lesssim \frac{2\tau^2\gamma_0^2}{n^2}\|\widetilde{\Sigma}_m\|_2 \mathbb{E}_{\boldsymbol{W}}\left([\mathrm{Tr}(\Sigma_m)]^2 \left\|I + \widetilde{\Sigma}_m^{-2}\Sigma_m^2\right\|_2\right)\sum_{t=0}^{n-1} t \\
&\leqslant \tau^2\gamma_0\|\widetilde{\Sigma}_m\|_2\sqrt{\mathbb{E}_{\boldsymbol{W}}[\mathrm{Tr}(\Sigma_m)]^2}\sqrt{\mathbb{E}_{\boldsymbol{W}}\left\|I + \widetilde{\Sigma}_m^{-2}\Sigma_m^2\right\|_2^2} \quad \text{[using Cauchy–Schwarz inequality]} \\
&\lesssim \tau^2\gamma_0\,, \quad \text{[using Lemma 2 and 3]}
\end{aligned}$$

which concludes the proof. □

F.3 Bound for V1

Here we aim to bound V1

$$\texttt{V1} := \mathbb{E}_{\boldsymbol{X},\boldsymbol{W},\boldsymbol{\varepsilon}}\left[\langle \bar{\eta}_n^{\texttt{var}} - \bar{\eta}_n^{\texttt{vX}}, \Sigma_m(\bar{\eta}_n^{\texttt{var}} - \bar{\eta}_n^{\texttt{vX}})\rangle\right].$$

Recall the definition of $\eta_t^{\texttt{var}}$ in Eq. 6 and $\eta_t^{\texttt{vX}}$ in Eq. 12, we define

$$\begin{aligned}
\alpha_t^{\texttt{v-X}} &:= \eta_t^{\texttt{var}} - \eta_t^{\texttt{vX}} = [I - \gamma_t\varphi(\boldsymbol{x}_t)\otimes\varphi(\boldsymbol{x}_t)]\alpha_{t-1}^{\texttt{v-X}} + \gamma_t[\Sigma_m - \varphi(\boldsymbol{x}_t)\otimes\varphi(\boldsymbol{x}_t)]\eta_{t-1}^{\texttt{vX}}\,, \quad \text{with } \alpha_0^{\texttt{v-X}} = 0\,. \\
&= [I - \gamma_t\varphi(\boldsymbol{x}_t)\otimes\varphi(\boldsymbol{x}_t)]\alpha_{t-1}^{\texttt{v-X}} + \gamma_t[\Sigma_m - \varphi(\boldsymbol{x}_t)\otimes\varphi(\boldsymbol{x}_t)]\sum_{k=1}^{t-1}\prod_{j=k+1}^{t-1}(I - \gamma_j\Sigma_m)\gamma_k\varepsilon_k\varphi(\boldsymbol{x}_k) \\
&= \sum_{s=1}^{t}\prod_{i=s+1}^{t}\gamma_s[I - \gamma_i\varphi(\boldsymbol{x}_i)\otimes\varphi(\boldsymbol{x}_i)][\Sigma_m - \varphi(\boldsymbol{x}_t)\otimes\varphi(\boldsymbol{x}_t)]\sum_{k=1}^{s-1}\prod_{j=k+1}^{s-1}(I - \gamma_j\Sigma_m)\gamma_k\varepsilon_k\varphi(\boldsymbol{x}_k)\,,
\end{aligned}$$

and thus the error bound for V1 is given by the following proposition.

Proposition 7. *Under Assumption 1, 2, 3, 4 with $r' \geqslant 1$, and Assumption 5 with $\tau > 0$, if the step-size $\gamma_t := \gamma_0 t^{-\zeta}$ with $\zeta \in [0, 1)$ satisfies*

$$\gamma_0 < \min\left\{\frac{1}{r'\mathrm{Tr}(\Sigma_m)}, \frac{1}{2\mathrm{Tr}(\Sigma_m)}\right\},$$

then V1 *can be bounded by*

$$\texttt{V1} \lesssim \frac{\tau^2 r'\gamma_0^2}{\sqrt{\mathbb{E}[1-\gamma_0 r'\mathrm{Tr}(\Sigma_m)]^2}}\begin{cases}\frac{m}{n^{1-\zeta}}, & \text{if } m \leqslant n \\ 1, & \text{if } m > n\,.\end{cases}$$

To prove Proposition 7, we need the following lemma. Define $C_t^{\texttt{v-X}} := \mathbb{E}_{\boldsymbol{X},\boldsymbol{\varepsilon}}[\alpha_t^{\texttt{v-X}}\otimes\alpha_t^{\texttt{v-X}}]$, we have the following lemma that is useful to bound $C_t^{\texttt{v-X}}$.

Lemma 12. *Denote $C_t^{\texttt{v-X}} := \mathbb{E}_{\boldsymbol{X},\boldsymbol{\varepsilon}}[\alpha_t^{\texttt{v-X}}\otimes\alpha_t^{\texttt{v-X}}]$, under Assumptions 1, 2, 3, 4 with $r' \geqslant 1$, and Assumption 5 with $\tau > 0$, if the step-size $\gamma_t := \gamma_0 t^{-\zeta}$ with $\zeta \in [0, 1)$ satisfies*

$$\gamma_0 < \min\left\{\frac{1}{r'\mathrm{Tr}(\Sigma_m)}, \frac{1}{2\mathrm{Tr}(\Sigma_m)}\right\},$$

we have

$$C_t^{\texttt{v-X}} \preccurlyeq \frac{\gamma_0^2 r'\tau^2[\mathrm{Tr}(\Sigma_m) + \gamma_0\mathrm{Tr}(\Sigma_m^2)]}{1-\gamma_0 r'\mathrm{Tr}(\Sigma_m)}I\,.$$

Proof. According to the definition of $C_t^{\texttt{v-X}}$, it admits the following expression

$$
\begin{aligned}
C_t^{\texttt{v-X}} &= \sum_{s=1}^{t}\prod_{i=s+1}^{t}\gamma_s^2\mathbb{E}_{\boldsymbol{x}}[I-\gamma_i\varphi(\boldsymbol{x}_i)\otimes\varphi(\boldsymbol{x}_i)]^2\mathbb{E}_{\boldsymbol{x}}[\Sigma_m-\varphi(\boldsymbol{x}_t)\otimes\varphi(\boldsymbol{x}_t)]^2\sum_{k=1}^{s-1}\prod_{j=k+1}^{s-1}(I-\gamma_j\Sigma_m)^2\gamma_k^2\Xi \\
&= (I-\gamma_t T^{\texttt{W}})\circ C_{t-1}^{\texttt{v-X}} + \gamma_t^2(S^{\texttt{W}}-\widetilde{S}^{\texttt{W}})\circ\sum_{k=1}^{t-1}\prod_{j=k+1}^{t-1}(I-\gamma_j\Sigma_m)^2\gamma_k^2\Xi \quad \text{[using PSD operators]} \\
&\preccurlyeq (I-\gamma_t T^{\texttt{W}})\circ C_{t-1}^{\texttt{v-X}} + \gamma_t^2 S^{\texttt{W}}\circ\sum_{k=1}^{t-1}\prod_{j=k+1}^{t-1}(I-\gamma_j\Sigma_m)^2\gamma_k^2\Xi \quad \text{[using } S^{\texttt{W}}\succcurlyeq\widetilde{S}^{\texttt{W}}\text{]} \\
&\preccurlyeq (I-\gamma_t T^{\texttt{W}})\circ C_{t-1}^{\texttt{v-X}} + \tau^2\gamma_t^2 S^{\texttt{W}}\circ\sum_{k=1}^{t-1}\prod_{j=k+1}^{t-1}(I-\gamma_j\Sigma_m)^2\gamma_k^2\Sigma_m \quad \text{[using Assumption 5]} \\
&\preccurlyeq (I-\gamma_t T^{\texttt{W}})\circ C_{t-1}^{\texttt{v-X}} + \tau^2\gamma_t^2 r'\mathrm{Tr}\left[\sum_{k=1}^{t-1}\prod_{j=k+1}^{t-1}(I-\gamma_j\Sigma_m)^2\gamma_k^2\Sigma_m^2\right]\Sigma_m\,. \quad \text{[using Assumption 4]}
\end{aligned}
\tag{47}
$$

Similar to Eq. 25, we have the following estimation

$$
\begin{aligned}
\mathrm{Tr}\left[\sum_{k=1}^{t-1}\prod_{j=k+1}^{t-1}(I-\gamma_j\Sigma_m)^2\Sigma_m^2\gamma_k^2\right] &= \sum_{i=1}^{m}\lambda_i^2\sum_{k=1}^{t-1}\gamma_k^2\prod_{j=k+1}^{t-1}(1-\gamma_j\lambda_i)^2 \leqslant \sum_{i=1}^{m}\lambda_i^2\sum_{k=1}^{t-1}\gamma_k^2\exp\left(-2\lambda_i\sum_{j=k+1}^{s-1}\gamma_j\right) \\
&\leqslant \gamma_0^2\sum_{i=1}^{m}\lambda_i^2\left[1+\int_1^{t-1}u^{-2\zeta}\exp\left(-2\lambda_i\gamma_0\frac{t^{1-\zeta}-(u+1)^{1-\zeta}}{1-\zeta}\right)\mathrm{d}u\right] \\
&\leqslant \gamma_0^2\mathrm{Tr}(\Sigma_m^2)+\sum_{i=1}^{m}\lambda_i^2\left(\frac{\gamma_0}{\lambda_i}\wedge\gamma_0^2 t\right) \quad \text{[using Eq. 26]} \\
&\leqslant \gamma_0^2\mathrm{Tr}(\Sigma_m^2)+\gamma_0\mathrm{Tr}(\Sigma_m)\,,
\end{aligned}
$$

where we use the error bound $\frac{\gamma_0}{\lambda_i}$ instead of the exact one $\gamma_0^2 t$ for tight estimation.

Taking the above equation back to Eq. 47, we have

$$
\begin{aligned}
C_t^{\texttt{v-X}} &\preccurlyeq (I-\gamma_t T^{\texttt{W}})\circ C_{t-1}^{\texttt{v-X}} + \gamma_t^2\tau^2 r'\gamma_0[\mathrm{Tr}(\Sigma_m)+\gamma_0\mathrm{Tr}(\Sigma_m^2)]\Sigma_m \\
&\preccurlyeq \tau^2 r'\gamma_0[\mathrm{Tr}(\Sigma_m)+\gamma_0\mathrm{Tr}(\Sigma_m^2)]\sum_{s=1}^{t}\prod_{i=s+1}^{t}(I-\gamma_i T^{\texttt{W}})\circ\gamma_s^2\Sigma_m \\
&\preccurlyeq \frac{\gamma_0^2 r'\tau^2[\mathrm{Tr}(\Sigma_m)+\gamma_0\mathrm{Tr}(\Sigma_m^2)]}{1-\gamma_0 r'\mathrm{Tr}(\Sigma_m)}I\,, \quad \text{[using Lemma 5]}
\end{aligned}
$$

which concludes the proof. ☐

Proof of Proposition 7. Accordingly, by virtue of $\mathbb{E}_{\boldsymbol{X},\boldsymbol{\varepsilon}}[\alpha_t^{\texttt{v-X}}|\alpha_{t-1}^{\texttt{v-X}}]=(I-\gamma_t\Sigma_m)\alpha_{t-1}^{\texttt{v-X}}$ and Lemma 12, `V1` can be bounded by

$$
\begin{aligned}
\texttt{V1} &= \mathbb{E}_{\boldsymbol{X},\boldsymbol{W},\boldsymbol{\varepsilon}}\left[\langle\bar{\eta}_n^{\texttt{var}}-\bar{\eta}_n^{\texttt{v-X}},\Sigma_m(\bar{\eta}_n^{\texttt{var}}-\bar{\eta}_n^{\texttt{v-X}})\rangle\right] = \mathbb{E}_{\boldsymbol{W}}\langle\Sigma_m,\mathbb{E}_{\boldsymbol{X},\boldsymbol{\varepsilon}}[\bar{\alpha}_n^{\texttt{v-X}}\otimes\bar{\alpha}_n^{\texttt{v-X}}]\rangle \\
&\leqslant \frac{2}{n^2}\sum_{t=0}^{n-1}\sum_{k=t}^{n-1}\mathbb{E}_{\boldsymbol{W}}\left\langle\prod_{j=t}^{k-1}(I-\gamma_j\Sigma_m)\Sigma_m,\underbrace{\mathbb{E}_{\boldsymbol{X},\boldsymbol{\varepsilon}}[\bar{\eta}_t^{\texttt{v-X}}\otimes\bar{\eta}_t^{\texttt{v-X}}]}_{:=C_t^{\texttt{v-X}}}\right\rangle \\
&\lesssim \frac{\tau^2\gamma_0^2 r'}{n^2}\mathbb{E}_{\boldsymbol{W}}\left[\frac{[\mathrm{Tr}(\Sigma_m)+\gamma_0\mathrm{Tr}(\Sigma_m^2)]}{1-\gamma_0 r'\mathrm{Tr}(\Sigma_m)}\sum_{i=1}^{m}\sum_{t=0}^{n-1}\lambda_i\left(\frac{n^\zeta}{\lambda_i\gamma_0}\wedge(n-t)\right)\right]\,, \quad \text{[using Lemma 12]}
\end{aligned}
$$

where the last inequality follows the integral estimation in Eq. 24.

For $m \leqslant n$, we use $\frac{n^\zeta}{\lambda_i \gamma_0}$, and thus

$$\texttt{V1} \lesssim \frac{\tau^2 \gamma_0 r' m}{n^{1-\zeta}} \mathbb{E}_{\boldsymbol{W}} \left[\frac{[\mathrm{Tr}(\Sigma_m) + \gamma_0 \mathrm{Tr}(\Sigma_m^2)]}{1 - \gamma_0 r' \mathrm{Tr}(\Sigma_m)} \right] \lesssim \frac{\tau^2 r' \gamma_0}{\sqrt{\mathbb{E}[1 - \gamma_0 r' \mathrm{Tr}(\Sigma_m)]^2}} \frac{m}{n^{1-\zeta}} \,,$$

where we use the Cauchy–Schwarz inequality and $\mathrm{Tr}(\Sigma_m)$ as a nonnegative sub-exponential random variable with the sub-exponential norm $\mathcal{O}(1)$ in Lemma 2.

For $m > n$, we use $n - t$, and thus

$$\texttt{V1} \lesssim \tau^2 \gamma_0^2 r' \mathbb{E}_{\boldsymbol{W}} \left[\frac{[\mathrm{Tr}(\Sigma_m) + \gamma_0 \mathrm{Tr}(\Sigma_m^2)]}{1 - \gamma_0 r' \mathrm{Tr}(\Sigma_m)} \right] \lesssim \frac{\tau^2 r' \gamma_0^2}{\sqrt{\mathbb{E}[1 - \gamma_0 r' \mathrm{Tr}(\Sigma_m)]^2}} \sim \mathcal{O}(1) \,.$$

□

F.4 Proof of Theorem 2

Proof. Combining the above results for three terms `V1`, `V2`, `V3`, we can directly obtain the result for `Variance`.

$$\begin{aligned}
\texttt{Variance} &\leqslant \left(\sqrt{\texttt{V1}} + \sqrt{\texttt{V2}} + \sqrt{\texttt{V3}} \right)^2 \leqslant \sqrt{3} (\texttt{V1} + \texttt{V2} + \texttt{V3}) \\
&\lesssim \frac{\gamma_0 r' \tau^2}{\sqrt{\mathbb{E}[1 - \gamma_0 r' \mathrm{Tr}(\Sigma_m)]^2}} \left\{ \begin{array}{l} m n^{\zeta - 1}, \text{ if } m \leqslant n \\ n^{\zeta-1} + \dfrac{n}{m}, \text{ if } m > n \end{array} \right. \\
&\sim \left\{ \begin{array}{l} \mathcal{O}\left(m n^{\zeta-1}\right), \text{ if } m \leqslant n \\ \mathcal{O}\left(n^{\zeta-1} + \dfrac{n}{m}\right), \text{ if } m > n \end{array} \right.
\end{aligned}$$

□