On the Double Descent of Random Features Models Trained with SGD

Fanghui Liu[*]
LIONS, EPFL
fanghui.liu@epfl.ch

Johan A.K. Suykens
ESAT-STADIUS, KU Leuven
johan.suykens@esat.kuleuven.be

Volkan Cevher
LIONS, EPFL
volkan.cevher@epfl.ch

Abstract

We study generalization properties of random features (RF) regression in high dimensions optimized by stochastic gradient descent (SGD) in under-/over-parameterized regime. In this work, we derive precise non-asymptotic error bounds of RF regression under both constant and polynomial-decay 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 for interpolation learning, 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 minimum-norm interpolator, as a theoretical justification of using SGD in practice.

1 Introduction

Over-parameterized models, e.g., linear/kernel regression [1, 2, 3, 4] and neural networks [5, 6, 7], still generalize well even if the labels are pure noise [8]. Such high-capacity models have received significant attention recently as they go against with classical generalization theory. A paradigm for understanding this important phenomenon is *double descent* [9], in which the test error first decreases with increasing number of model parameters in the under-parameterized regime. They large error is yielded until interpolating the data, which is called the interpolation threshold. Finally, the test error decreases again in the over-parameterized regime.

Our work partakes in this research vein and studies the random features (RF) model [10], as a simplified version of neural networks, in the context of double descent phenomenon. 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 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

[*]Most of this work was done when Fanghui was at KU Leuven. Correspondence to: Fanghui Liu <fanghui.liu@epfl.ch>.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

[4, 11]. 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 [4, 12, 11, 13, 14, 15] (see comparisons in Table 1 in Appendix A). The existing theoretical results demonstrate that the excess risk curve exhibits double descent.

Nevertheless, the analysis framework of previous work on RF regression mainly relies on the least-squares closed-form solution, including *minimum-norm* interpolator and ridge regressor. Besides, they often assume the data with specific distribution, e.g., to be Gaussian or uniformly spread on a sphere. Such dependency on the analytic solution and relatively strong data assumption in fact mismatches practical neural networks optimized by stochastic gradient descent (SGD) based algorithms. Our work precisely bridges this gap: We provide a new analysis framework for the generalization properties of RF models trained with SGD and general activation functions, also accommodating adaptive (i.e., polynomial decay) step-size selection, and provide non-asymptotic results in under-/over-parameterized regimes. We make the following contributions and findings:

First, we characterize statistical properties of 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, if the activation function $\sigma(\cdot) : \mathbb{R} \mapsto \mathbb{R}$ is Lipschitz continuous, $\mathrm{Tr}(\Sigma_m)$ 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. This is different from the least squares setting based on effective dimension [2, 16].

Second, based on the bias-variance decomposition in stochastic approximation, we take into account multiple randomness sources of initialization, label noise, and data sampling as well as stochastic gradients. We (partly) disentangle these randomness sources and derive non-asymptotic error bounds under the optimization effect: 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 polynomial-decay step-size SGD setting, and is valid under sub-Gaussian data and general activation functions.

Third, our non-asymptotic results show that, RF regression trained 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 (technical challenges are discussed in Section 4) sheds light on the effect of SGD on high dimensional RF models in under-/over-parameterized regimes, and bridges the gap between the minimum-norm solution and numerical iteration solution in terms of optimization and generalization on double descent. It 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 and introduces our problem setting of RF regression with SGD.

Notation: The notation $\boldsymbol{a} \otimes \boldsymbol{a}$ denotes the tensor product of a vector $\boldsymbol{a}$. 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.

2.1 Related work

A flurry of research papers are devoted to analysis of over-parameterized models on optimization [17, 18, 19], generalization (or their combination) under neural tangent kernel [20, 21, 22] and mean-field analysis regime [23, 24]. 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 [4, 12, 13, 11].

Double descent in random features model: Characterizing the double descent of the RF model often derives from random matrix theory (RMT) in high dimensional statistics [1, 4, 12, 13, 25] and from the replica method [11, 26, 14]. 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 $m/n > 1$. Further, refined results are developed on the *analysis of variance* due to multiple randomness sources [11, 27, 15]. 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 [28].

Non-asymptotic stochastic approximation: Many papers on linear least-squares regression [29, 30], kernel regression [31, 32], random features [33] with SGD often work in the under-parameterized regime, where d is finite and much smaller than n. In the over-parameterized regime, under GD setting, the excess risk of least squares is controlled by the smallest positive eigenvalue in [34] via random matrix theory. Under the averaged constant step-size SGD setting, the excess risk in [35] on least squares in high dimensions can be independent of d, and the convergence rate is built in [16]. This convergence rate is also demonstrated under the minimal-iterate [36] or last-iterate [37] setting in step-size SGD for noiseless least squares. We also notice a concurrent work [38] on last-iterate SGD with decaying step-size on least squares. Besides, the existence of multiple descent [39, 40] beyond double descent and SGD as implicit regularizer [41, 42] can be traced to the above two lines of work. Our work shares some similar technical tools with [31] and [16] but differs from them in several aspects. We detail the differences in Section 4.

2.2 Problem setting

We study the standard problem setting for RF least-squares regression and adopt the relevant terminologies from learning theory: *cf.*, [43, 31, 33, 25] for details. Let $X \subseteq \mathbb{R}^d$ be a metric space and $Y \subseteq \mathbb{R}$. The training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ are 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:

$$\min_{f \in \mathcal{H}} \mathcal{E}(f), \quad \mathcal{E}(f) := \int (f(\boldsymbol{x}) - y)^2 \mathrm{d}\rho(\boldsymbol{x}, y) = \|f - f_\rho\|_{L^2_{\rho_X}}^2, \text{ with } f(\boldsymbol{x}) = \langle \boldsymbol{\theta}, \varphi(\boldsymbol{x}) \rangle \,,$$

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 $\boldsymbol{W} = [\boldsymbol{\omega}_1, \cdots, \boldsymbol{\omega}_m]^\top \in \mathbb{R}^{m \times d}$ 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

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

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

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

actually defined in $\mathcal{H}$ that is isomorphic to $\mathbb{R}^m$. This is the usually (uncentered) covariance matrix in finite dimensions,[2] i.e., $\boldsymbol{\Sigma}_m = \mathbb{E}_{\boldsymbol{x}}[\varphi(\boldsymbol{x}) \otimes \varphi(\boldsymbol{x})]$. Define $J_m : \mathbb{R}^m \to L^2_{\rho_X}$ such that$(J_m \boldsymbol{v})(\cdot) = \langle \boldsymbol{v}, \varphi(\cdot) \rangle, \quad \forall \boldsymbol{v} \in \mathbb{R}^m$, we have $\Sigma_m = J_m^* J_m$, where J_m^* denotes the adjoint operator of J_m.

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

Clearly, Σ_m is random with respect to $\boldsymbol{W}$, and thus its deterministic version is defined as $\widetilde{\Sigma}_m = \mathbb{E}_{\boldsymbol{x},\boldsymbol{W}}[\varphi(\boldsymbol{x}) \otimes \varphi(\boldsymbol{x})]$.

SGD with averaging: Regarding the stochastic approximation, we consider the one pass SGD with iterate averaging and adaptive step-size at each iteration t: after a training sample $(\boldsymbol{x}_t, y_t) \sim \rho$ is observed, we update the decision variable as below (initialized at $\boldsymbol{\theta}_0$)

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

where we use the polynomial decay step size $\gamma_t := \gamma_0 t^{-\zeta}$ with $\zeta \in [0, 1)$, following [31]. This setting also holds for the constant step-size case by taking $\zeta = 0$. Besides, we employ the bath size $= 1$ in an online setting style, which is commonly used in theory [31, 16, 44] for ease of analysis, which captures the key idea of SGD by combining stochastic gradients and data sampling.

The final output is defined as the average of the iterates: $\bar{\boldsymbol{\theta}}_n := \frac{1}{n}\sum_{t=0}^{n-1} \boldsymbol{\theta}_t$. Here we sum up $\{\theta_t\}_{t=0}^{n-1}$ with n terms for notational simplicity. 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 $f^* = \operatorname{argmin}_{f\in\mathcal{H}} \mathcal{E}(f)$ exists (see Assumption 2 in the next section). 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 averaged 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 [31, 45, 33, 25], 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).

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, refer to Appendix B for more discussions.

Assumption 1. *[46, 1, high dimensional setting] We work in the large d, n, m regime 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 sample 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).*

Assumption 2. *There exists $f^* \in \mathcal{H}$ such that $f^* = \operatorname{argmin}_{f\in\mathcal{H}} \mathcal{E}(f)$ with bounded Hilbert norm.*

Remark: This bounded Hilbert norm assumption is commonly used in [47, 40, 48] even though n and d tend to infinity. It holds true for linear functions with $\|f\|_{\mathcal{H}} \leqslant 4\pi$ [49], see Appendix B for details.

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, Logistic, and sine/cosine functions.

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 that follows [29, 16, 37] to analyse SGD for least squares.

Assumption 4 (Fourth moment condition)**.** *Assume there exists some positive constants $r', r \geqslant 1$, such that for any PSD operator A, it holds that*

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

Remark: This assumption requires the data are drawn from some not-too-heavy-tailed distribution, e.g., $\Sigma_m^{-\frac{1}{2}}\boldsymbol{x}$ has sub-Gaussian tail, common in high dimensional statistics. This condition is weaker than most previous work on double descent that requires the data to be Gaussian [1, 11, 27, 12], or uniformly spread on a sphere [4, 50], see comparisons in Table 1 in Appendix A. Note that the

assumption for any PSD operator is just for ease of description. In fact some certain PSD operators satisfying this assumption are enough for our proof. Besides, a special case of this assumption with $A := I$ is proved by Lemma 3, and thus this assumption can be regarded as a natural extension, with more discussions in Appendix B.

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 [31, 16] and holds for the standard noise model $y = f^*(\boldsymbol{x}) + \varepsilon$ with $\mathbb{E}[\varepsilon] = 0$ and $\mathbb{V}[\varepsilon] < \infty$ [1].

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 C), that will be needed for our main result. This is different from the least squares setting [2, 16] that introduces the effective dimension to separate the entire space into a "head" subspace where the error decays more quickly than the complement "tail" subspace. Instead, the following lemma shows that $\widetilde{\Sigma}_m$ has only two distinct eigenvalues at $\mathcal{O}(1)$ and $\mathcal{O}(1/m)$ order, respectively. Such fast eigenvalue decay can avoid extra data spectrum assumption for tight bound. For description simplicity, we consider the single-output activation function: $\sigma(\cdot) : \mathbb{R} \to \mathbb{R}$. Our results can be extended to multiple-output activation functions, see Appendix C.1.2 for details.

Lemma 1. *Under Assumption 1 and 3, the expected covariance operator $\widetilde{\Sigma}_m := \mathbb{E}_{\boldsymbol{x},\boldsymbol{W}}[\varphi(\boldsymbol{x}) \otimes \varphi(\boldsymbol{x})] \in \mathbb{R}^{m\times m}$ has the same diagonal elements and the same non-diagonal element*

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

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

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

Remark: Lemma 1 implies $\mathrm{tr}(\widetilde{\Sigma}_m) < \infty$. 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.

Here we take the ReLU activation $\sigma(x) = \max\{x, 0\}$ as one example, RF actually approximates the first-order arc-cosine kernel [51] with $\varphi(\boldsymbol{x}) \in \mathbb{R}^m$. We have $(\widetilde{\Sigma}_m)_{ii} = \frac{1}{2md}\mathrm{Tr}(\Sigma_d)$ and $(\widetilde{\Sigma}_m)_{ij} = \frac{1}{2md\pi}\mathrm{Tr}(\Sigma_d)$ by recalling $\Sigma_d := \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}\boldsymbol{x}^\top]$ and $\mathrm{Tr}(\Sigma_d)/d \sim \mathcal{O}(1)$. More examples can be found in Appendix C.1.2.

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

Remark: This lemma characterizes the sub-exponential property of covariance operator Σ_m, which is a fundamental result for our proof since the bias and variance involve them.

The following lemma demonstrates that the behavior of the fourth moment can be bounded.

Lemma 3. *Under Assumptions 1,and 3, there exists a constant $r > 0$ such that $\mathbb{E}_{\boldsymbol{W}}\left(\Sigma_m^2\right) \preccurlyeq \mathbb{E}_{\boldsymbol{x},\boldsymbol{W}}[\varphi(\boldsymbol{x}) \otimes \varphi(\boldsymbol{x}) \otimes \varphi(\boldsymbol{x}) \otimes \varphi(\boldsymbol{x})] \preccurlyeq r\mathrm{Tr}(\widetilde{\Sigma}_m)\widetilde{\Sigma}_m$.*

Lemma 4. *Under Assumptions 1 and 3, we have $\mathrm{Tr}[\widetilde{\Sigma}_m^{-1}\mathbb{E}_{\boldsymbol{W}}(\Sigma_m^2)] \sim \mathcal{O}(1)$.*

We remark here that Lemma 3 is a special case of Assumption 4 if we take $A := I$ and $r := 1+\mathcal{O}\left(\frac{1}{m}\right)$; and Lemma 4 is a direct corollary of Lemma 3.

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 [31, 30, 16], it admits

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

where the first term corresponds to the bias

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

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 $\|f\|_{L^2_{\rho_X}}^2 = \langle f, \Sigma_m f\rangle$.

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

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

By (partly) 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 $\gamma_0 \lesssim \frac{1}{r'\mathrm{Tr}(\widetilde{\Sigma}_m)} \sim \mathcal{O}(1)$, the* Bias *in Proposition 1 holds by*

$$\texttt{Bias} \lesssim \gamma_0 r' n^{\zeta-1}\|f^*\|^2 \sim \mathcal{O}\left(n^{\zeta-1}\right)\,.$$

Remark: The error bound for Bias is monotonically decreasing at $\mathcal{O}(n^{\zeta-1})$ rate. For the constant step-size setting, it converges at $\mathcal{O}(1/n)$ rate, which is better than $\mathcal{O}(\sqrt{\log n/n})$ in [25] relying on closed-form solution under correlated features with polynomial decay on Σ_d. Besides, our result on bias matches the exact formulation in [11] under the closed-form solution, i.e., monotonically decreasing bias. One slight difference is, their result on bias tends to a constant under the over-parameterized regime while our bias result can converge to zero.

Theorem 2. *(Error bound for variance) Under Assumptions 1, 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 $\gamma_0 \lesssim \frac{1}{r'\mathrm{Tr}(\widetilde{\Sigma}_m)} \sim \mathcal{O}(1)$, the* Variance *defined in Proposition 1 holds*

$$\texttt{Variance} \lesssim \gamma_0 r'\tau^2 \begin{cases} mn^{\zeta-1}, \text{ if } m \leqslant n \\ 1 + n^{\zeta-1} + \dfrac{n}{m}, \text{ if } m > n \end{cases}$$

Remark: We make the following remarks:
i) The error bound for Variance is demonstrated to be unimodal: increasing with m in the under-parameterized regime and decreasing with m in the over-parameterized regime, and finally converge to a constant order (that depends on noise parameter τ^2), which matches recent results relying on closed-form solution for (refined) variance, e.g., [11, 27, 15].
ii) When compared to least squares, our result can degenerate to this setting by choosing $m := d$. Our upper bound is able to match the lower bound in [1, Corollary 1] with the same order, which demonstrates the tightness of our upper bound. Besides, our results can recover the result of [16] by taking the effective dimension $k^* = \min\{n, d\}$ (no data spectrum assumption is required here). More discussion on our derived results refers to Appendix A.

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.

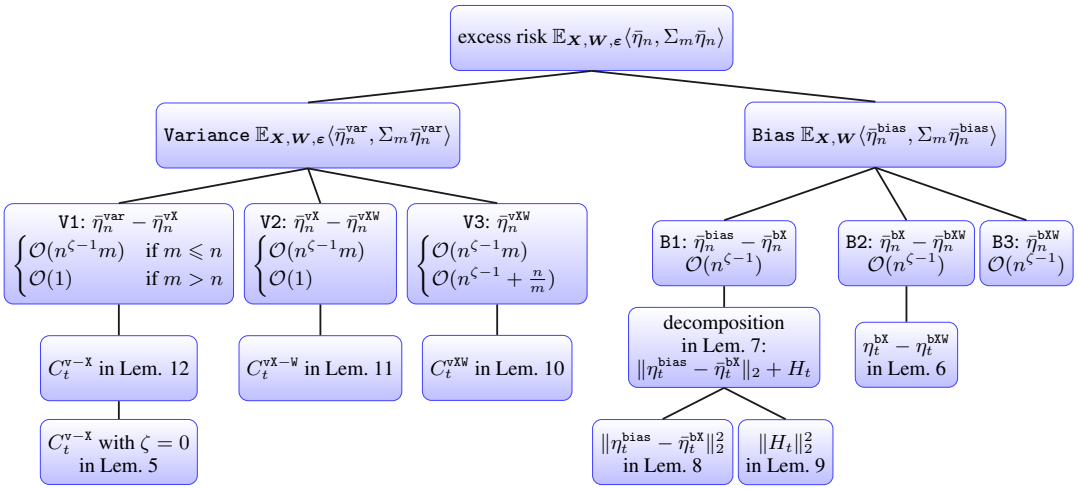


Figure 1: The roadmap of proofs.

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. Recall $\Sigma_m = \mathbb{E}_{\boldsymbol{x}}[\varphi(\boldsymbol{x}) \otimes \varphi(\boldsymbol{x})]$ and $\widetilde{\Sigma}_m = \mathbb{E}_{\boldsymbol{x},\boldsymbol{W}}[\varphi(\boldsymbol{x}) \otimes \varphi(\boldsymbol{x})]$, define

$$\eta_t^{\mathtt{bX}} = (I - \gamma_t\Sigma_m)\eta_{t-1}^{\mathtt{bX}}, \quad \eta_0^{\mathtt{bX}} = f^*, \qquad \eta_t^{\mathtt{bXW}} = (I - \gamma_t\widetilde{\Sigma}_m)\eta_{t-1}^{\mathtt{bXW}}, \quad \eta_0^{\mathtt{bXW}} = f^*, \tag{7}$$

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 Minkowski inequality, the `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 `B3` is a deterministic quantity that is closely connected to model (intrinsic) bias without any randomness; while `B1` and `B2` evaluate the effect of randomness from $\boldsymbol{X}$ and $\boldsymbol{W}$ on the bias, respectively. The error bounds for them can be directly found in Figure 1.

To bound `B3`, we directly focus on its formulation by virtue of spectrum decomposition and integral estimation. To bound `B2`, we have $\mathtt{B2} = \frac{1}{n^2}\mathbb{E}_{\boldsymbol{W}}\Big\|\Sigma_m^{\frac{1}{2}}\sum_{t=0}^{n-1}(\eta_t^{\mathtt{bX}} - \eta_t^{\mathtt{bXW}})\Big\|^2$, where the key part $\eta_t^{\mathtt{bX}} - \eta_t^{\mathtt{bXW}}$ can be estimated by Lemma 6. To bound `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 `Variance`, we need some auxiliary notations.

$$\eta_t^{\mathtt{vX}} := (I - \gamma_t\Sigma_m)\eta_{t-1}^{\mathtt{vX}} + \gamma_t\varepsilon_t\varphi(\boldsymbol{x}_t), \quad \eta_0^{\mathtt{vX}} = 0, \tag{8}$$

$$\eta_t^{\mathtt{vXW}} := (I - \gamma_t\widetilde{\Sigma}_m)\eta_{t-1}^{\mathtt{vXW}} + \gamma_t\varepsilon_t\varphi(\boldsymbol{x}_t), \quad \eta_0^{\mathtt{vXW}} = 0, \tag{9}$$

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

4.2 Discussion on techniques

Our proof framework follows [31] 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 $\boldsymbol{W}$. For instance, their results depend on [29, 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 $\mathtt{B1}$ but expires on other quantities.

Some technical tools used in this paper follow [16] that focuses on linear regression with constant step-size SGD for benign overfitting. However, our results differ from it in 1) tackling multiple randomness, e.g., stochastic gradients, random features (Gaussian initialization), by introducing another type of error decomposition and several deterministic/randomness covariance operators. We prove nice statistical properties of them for proof, which gets rid of data spectrum assumption in [16]. 2) tackling non-constant step-size SGD setting by introducing new integral estimation techniques. Original techniques on constant step-size in [16] are invalid due to non-homogeneous update rules. The above two points make our proof relatively more intractable and largely different. 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.

Here we take the estimation for the variance in [16] under the least squares setting as an example to illustrate this.

$$\text{Variance} \lesssim \sum_{t=0}^{n-1} \Big\langle I - (I - \gamma\Sigma_d)^{n-t}, I - (I - \gamma\Sigma_d)^{t} \Big\rangle \qquad \text{[Eq. (4.10) in [16]]}$$

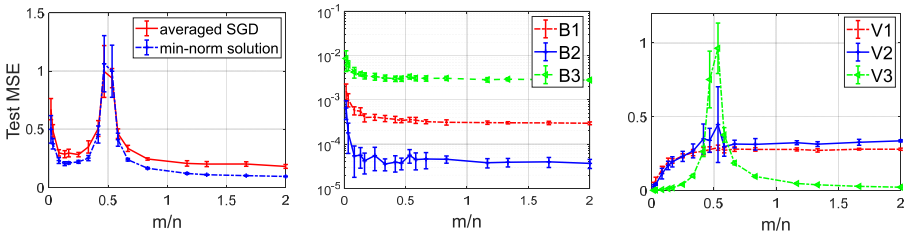
In this setting, the effective dimension to tackle $I - (I - \gamma\Sigma_d)^{n-t}$; while our result is based on fast eigenvalue decay of $\widetilde{\Sigma}_m$ in Lemma 1 can direct to bound this. Besides, the homogeneous markov chain under the constant step-size setting is employed [16] for $(I - \gamma\Sigma_d)^{n-t}$, which is naturally invalid under our decaying step-size setting. Instead, we introduce integral estimation techniques to tackle adaptive step-size, see Appendix E for details.

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. More experiments can be found in Appendix H.

5.1 Behavior of RF for interpolation learning

Here we evaluate the test mean square error (MSE) of RFF regression on the MNIST data set [52], following the experimental setting of [13, 53], to study the generalization performance of minimum-norm solution, see Figure 2(a). More results on regression dataset refer to Appendix H.



(a) SGD vs. min-norm solution (b) `Bias` (c) `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). The interpolation threshold occurs at $m/n = 0.5$ as the Gaussian kernel outputs the $2m$-feature mapping (instead of m), i.e., $\sigma(\boldsymbol{W}\boldsymbol{x}) \in \mathbb{R}^{2m}$. Under this setting, the trends of `Bias` and `Variance` are empirically given in (b) and (c).

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 $n = 600$, $d = 784$, and tuning m 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 1. In our experiment, the initial step-size is set to $\gamma_0 = 1$ and we take the initial point $\boldsymbol{\theta}_0$ near the min-norm solution[3] corrupted with zero-mean, unit-variance Gaussian noise. The experiments are repeated 10 times and the test MSE (mean±std.) can be regarded as a function of the ratio m/n by tuning m. Results on different initialization and more epochs of SGD refer to Appendix H.

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 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.

5.2 Behavior of our error bounds

We have experimentally validated 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(b) and 2(c), respectively. Results of these quantities on different step-size refer to Appendix H.

Experimental settings: Since the target function f^*, the covariance operators Σ_d, Σ_m, and the noise ε are unknown on the MNIST data set, our experimental evaluation need some assumptions to calculate `Bias` and `Variance`. 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, see the remark in Lemma 1, where the expectation on $\boldsymbol{x}$ is approximated by Monte Carlo sampling with n training samples. Accordingly, based on the above results, we are ready to calculate $\eta_t^{\mathtt{bias}}$ in Eq. (5), $\eta_t^{\mathtt{bX}}$, and $\eta_t^{\mathtt{bXW}}$ in Eq. (7), respectively, which is further used to approximately compute $\mathtt{B1} := \mathbb{E}_{\boldsymbol{X},\boldsymbol{W}}\left[\langle \bar{\eta}_n^{\mathtt{bias}} - \bar{\eta}_n^{\mathtt{bX}}, \Sigma_m(\bar{\eta}_n^{\mathtt{bias}} - \bar{\eta}_n^{\mathtt{bX}})\rangle\right]$ (red line) and $\mathtt{B2} := \mathbb{E}_{\boldsymbol{W}}\left[\langle \bar{\eta}_n^{\mathtt{bX}} - \bar{\eta}_n^{\mathtt{bXW}}, \Sigma_m(\bar{\eta}_n^{\mathtt{bX}} - \bar{\eta}_n^{\mathtt{bXW}})\rangle\right]$ (blue line) and $\mathtt{B3} := \langle \bar{\eta}_n^{\mathtt{bXW}}, \widetilde{\Sigma}_m \bar{\eta}_n^{\mathtt{bXW}}\rangle$ (green line). The (approximate) computation for `Variance` can be similar achieved by this process.

Error bounds for bias: Figure 2(b) shows the trends of (scaled) `B1`, `B2`, and `B3`. Recall our error bound: `B1`, `B2`, `B3` $\sim \mathcal{O}(n^{\zeta-1})$, we find that, all of them monotonically decreases at a certain convergence rate when m increases from the under-parameterized regime to the over-parameterized regime. These experimental results coincide with our error bound on them.

[3] In our numerical experiments, we only employ single-pass SGD, and thus the initialization is chosen close to minimum norm solution, with more discussion in Appendix H.

Error bounds for variance: Figure 2(c) shows the trends of (scaled) `V1`, `V2`, and `V3`. Recall our error bound: in the under-parameterized regime, `V1`, `V2`, and `V3` increase 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(c) 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 [11, 27, 15]: the unimodality of variance is a prevalent phenomenon.

6 Conclusion

We present non-asymptotic results for RF regression under the averaged SGD setting for understanding double descent under the optimization effect. 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 polynomial-decay step-size case will slow down the learning rate, but does not change the error bound for variance in over-parameterized regime that converges to $\mathcal{O}(1)$ order, that depends on noise parameter(s).

Our work centers around the RF model, which is still a bit far away from practical neural networks. Theoretical understanding the generalization properties of over-parameterized neural networks is a fundamental but difficult problem. We believe that a comprehensive and thorough understanding of shallow neural networks, e.g., the RF model, is a necessary first step. Besides, we consider the single-pass SGD in our work for simplicity rather than multiple-pass SGD used in practice. This is also an interesting direction for understanding the optimization effect of SGD in the double descent.

Besides, our results obtain the dimension-free bound under both *non-asymptotic* and *asymptotic* regimes. We also need to mention that, our results are also valid under the fixed d setting (which can be larger or smaller than n). This is more practical for real-world applications.

Acknowledgment

The research leading to these results has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation program: ERC Advanced Grant E-DUALITY (787960) and grant agreement n° 725594 - time-data. This paper reflects only the authors' views and the Union is not liable for any use that may be made of the contained information. This work was supported by SNF project – Deep Optimisation of the Swiss National Science Foundation (SNSF) under grant number 200021_205011; Research Council KU Leuven: Optimization frameworks for deep kernel machines C14/18/068; Flemish Government: FWO projects: GOA4917N (Deep Restricted Kernel Machines: Methods and Foundations), PhD/Postdoc grant. This research received funding from the Flemish Government (AI Research Program). This work was supported in part by Ford KU Leuven Research Alliance Project KUL0076 (Stability analysis and performance improvement of deep reinforcement learning algorithms), EU H2020 ICT-48 Network TAILOR (Foundations of Trustworthy AI - Integrating Reasoning, Learning and Optimization), Leuven.AI Institute.
We also thank Zhenyu Liao and Leello Dadi for their helpful discussions on this work.

References

[1] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Annals of Statistics*, 50(2):949–986, 2022. 1, 3, 4, 5, 6, 16, 18, 19

[2] Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *the National Academy of Sciences*, 2020. 1, 2, 5, 17, 18

[3] Denny Wu and Ji Xu. On the optimal weighted ℓ_2 regularization in overparameterized linear regression. In *Advances in Neural Information Processing Systems*, pages 10112–10123, 2020. 1

[4] Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766, 2022. 1, 2, 3, 4, 16, 19, 41

[5] 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. 1

[6] 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. 1

[7] 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*, pages 5137–5147. PMLR, 2020. 1

[8] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021. 1

[9] 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. 1

[10] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, pages 1177–1184, 2007. 1

[11] Stéphane d'Ascoli, Maria Refinetti, Giulio Biroli, and Florent Krzakala. Double trouble in double descent: Bias and variance (s) in the lazy regime. In *International Conference on Machine Learning*, pages 2280–2290, 2020. 2, 3, 4, 6, 10, 16, 17, 19

[12] 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*, pages 1–8, 2020. 2, 3, 4, 16, 17, 19

[13] 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. 2, 3, 8, 16, 17, 40

[14] 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*, pages 3452–3462, 2020. 2, 3, 16

[15] 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. 2, 3, 6, 10, 16, 17

[16] 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. 2, 3, 4, 5, 6, 8, 17, 18, 22, 23, 24

[17] 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*, pages 92–99. IEEE, 2019. 2

[18] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pages 242–252. PMLR, 2019. 2

[19] 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. 2

[20] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, pages 8571–8580, 2018. 2

[21] 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*, pages 322–332, 2019. 2

[22] Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, pages 2933–2943, 2019. 2

[23] 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*, pages 2388–2464. PMLR, 2019. 2

[24] 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*, pages 1305–1338, 2020. 2

[25] Zhu Li, Zhi-Hua Zhou, and Arthur Gretton. Towards an understanding of benign overfitting in neural networks. *arXiv preprint arXiv:2106.03212*, 2021. 3, 4, 6, 16, 20

[26] Jason W Rocks and Pankaj Mehta. Memorizing without overfitting: Bias, variance, and interpolation in over-parameterized models. *arXiv preprint arXiv:2010.13933*, 2020. 3

[27] Ben Adlam and Jeffrey Pennington. Understanding double descent requires a fine-grained bias-variance decomposition. In *Advances in Neural Information Processing Systems*, 2020. 3, 4, 6, 10, 16, 17, 19

[28] Hong Hu and Yue M Lu. Universality laws for high-dimensional learning with random features. *arXiv preprint arXiv:2009.07669*, 2020. 3, 16

[29] 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. 3, 4, 8, 19, 31

[30] 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. 3, 6, 19

[31] Aymeric Dieuleveut and Francis Bach. Nonparametric stochastic approximation with large step-sizes. *Annals of Statistics*, 44(4):1363–1399, 2016. 3, 4, 5, 6, 8, 19, 31

[32] 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. 3

[33] Luigi Carratino, Alessandro Rudi, and Lorenzo Rosasco. Learning with SGD and random features. In *Advances in Neural Information Processing Systems*, pages 10212–10223, 2018. 3, 4, 17

[34] Ilja Kuzborskij, Csaba Szepesvári, Omar Rivasplata, Amal Rannen-Triki, and Razvan Pascanu. On the role of optimization in double descent: A least squares study. In *Advances in Neural Information Processing Systems*, 2021. 3

[35] Xi Chen, Qiang Liu, and Xin T Tong. Dimension independent generalization error by stochastic gradient descent. *arXiv preprint arXiv:2003.11196*, 2020. 3

[36] 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, pages 2576–2586, 2020. 3, 19

[37] Aditya Vardhan Varre, Loucas Pillaud-Vivien, and Nicolas Flammarion. Last iterate convergence of sgd for least-squares in the interpolation regime. In *Advances in Neural Information Processing Systems*, volume 34, pages 21581–21591, 2021. 3, 4, 19

[38] Jingfeng Wu, Difan Zou, Vladimir Braverman, Quanquan Gu, and Sham Kakade. Last iterate risk bounds of sgd with decaying stepsize for overparameterized linear regression. In *International Conference on Machine Learning*, pages 24280–24314. PMLR, 2022. 3, 18

[39] Lin Chen, Yifei Min, Mikhail Belkin, and Amin Karbasi. Multiple descent: Design your own generalization curve. In *Advances in Neural Information Processing Systems*, volume 34, pages 8898–8912, 2021. 3

[40] Tengyuan Liang, Alexander Rakhlin, and Xiyu Zhai. On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels. In *Conference on Learning Theory*, pages 2683–2711, 2020. 3, 4, 18

[41] Behnam Neyshabur, Ryota Tomioka, Ruslan Salakhutdinov, and Nathan Srebro. Geometry of optimization and implicit regularization in deep learning. *arXiv preprint arXiv:1705.03071*, 2017. 3

[42] 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. 3

[43] Felipe Cucker and Dingxuan Zhou. *Learning theory: an approximation theory viewpoint*, volume 24. Cambridge University Press, 2007. 3

[44] Atsushi Nitanda and Taiji Suzuki. Optimal rates for averaged stochastic gradient descent under neural tangent kernel regime. In *International Conference on Learning Representations*, 2020. 4

[45] Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features. In *Advances in Neural Information Processing Systems*, pages 3215–3225, 2017. 4, 17

[46] Noureddine El Karoui. The spectrum of kernel random matrices. *Annals of Statistics*, 38(1):1–50, 2010. 4

[47] Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel "ridgeless" regression can generalize. *Annals of Statistics*, 48(3):1329–1347, 2020. 4, 18

[48] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Generalization error of random feature and kernel methods: hypercontractivity and kernel matrix concentration. *Applied and Computational Harmonic Analysis*, 2021. 4, 18

[49] Francis Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(1):629–681, 2017. 4, 18

[50] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. *Annals of Statistics*, 49(2):1029–1054, 2021. 4, 19

[51] Youngmin Cho and Lawrence K Saul. Kernel methods for deep learning. In *Advances in Neural Information Processing Systems*, pages 342–350, 2009. 5, 18

[52] 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. 8

[53] Michał Dereziński, Feynman Liang, and Michael W Mahoney. Exact expressions for double descent and implicit regularization via surrogate random design. In *Advances in Neural Information Processing Systems*, volume 33, pages 5152–5164, 2020. 8, 40

[54] Oussama Dhifallah and Yue M Lu. A precise performance analysis of learning with random features. *arXiv preprint arXiv:2008.11904*, 2020. 16

[55] Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007. 17

[56] 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*, pages 649–657, 2021. 18

[57] Konstantin Donhauser, Mingqi Wu, and Fanny Yang. How rotational invariance of common kernels prevents generalization in high dimensions. In *International Conference on Machine Learning*, pages 2804–2814. PMLR, 2021. 18

[58] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018. 18

[59] Christopher KI Williams. Computation with infinite neural networks. *Neural Computation*, 10(5):1203–1216, 1998. 19

[60] Cosme Louart, Zhenyu Liao, and Romain Couillet. A random matrix approach to neural networks. *The Annals of Applied Probability*, 28(2):1190–1248, 2018. 19

[61] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019. 20

[62] Salomon Bochner. *Harmonic Analysis and the Theory of Probability*. Courier Corporation, 2005. 20

[63] 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. 22, 23, 24

[64] 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*, 2021. 41

[65] Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *International Conference on Artificial Intelligence and Statistics*, pages 1195–1204. PMLR, 2019. 41

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] We clearly discuss the limitation of this work in Conclusion.
   (c) Did you discuss any potential negative societal impacts of your work? [No] Our work is theoretical and generally will have no negative societal impacts.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] The assumptions are clearly stated and well discussed.
   (b) Did you include complete proofs of all theoretical results? [Yes] All of the proofs can be found in the Appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]