Overparameterized Random Feature Regression with Nearly Orthogonal Data

Zhichao Wang
University of California, San Diego

Yizhe Zhu
University of California, Irvine

Abstract

We investigate the properties of random feature ridge regression (RFRR) given by a two-layer neural network with random Gaussian initialization. We study the non-asymptotic behaviors of the RFRR with nearly orthogonal deterministic unit-length input data vectors in the overparameterized regime, where the width of the first layer is much larger than the sample size. Our analysis shows high-probability non-asymptotic concentration results for the training errors, cross-validations, and generalization errors of RFRR centered around their respective values for a kernel ridge regression (KRR). This KRR is derived from an expected kernel generated by a nonlinear random feature map. We then approximate the performance of the KRR by a polynomial kernel matrix obtained from the Hermite polynomial expansion of the activation function, whose degree only depends on the orthogonality among different data points. This polynomial kernel determines the asymptotic behavior of the RFRR and the KRR. Our results hold for a wide variety of activation functions and input data sets that exhibit nearly orthogonal properties. Based on these approximations, we obtain a lower bound for the generalization error of the RFRR for a nonlinear student-teacher model.

1 INTRODUCTION

Random feature regression is closely linked to deep learning theory as a linear model with respect to random features. Training the output layer weight with ridge regression for a neural network with *random* first-layer weight is equivalent to a random feature ridge regression model (RFRR) (Rahimi and Recht, 2007; Cho and Saul, 2009; Daniely et al., 2016; Poole et al., 2016; Schoenholz et al., 2017; Lee et al., 2018; Matthews et al., 2018). The conjugate kernel (CK), whose spectrum has been exploited to study the generalization of random feature regression (Mei et al., 2022), is the Gram matrix of the output of the last hidden layer on the training dataset. The performances (e.g., prediction risk) have been studied by Rahimi and Recht (2007, 2008); Rudi and Rosasco (2017); Mei and Montanari (2019); Mei et al. (2022); Ghorbani et al. (2021). As the width of the neural network increases to infinity, we expect the empirical CK concentrates around its expectation, analogously to the neural tangent kernel (NTK) theory from Jacot et al. (2018). In this overparameterized (or ultra-wide (Arora et al., 2019)) regime, RFRR is asymptotically equivalent to a kernel ridge regression (KRR) model.

In this paper, we focus on the random CK generated by a two-layer fully-connected neural network at random initialization $f : \mathbb{R}^{d\times n} \to \mathbb{R}^n$ such that

$$f(\boldsymbol{X}) := \frac{1}{\sqrt{N}}\boldsymbol{\theta}^\top \sigma\left(\boldsymbol{W}\boldsymbol{X}\right), \tag{1.1}$$

where $\boldsymbol{X} \in \mathbb{R}^{d\times n}$ is the input data matrix, $\boldsymbol{W} \in \mathbb{R}^{N\times d}$ is the weight matrix for the first layer, $\boldsymbol{\theta} \in \mathbb{R}^N$ is the second layer weight, and σ is a nonlinear activation function. Here d is the feature dimension, n is the sample size of the input data, and N is the width of the first layer.

This work focuses on the behavior of the two-layer network under the random initialization of $\boldsymbol{W}$ with sufficiently large width N. We will always view the input data $\boldsymbol{X}$ as a deterministic matrix (independent of the random weights in $\boldsymbol{W}$) with certain assumptions. We fix the random matrix $\boldsymbol{W}$ and only train the second layer $\boldsymbol{\theta}$ via training data $\boldsymbol{X}$. This procedure is the same as the linear regression of random feature vectors $\{\sigma(\boldsymbol{W}\boldsymbol{x}_i) \in \mathbb{R}^N, i \in [n]\}$. The empirical CK matrix is defined by

$$\boldsymbol{K}_N := \frac{1}{N}\sigma\left(\boldsymbol{W}\boldsymbol{X}\right)^\top \sigma\left(\boldsymbol{W}\boldsymbol{X}\right) \in \mathbb{R}^{n\times n}. \tag{1.2}$$

We will show that this random CK matrix will be concentrated around its expected $n \times n$ kernel matrix

$$\boldsymbol{K} := \mathbb{E}\boldsymbol{K}_N = \mathbb{E}_{\boldsymbol{w}}[\sigma(\boldsymbol{w}^\top\boldsymbol{X})^\top\sigma(\boldsymbol{w}^\top\boldsymbol{X})], \tag{1.3}$$

under the spectral norm when width N is sufficiently large, where $\boldsymbol{w}$ is the standard normal random vector in $\mathbb{R}^d$.

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

Random feature regression has already attracted as a random approximation of the reproducing kernel Hilbert space (RKHS) defined by population kernel function $\boldsymbol{K} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ such that

$$\boldsymbol{K}(\boldsymbol{x}_1, \boldsymbol{x}_2) := \mathbb{E}_{\boldsymbol{w}}[\sigma(\langle \boldsymbol{w}, \boldsymbol{x}_1 \rangle)\sigma(\langle \boldsymbol{w}, \boldsymbol{x}_2 \rangle)], \tag{1.4}$$

when width N is sufficiently large (Rahimi and Recht, 2007; Bach, 2013; Rudi and Rosasco, 2017; Bach, 2017; Mei et al., 2022). By an abuse of notation, we use $\boldsymbol{K}$ to represent both the $n \times n$ kernel matrix $\boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X})$ depending on dataset $\boldsymbol{X}$ and the kernel function in (1.4). Denote the output of the first layer by

$$\boldsymbol{\Phi} := \sigma(\boldsymbol{W}\boldsymbol{X}) \in \mathbb{R}^{N \times n}. \tag{1.5}$$

Observe that the rows of the matrix $\boldsymbol{\Phi}$ are independent and identically distributed since only $\boldsymbol{W}$ is random and $\boldsymbol{X}$ is deterministic. Let the i-th row of $\boldsymbol{\Phi}$ be $\boldsymbol{\phi}_i^\top = \sigma(\boldsymbol{w}_i \boldsymbol{X})$ for $1 \le i \le N$, where we denote $\boldsymbol{w}_i \in \mathbb{R}^d$ as the i-th row of weight $\boldsymbol{W}$. Then, CK can be written as $\boldsymbol{K}_N = \frac{1}{N}\sum_{i=1}^N \boldsymbol{\phi}_i \boldsymbol{\phi}_i^\top$, which is a sum of N independent rank-one random matrices in $\mathbb{R}^{n \times n}$. The second moment of any row $\boldsymbol{\phi}_i$ is given by (1.3).

Most of the recent results considered the RFRR with the data points $\boldsymbol{X}$ independently drawn from a specific high-dimensional distribution, e.g., uniform measure on the hypercube or the unit sphere (Misiakiewicz, 2022; Hu and Lu, 2022a; Xiao and Pennington, 2022; Ghorbani et al., 2021) or under the hypercontractivity assumption from (Mei et al., 2022). The analysis of this RFRR usually requires strong assumptions on the data distribution and specific orthogonal polynomial expansions with respect to the distribution. In practice, real-world data cannot satisfy these ideal assumptions, or it is hard to verify them. In this paper, we consider a general deterministic dataset for RFRR. Inspired by Du et al. (2019); Fan and Wang (2020); Wang and Zhu (2021); Donhauser et al. (2021), we point out that the inner products among different unit-length data points, namely the degree of the orthogonality, play an important role in the performances of the RFRR. More precisely, it affects how many degrees of the polynomial this RFRR can consistently learn from the teacher models. The expected kernel model can be truncated as a polynomial inner-product kernel based on this approximate orthogonality of the data points. Combing the concentration of RFRR and this polynomial truncation, we can obtain a lower bound of the generalization error (out-of-sample prediction risk) for RFRR induced by an ultra-wide neural network ($N \gg n$). Since we consider a general distribution-free dataset, we can also analyze cross-validations of RFRR approximated by corresponding cross-validations of the KRR. Our assumptions on the dataset are verifiable even for real-world datasets, and our theory exhibits new ingredients to the study of neural networks with general real-world datasets (Liao and Couillet, 2018; Goldt et al., 2022; Wei et al., 2022).

1.1 Our Contributions

We prove a sequence of sharp concentrations for RFRR around its expected KRR for a general *distribution-free* dataset satisfying an ℓ-orthonormal property (see Assumption 2.3). As long as the width N of the neural network is much larger than sample size n, we can use a KRR to approximate RFRR in terms of in-sample prediction risks, cross-validations, and out-of-sample prediction risks. With a qualitative control of the approximate orthogonality among different data points measured by $\left\|(\boldsymbol{X}^\top \boldsymbol{X})^{\odot(\ell+1)} - \mathrm{Id}\right\|_F$, we can further approximate this KRR by a truncated polynomial inner-product KRR. Meanwhile, we reveal that both RFRR and its corresponding KRR can only consistently learn a polynomial teacher model with a degree at most ℓ. To the best of our knowledge, this is the first work making a connection between the lower bound of the generalization errors of RFRR and KRR, and the orthogonality of deterministic data points. Our main results are stated in Section 2 and proved in Appendix C. The empirical simulations on both synthetic and real-world datasets are presented in Section 3.

1.2 Related Work

Nonlinear Random Matrix Theory When $N \asymp n$, the concentration of the CK matrix around its expectation fails, and the limiting spectrum of the CK with random input dataset has been investigated by Pennington and Worah (2017); Benigni and Péché (2021); Louart et al. (2018); Benigni and Péché (2022); whereas Fan and Wang (2020) studied the spectrum of the CK with similar but stronger assumptions compared to ours on input data and activation functions, and obtained a deformed Marchenko-Pastur distribution (Fan and Wang, 2020). As an application, when $N \asymp n$, the behavior of RFRR is determined by the limiting spectra of the CK (Gerace et al., 2020; Mei and Montanari, 2019; Adlam and Pennington, 2020; Chouard, 2022). Specifically, Louart et al. (2018); Liao et al. (2020); Hu and Lu (2022b); Chouard (2022) studied the training error and empirical test error of RFRR in the proportional limit.

Concentrations of RFRR Rudi and Rosasco (2017) proved the approximation of RFRR when the sample size n and the number of neurons (width) N satisfy $N \asymp \sqrt{n}\log n$. This condition only considered fixed d with i.i.d. data. Moreover, Louart et al. (2018); Wang and Zhu (2021) considered similar concentrations of RFRR for more general datasets. The concentration of random Fourier feature matrices was considered by Chen et al. (2022). The sharp analysis of RFRR (Mei et al., 2022, Theorem 1) gave the precise asymptotic behavior of RFRR and only required $N \gg n$. Our results are consistent with their results on the training errors but relax the assumption on the dataset.

Rotational Invariant Kernels The expected CK and NTK are rotational invariant kernels (Liang et al., 2020),

whence the kernel theory plays a crucial role in analyzing ultra-wide neural networks. In general, the spectra of rotational invariant kernels have been analyzed by El Karoui (2010); Liao and Couillet (2019); Ali et al. (2021) when $n \asymp d$ and such results have been applied in the study of kernel ridge regression in Bartlett et al. (2021); Sahraee-Ardakan et al. (2022). Liao and Couillet (2018, 2019) studied the inner-product kernel induced by random features in the proportional limit, where they can further decompose the expected kernel and extract the useful structure from the data. When $n \asymp d^k$, for $k \in \mathbb{N}$, the performance of inner-product kernel with data uniformly drawn from the unit sphere has been recently studied by Misiakiewicz (2022); Hu and Lu (2022a); Lu and Yau (2022); Xiao and Pennington (2022).

Cross-validations in High Dimensions There is a line of research on cross-validations (Liu and Dobriban, 2019; Jacot et al., 2020b; Miolane and Montanari, 2021; Xu et al., 2021; Hastie et al., 2022; Meanti et al., 2022) for ridge regressions. In high dimensional linear ridge regressions, Hastie et al. (2022) shows precise asymptotic behaviors of cross-validations as $n/d \to \gamma \in (0, \infty)$. Cross-validations help us to tune the hyperparameters and approximate the generalization error of the model (Jacot et al., 2020b; Wei et al., 2022). Most of the above works only focus on linear regression, while our work considers the cross-validations of both nonlinear RFRR and KRR on general datasets.

2 MAIN RESULTS

Notations We use $\mathrm{tr}(A) = \frac{1}{n}\sum_i A_{ii}$ as the normalized trace of a matrix $A \in \mathbb{R}^{n\times n}$ and $\mathrm{Tr}(A) = \sum_i A_{ii}$. Denote vectors by lowercase boldface. $\|A\|$ is the spectral norm for any matrix A, $\|A\|_F$ denotes the Frobenius norm, and $\|\boldsymbol{x}\|$ is the ℓ_2-norm of any vector $\boldsymbol{x}$. Denote $A \odot B$ as the Hadamard product of two matrices A, B of the same size defined by $(A \odot B)_{ij} = A_{ij}B_{ij}$, and $A^{\odot k}$ is the k-th Hadamard product of A with itself. Let $\mathbb{E}_{\boldsymbol{w}}[\cdot]$ be the expectation with respect to the random vector $\boldsymbol{w}$.

2.1 Model Assumptions

Before stating our main results, we list the following assumptions for the random weights $\boldsymbol{W}$, the activation function σ, and input data $\boldsymbol{X}$.

Assumption 2.1. The entries of weight matrix $\boldsymbol{W} \in \mathbb{R}^{N\times d}$ are i.i.d. standard normal random variables $\mathcal{N}(0,1)$.

Let h_k be the k-th normalized Hermite polynomial and $\zeta_k(\sigma)$ be the k-th Hermite coefficient for nonlinear function σ. For more details, see Definition B.1.

Assumption 2.2. Assume $\sigma(x) \in L^4(\mathbb{R}, \Gamma)$, where we denote the standard Gaussian measure denoted by Γ. Define the $L^2(\Gamma)$ and $L^4(\Gamma)$-norm of σ by $\|\sigma\|_2 = (\mathbb{E}[\sigma^2(\xi)])^{1/2}$, $\|\sigma\|_4 = (\mathbb{E}[\sigma^4(\xi)])^{1/4}$, where $\xi \sim \mathcal{N}(0,1)$. In particular, Assumption 2.2 covers many commonly used activation functions, including sigmoid, tanh, ReLU, and leaky ReLU. This is a more general condition compared to previous works by Montanari and Zhong (2022); Wang and Zhu (2021) which assume that σ is Lipschitz or has a polynomial growth rate, and Assumption 2.2 is actually sufficient for the concentrations of training and generalization errors for RFRR.

We consider a sequence of $\boldsymbol{X}_n \in \mathbb{R}^{d_n\times n}$ with growing d_n as $n \to \infty$, where all $\boldsymbol{X}_n$ satisfy the following assumption. Below we drop the dependence on n for the ease of notations. We treat $\boldsymbol{X}$ as a deterministic matrix under the following asymptotic condition.

Assumption 2.3 (ℓ-orthonormal dataset)**.** Suppose that the input data $\boldsymbol{X} \in \mathbb{R}^{d\times n}$ satisfies $\|\boldsymbol{x}_i\| = 1, \forall i \in [n]$. Let $\ell \in \mathbb{N}$ be the smallest integer such that

$$\lim_{n\to\infty} \left\|(\boldsymbol{X}^\top \boldsymbol{X})^{\odot(\ell+1)} - \mathrm{Id}\right\|_F = 0. \tag{2.1}$$

We further assume $\sigma_{>\ell}^2 := \|\sigma\|_2^2 - \sum_{k=1}^{\ell} \zeta_k^2(\sigma) > 0$.

Different from previous work that requires an upper bound on the maximal angle $\varepsilon_n := \max_{i\neq j} |\langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle|$ (Fan and Wang, 2020; Wang and Zhu, 2021; Nguyen and Mondelli, 2020; Hu et al., 2020; Frei et al., 2022), our relaxed Condition (2.1) measures how data points separate from each other *on average*. In particular,

$$\left\|(\boldsymbol{X}^\top \boldsymbol{X})^{\odot(\ell+1)} - \mathrm{Id}\right\|_F \le n\varepsilon_n^{\ell+1}, \tag{2.2}$$

whence (2.1) holds if $n\varepsilon_n^{\ell+1} \to 0$. Here, feature dimension d of the data is implicitly governed by (2.1). In a word, degree ℓ in (2.2) exhibits the average degree of the orthogonality among different data points.

We can also verify Assumption 2.3 for a random dataset. For example, if $\{\boldsymbol{x}_i\}_{i\in[n]}$ are i.i.d. uniformly distributed on $\mathbb{S}^{d-1}$ and $n = \Theta(d^\alpha)$ for $\alpha \in \mathbb{R}_+$, then $\varepsilon_n = O\left(\frac{\log^{1/2} n}{d^{1/2}}\right)$ with high probability (see, for example, Vershynin (2018)), and we can take $\ell = 2\lfloor \alpha \rfloor$ and condition on the high probability event to make $\boldsymbol{X}$ deterministic. A similar argument is also applied by Donhauser et al. (2021), where the distribution of random data can have some covariance structure.

2.2 Power Expansion of the Expected Kernel

For any two unit-length column vectors $\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta$ in $\boldsymbol{X}$, and any two Hermite polynomials h_j, h_k, we have (Nguyen and Mondelli, 2020, Lemma D.2)

$$\mathbb{E}_{\boldsymbol{w}}[h_j(\langle \boldsymbol{w}, \boldsymbol{x}_\alpha\rangle) h_k(\langle \boldsymbol{w}, \boldsymbol{x}_\beta\rangle)] = \delta_{jk}\langle \boldsymbol{x}_\alpha, \boldsymbol{x}_\beta\rangle^k. \tag{2.3}$$

This relation also appears in Oymak and Soltanolkotabi (2020), which directly gives the following power expansion of the expected kernel $\boldsymbol{K}$ in (1.3): $\boldsymbol{K} =$

$\sum_{k=0}^{\infty} \zeta_k^2(\sigma) \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{\odot k}$. Hence, the kernel function $\boldsymbol{K}$ defined in (1.4) is an inner-product kernel. In a concurrent work by Murray et al. (2022), the same power series expansion was applied to the NTK.

In high-dimensional statistics, invariant kernels can be approximated by some simpler models. For instance, El Karoui (2010) proved that the inner-product random kernel matrices with a random dataset could be approximated by a linear random matrix model when $d \asymp n$. The proof by El Karoui (2010) utilized the Taylor approximation of the nonlinear function. In this work, beyond the first-order approximation in El Karoui (2010), we define a degree-ℓ polynomial inner-product kernel by

$$\boldsymbol{K}_\ell := \sum_{k=0}^{\ell} \zeta_k^2(\sigma) \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{\odot k} + \sigma_{>\ell}^2 \operatorname{Id}, \tag{2.4}$$

Here $\sigma_{>\ell}^2$ is an extra ridge parameter added to the polynomial kernel $\sum_{k=0}^{\ell} \zeta_k^2(\sigma) \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{\odot k}$. This extra ridge can be viewed as an *implicit regularization*, especially for the minimum-norm interpolators (Liang et al., 2020; Jacot et al., 2020a; Bartlett et al., 2021).

Assumption 2.3 implies that the off-diagonal entries of $\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{\odot k}$ become negligible when the power k is sufficiently large. Hence, we can truncate $\boldsymbol{K}$ and employ $\boldsymbol{K}_\ell$ as an approximation of $\boldsymbol{K}$ as follows.

Proposition 2.4. Under Assumptions 2.2 and 2.3, let n_0 be the smallest integer such that for all $n \geq n_0$, $\max_{i \neq j} \left|\boldsymbol{x}_i^\top \boldsymbol{x}_j\right| \leq 1/\sqrt{2}$ and

$$\left\| (\boldsymbol{X}^\top \boldsymbol{X})^{\odot(\ell+1)} - \operatorname{Id} \right\|_F \leq \frac{\sigma_{>\ell}^2}{4\|\sigma\|_4^2}. \tag{2.5}$$

We have for all $n \geq n_0$, $\lambda_{\min}(\boldsymbol{K}) \geq \lambda_0 := \frac{1}{2}\sigma_{>\ell}^2$, and

$$\begin{aligned} \|\boldsymbol{K}_\ell - \boldsymbol{K}\| &\leq \sqrt{2}\|\sigma\|_4^2 \left\| \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{\odot \ell+1} - \operatorname{Id} \right\|_F \\ &\leq \frac{\sigma_{>\ell}^2}{2\sqrt{2}}. \end{aligned} \tag{2.6}$$

Remark 2.5 (Comparison to previous work with random dataset). (2.6) is proved by using the inequality $\|\boldsymbol{K}_\ell - \boldsymbol{K}\| \leq \|\boldsymbol{K}_\ell - \boldsymbol{K}\|_F$ and performing an entry-wise expansion of $(\boldsymbol{K}_\ell - \boldsymbol{K})$. Such a Hermite polynomial expansion approach might not be optimal if we know the exact distribution of the random dataset. Previous work from Ghorbani et al. (2021); Mei et al. (2022); Montanari and Zhong (2022); Hu and Lu (2022a) assumed random datasets and random weights with specific distributions. The authors obtained better approximation error bounds using a harmonic analysis approach, where the activation functions and the kernel $\boldsymbol{K}$ were expanded in terms of an orthogonal basis with respect to the distribution of random $\boldsymbol{X}$ and $\boldsymbol{W}$. In many examples, these two distributions are assumed to be the same, which provides a convenient way to expand and approximate $\boldsymbol{K}$ with some degree-ℓ polynomial kernel. Since we do not have any specific data distribution assumption, such an approach cannot be applied to deterministic datasets.

Remark 2.6 (Optimality). In fact, under our Assumption 2.3, the bound (2.6) is tight up to a constant factor. For example, let $\sigma(x) = \sum_{k=0}^{\ell+1} \zeta_k(\sigma) h_k(x)$ be an order-$(\ell+1)$ polynomial with $\zeta_{\ell+1}(\sigma) \neq 0$. Assume $|\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle| = \varepsilon$ for all $i \neq j$ and ℓ is an odd integer. Then

$$\begin{aligned} \|\boldsymbol{K}_\ell - \boldsymbol{K}\| &= \xi_{\ell+1}^2(\sigma) \left\| \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{\odot(\ell+1)} - \operatorname{Id} \right\| \\ &\geq \xi_{\ell+1}^2(\sigma) \varepsilon^{\ell+1} (n-1) \\ &\geq \frac{1}{2} \xi_{\ell+1}^2(\sigma) \left\| \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{\odot(\ell+1)} - \operatorname{Id} \right\|_F. \end{aligned}$$

Proposition 2.4 can be viewed as an extension of (El Karoui, 2010, Theorem 2.1) and (Donhauser et al., 2021, Lemma C.7) for a specific inner-product kernel $\boldsymbol{K}$ induced from the random CK with Gaussian weights, although El Karoui (2010) and Donhauser et al. (2021) considered general rotational invariant random kernels. Our result reveals that we can simply employ such a truncated kernel to approximate the nonlinear kernel because of the ℓ-orthonormal property in Assumption 2.3. In the proof of Donhauser et al. (2021), the authors verified that such property holds for random data with high probability. The same form of $\boldsymbol{K}$ has also been studied by Liang et al. (2020) for the ridgeless regression on some random data $\boldsymbol{X}$ under the polynomial regime ($n \asymp d^\alpha$). Under a stronger regularity assumption on the kernel function, the authors first applied Taylor expansion to get truncated kernel $\boldsymbol{K}_\ell$, then took the Gram-Schmidt process to obtain an orthogonal polynomial basis, which implied a sharper bound on the generalization error for random datasets.

2.3 Concentrations of the RFRR When $N \gg n$

We first consider a two-layer neural network at random initialization defined in (1.1) and estimate the performance of random feature ridge regression in the ultra-high dimensional limit where $N \gg n$. We focus on the linear regression with respect to $\boldsymbol{\theta} \in \mathbb{R}^N$ for predictors of the form $f_{\boldsymbol{\theta}}(\boldsymbol{X}) := \frac{1}{\sqrt{N}} \boldsymbol{\theta}^\top \sigma(\boldsymbol{W}\boldsymbol{X})$, with training data $\boldsymbol{X} \in \mathbb{R}^{d \times n}$ and training labels $\boldsymbol{y} \in \mathbb{R}^n$. The loss function of the ridge regression with a ridge parameter $\lambda \geq 0$ is defined by

$$L(\boldsymbol{\theta}) := \frac{1}{n} \|f_{\boldsymbol{\theta}}(\boldsymbol{X})^\top - \boldsymbol{y}\|^2 + \frac{\lambda}{n} \|\boldsymbol{\theta}\|^2. \tag{2.7}$$

The minimizer of (2.7) denoted by $\hat{\boldsymbol{\theta}} := \arg\min_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$ has an explicit expression $\hat{\boldsymbol{\theta}} = \frac{1}{\sqrt{N}} \boldsymbol{\Phi} \left(\frac{1}{N} \boldsymbol{\Phi}^\top \boldsymbol{\Phi} + \lambda \operatorname{Id}\right)^{-1} \boldsymbol{y}$,

where $\boldsymbol{\Phi}$ is defined in (1.5). The optimal predictor for this RFRR with respect to the loss function in (2.7) is given by

$$\hat{f}_\lambda^{(\mathsf{RF})}(\boldsymbol{x}) := \frac{1}{\sqrt{N}} \hat{\boldsymbol{\theta}}^\top \sigma\left(\boldsymbol{W}\boldsymbol{x}\right), \tag{2.8}$$

where we define an empirical kernel $\boldsymbol{K}_N(\cdot,\cdot) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ as $\boldsymbol{K}_N(\boldsymbol{x},\boldsymbol{z}) := \frac{1}{N}\sigma(\boldsymbol{W}\boldsymbol{x})^\top\sigma(\boldsymbol{W}\boldsymbol{z})$, and the n-dimension row vector is given by $\boldsymbol{K}_N(\boldsymbol{x},\boldsymbol{X}) = [\boldsymbol{K}_N(\boldsymbol{x},\boldsymbol{x}_1),\ldots,\boldsymbol{K}_N(\boldsymbol{x},\boldsymbol{x}_n)]$.

Analogously, consider any kernel function $\boldsymbol{K}(\cdot,\cdot) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ defined in (1.4). Similar to (2.8), the optimal kernel predictor with ridge parameter λ for kernel ridge regression is given by

$$\hat{f}_\lambda^{(\mathsf{K})}(\boldsymbol{x}) := \boldsymbol{K}(\boldsymbol{x},\boldsymbol{X})(\boldsymbol{K}+\lambda\,\mathrm{Id})^{-1}\boldsymbol{y}. \tag{2.9}$$

See Rahimi and Recht (2007); Avron et al. (2017); Liang and Rakhlin (2020); Jacot et al. (2020a); Liu et al. (2021); Bartlett et al. (2021) for additional descriptions about KRR.

We compare the behavior of the two different predictors $\hat{f}_\lambda^{(\mathsf{RF})}(\boldsymbol{x})$ in (2.8) and $\hat{f}_\lambda^{(\mathsf{K})}(\boldsymbol{x})$ in (2.9) with the kernel $\boldsymbol{K}$ defined in (1.4). As N is sufficiently large, the empirical kernel $\boldsymbol{K}_N$ defined in (1.2) will concentrate around its expectation (1.4). From (2.8) and (2.9), the predictors of RFRR and KRR are determined by $\boldsymbol{K}_N$ and $\boldsymbol{K}$, respectively. Therefore, our concentration inequality will help us conclude that the performances of these two predictors are also close to each other as long as the width N is sufficiently larger than sample size n. In the following subsections, we will show that the training error, cross-validations, and generalization error of RFRR can be approximated by the corresponding quantities of KRR defined in (2.9) when N is sufficiently large.

2.3.1 Training Error Approximation

Denote the optimal predictors for the random feature and kernel ridge regressions on the training data $\boldsymbol{X}$ with the ridge parameter $\lambda \geq 0$ by

$$\hat{f}_\lambda^{(\mathsf{RF})}(\boldsymbol{X}) := \left(\hat{f}_\lambda^{(\mathsf{RF})}(\boldsymbol{x}_1),\ldots,\hat{f}_\lambda^{(\mathsf{RF})}(\boldsymbol{x}_n)\right)^\top,$$
$$\hat{f}_\lambda^{(\mathsf{K})}(\boldsymbol{X}) := \left(\hat{f}_\lambda^{(\mathsf{K})}(\boldsymbol{x}_1),\ldots,\hat{f}_\lambda^{(\mathsf{K})}(\boldsymbol{x}_n)\right)^\top,$$

respectively. We first compare the training errors for these two predictors. Let the *training errors* (empirical risks) of these two predictors be

$$E_{\text{train}}^{(\mathsf{K},\lambda)} = \frac{1}{n}\|\hat{f}_\lambda^{(\mathsf{K})}(\boldsymbol{X}) - \boldsymbol{y}\|_2^2, \tag{2.10}$$

$$E_{\text{train}}^{(\mathsf{RF},\lambda)} = \frac{1}{n}\|\hat{f}_\lambda^{(\mathsf{RF})}(\boldsymbol{X}) - \boldsymbol{y}\|_2^2. \tag{2.11}$$

With high probability, the training error of a random feature model and the corresponding kernel model with the same ridge parameter λ can be approximated as follows.

Theorem 2.7 (Training error approximation)**.** *Suppose that Assumptions 2.1, 2.2, and 2.3 hold. Then, with probability at least $1 - N^{-2}$, for any $\lambda \geq 0$, $N/\log^2(N) > C_1 n$, and $n \geq n_0$,*

$$\left|E_{train}^{(\mathsf{RF},\lambda)} - E_{train}^{(\mathsf{K},\lambda)}\right| \leq \frac{C_2\lambda^2\log N\|\boldsymbol{y}\|^2}{\sqrt{nN}}, \tag{2.12}$$

where C_1 and C_2 are positive constants depending only on $\|\sigma\|_4$ and λ_0.

Our bound (2.12) provides a non-asymptotic estimate on the training error approximation, including the case when $\lambda = 0$. From (2.12), assuming $y_i = O(1)$ for all $i \in [n]$, we can conclude that the training error (2.10) concentrates around (2.11) as long as $N/\log^2(N) \gg n$. This result does not rely on the distribution of the data $\boldsymbol{X}$ and how we generate the labels $\boldsymbol{y}$.

The random matrix tool we employ to prove Theorem 2.7 is a *normalized* kernel matrix concentration inequality (Proposition C.1 in Appendix C.2). In contrast to other kernel random matrix concentration results with deterministic $\boldsymbol{X}$ in Louart et al. (2018); Wang and Zhu (2021), a crucial property of our concentration inequality is that it does not depend on $\|\boldsymbol{X}\|$, which guarantees an $o(1)$ approximation error in (2.12) as long as $N/\log^2(N) \gg n$.

2.3.2 Cross-validations Approximation

In the overparameterized regime, the training error approximation in Theorem 2.7 does not directly imply a good approximation of the generalization, but the above analysis of training errors assists us in getting similar approximations on cross-validations of RFRR. Cross-validation (CV) is a common method of model selection and parameter tuning in practice. Especially when practitioners have no access to the data distributions, one can employ CV to approximate the generalization errors of the model (Patil et al., 2022; Jacot et al., 2020b). For more background on cross-validations, we further refer to Arlot and Celisse (2010).

In this subsection, we focus on leave-one-out cross-validation (LOOCV) and generalized cross-validation (GCV) for the predictors $\hat{f}_\lambda^{(\mathsf{RF})}$ and $\hat{f}_\lambda^{(\mathsf{K})}$. Following Hastie et al. (2009), LOOCV is defined by

$$\begin{aligned}\mathrm{CV}_n^{(\mathsf{K},\lambda)} &:= \frac{1}{n}\sum_{i=1}^n (y_i - \hat{f}_{\lambda,-i}^{(\mathsf{K})}(\boldsymbol{x}_i))^2,\\ \mathrm{CV}_n^{(\mathsf{RF},\lambda)} &:= \frac{1}{n}\sum_{i=1}^n (y_i - \hat{f}_{\lambda,-i}^{(\mathsf{RF})}(\boldsymbol{x}_i))^2,\end{aligned} \tag{2.13}$$

where $\hat{f}_{\lambda,-i}^{(\mathsf{K})}$ and $\hat{f}_{\lambda,-i}^{(\mathsf{RF})}$ are KRR and RFRR estimators, respectively, on training data set $\boldsymbol{X}$ with the data point $\boldsymbol{x}_i$ removed. For simplicity, denote $\boldsymbol{K}_\lambda = \boldsymbol{K} + \lambda\,\mathrm{Id}$ and $\boldsymbol{K}_{N,\lambda} = \boldsymbol{K}_N + \lambda\,\mathrm{Id}$. With Schur complement, we can

obtain the "shortcut" formulae for LOOCV as

$$\mathrm{CV}_n^{(\mathsf{K},\lambda)} = \frac{1}{n}\boldsymbol{y}^\top \boldsymbol{K}_\lambda^{-1}\boldsymbol{D}^{-2}\boldsymbol{K}_\lambda^{-1}\boldsymbol{y}, \tag{2.14}$$

$$CV_n^{(\mathsf{RF},\lambda)} = \frac{1}{n}\boldsymbol{y}^\top \boldsymbol{K}_{N,\lambda}^{-1}\boldsymbol{D}_N^{-2}\boldsymbol{K}_{N,\lambda}^{-1}\boldsymbol{y}, \tag{2.15}$$

where $\boldsymbol{D}$ and $\boldsymbol{D}_N$ are diagonal matrices with diagonals $[\boldsymbol{D}]_{ii} = [\boldsymbol{K}_\lambda^{-1}]_{ii}$ and $[\boldsymbol{D}_N]_{ii} = [\boldsymbol{K}_{N,\lambda}^{-1}]_{ii}$, for $i \in [n]$ respectively. The derivations of (2.14) and (2.15) are given in Lemma C.5 of Appendix C.4.

Under certain assumptions, we expect $[\boldsymbol{D}]_{ii}$ and $[\boldsymbol{D}_N]_{ii}$ to concentrate around $\operatorname{tr}\boldsymbol{K}_\lambda^{-1}$ and $\operatorname{tr}\boldsymbol{K}_{N,\lambda}^{-1}$ respectively. Therefore, as an approximation of LOOCV, we define GCV

$$\begin{aligned}\mathrm{GCV}_n^{(\mathsf{K},\lambda)} &:= \left(\lambda \operatorname{tr}(\boldsymbol{K}+\lambda \mathrm{Id})^{-1}\right)^{-2} E_{\text{train}}^{(\mathsf{K},\lambda)},\\ \mathrm{GCV}_n^{(\mathsf{RF},\lambda)} &:= \left(\lambda \operatorname{tr}(\boldsymbol{K}_N+\lambda \mathrm{Id})^{-1}\right)^{-2} E_{\text{train}}^{(\mathsf{RF},\lambda)}.\end{aligned} \tag{2.16}$$

For linear ridge regression models (Hastie et al., 2022), the such approximation is done by applying random matrix theory to replace $\boldsymbol{D}_{ii}$ with $\operatorname{tr}\boldsymbol{K}_\lambda^{-1}$ and $[\boldsymbol{D}_N]_{ii}$ with $\operatorname{tr}\boldsymbol{K}_{N,\lambda}^{-1}$ in (2.14) and (2.15), respectively.

Since these cross-validation estimators are determined by training errors, with Theorem 2.7, we obtain the concentrations of LOOCV and GCV. Theorem 2.8 reveals that under the ultra-wide regime, i.e., $N/\log^2 N \gg n$, GCV and CV estimators of RFRR are close to the corresponding cross-validations of KRR, respectively.

Theorem 2.8 (LOOCV and GCV approximations)**.** *Under the same assumptions as Theorem 2.7, with probability at least* $1-N^{-2}$*, for any* $\lambda \geq 0$*, when* $N/\log^2(N) \geq C(1+\lambda^2)n$ *and* $n \geq n_0$*,*

$$\left|GCV_n^{(\mathsf{K},\lambda)} - GCV_n^{(\mathsf{RF},\lambda)}\right| \leq \frac{c(1+\lambda^4)\log N\,\|\boldsymbol{y}\|^2}{\sqrt{nN}} \tag{2.17}$$

$$\left|CV_n^{(\mathsf{K},\lambda)} - CV_n^{(\mathsf{RF},\lambda)}\right| \leq \frac{c(1+\lambda^4)\log N\|\boldsymbol{y}\|^2}{\sqrt{nN}} \tag{2.18}$$

where $C, c > 0$ *are constants depending only on* σ*.*

The LOOCV and GCV of the linear model have been analyzed by Liu and Dobriban (2019); Xu et al. (2021); Hastie et al. (2022); Patil et al. (2022); Wei et al. (2022). As shown by Hastie et al. (2022), the advantage of LOOCV and GCV is that the optimal ridge parameter tuned by CV is asymptotically the same as the optimal ridge parameter in the high dimensional case. Unlike the results mentioned above, Theorem 2.8 does not require any assumption on data distribution, which opens the door to studying LOOCV and GCV on more general datasets.

In Jacot et al. (2020b), GCV is also called Kernel Alignment Risk Estimator (KARE), and the authors verified that GCV could be used to approximate the generalization error for KRR under a Gaussian universality hypothesis. In addition, Wei et al. (2022) proved that GCV is a good approximation of the generalization error of the linear ridge regression model when a local law for data distribution holds. This may imply that $\mathrm{GCV}_n^{(\mathsf{K},\lambda)}$ also asymptotically approaches the generalization error of KRR when the deterministic matrix $\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X})$ satisfies a local law property. This suggests that the concentrations in Theorem 2.8 could be useful in approximating the generalization error of RFRR. Notably, Wei et al. (2022) considered general datasets under an anisotropic local law hypothesis, while our deterministic data only possesses some orthogonal structures. The proof of Theorem 2.8 in Appendix C.4 opens a new avenue for analyzing LOOCV and GCV for kernel regression (Patil et al., 2022). Following Wei et al. (2022), as a future research direction, we also expect that the GCV estimator of RFRR will converge to its generalization error under certain extra conditions.

2.3.3 Generalization Error Approximation

Different from the controls of in-sample prediction risks and cross-validations in Sections 2.3.1 and 2.3.2, to investigate the generalization error, we introduce further assumptions on the model and the target function under a student-teacher model. The student-teacher model has been investigated in recent works (Gerace et al., 2020; Dhifallah and Lu, 2020; Hu and Lu, 2022b; Goldt et al., 2020; Loureiro et al., 2021; Lin and Dobriban, 2021; Damian et al., 2022; Ba et al., 2022). Since all the data points $\boldsymbol{x}_i$ are deterministic, our model is a fixed design rather than random design (Hsu et al., 2012).

Denote an unknown teacher function by $f^* : \mathbb{R}^d \to \mathbb{R}$. The training labels are generated by $\boldsymbol{y} = f^*(\boldsymbol{X}) + \boldsymbol{\varepsilon}$, where $f^*(\boldsymbol{X}) = (f^*(\boldsymbol{x}_1),\ldots,f^*(\boldsymbol{x}_n))^\top$, and $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma_\varepsilon^2 \mathrm{Id})$. We impose the following assumptions.

Assumption 2.9. Assume that the target function is a nonlinear function with one neuron defined by $f^*(\boldsymbol{x}) = \tau(\langle\boldsymbol{\beta},\boldsymbol{x}\rangle)$, where the random vector $\boldsymbol{\beta} \sim \mathcal{N}(0,\mathrm{Id}_d)$ and $\tau \in L^4(\mathbb{R},\Gamma)$. Suppose that $\zeta_k(\tau) \neq 0$ as long as $\zeta_k(\sigma) \neq 0$, for $0 \leq k \leq \ell$. Training labels are given by $\boldsymbol{y} = \tau(\boldsymbol{X}^\top\boldsymbol{\beta}) + \boldsymbol{\varepsilon} \in \mathbb{R}^n$.

In particular, such an assumption includes the case when σ and τ are the same activation function.

Assumption 2.10. Suppose the test data $\boldsymbol{x} \in \mathbb{R}^d$ satisfies almost surely, $\|\boldsymbol{x}\| = 1$ and

$$\lim_{n\to\infty} \sqrt{n}\left\|(\boldsymbol{X}^\top\boldsymbol{x})^{\odot(\ell+1)}\right\|_2 = 0. \tag{2.19}$$

Assumption 2.10 of the test data $\boldsymbol{x}$ guarantees similar statistical behavior as the training data points in $\boldsymbol{X}$, but we do not impose any specific assumption on its distribution. It is promising to utilize such assumption further to handle statistical models with real-world data (Liao et al., 2020; Seddik et al., 2020).

Assumption 2.10 holds with high probability in many cases when $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n$ are i.i.d. samples from some dis-

tribution $\mathcal{P}$: e.g. Unif$(\mathbb{S}^{d-1})$ and Unif $\left(\{\frac{-1}{\sqrt{d}}, \frac{+1}{\sqrt{d}}\}\right)$ with $n \asymp d^{\alpha}$ and $\ell = 2\lfloor \alpha \rfloor$; an arbitrary distribution such that (2.19) holds almost surely through reject sampling (Casella et al., 2004); or an empirical distribution $\mu_{\hat{n}}$ where $\mu_{\hat{n}} = \frac{1}{\hat{n}} \sum_{i=1}^{\hat{n}} \delta_{\hat{\boldsymbol{x}}_i}$, and $\hat{\boldsymbol{x}}_1, \ldots, \hat{\boldsymbol{x}}_{\hat{n}}$ are deterministic unit vectors such that (2.19) holds for each $\hat{\boldsymbol{x}}_i, i \in [\hat{n}]$.

For any predictor denoted by $\hat{f}$, define the *generalization error* (also called test error) to be the following conditional expectation

$$\mathcal{L}(\hat{f}) := \mathbb{E}\left[|\hat{f}(\boldsymbol{x}) - f^*(\boldsymbol{x})|^2 \,\middle|\, \boldsymbol{X} \right], \tag{2.20}$$

where the expectation is taken over noise $\boldsymbol{\varepsilon}$, test data $\boldsymbol{x}$, and signal $\boldsymbol{\beta}$. Since the dataset $\boldsymbol{X}$ is deterministic in our setting, the conditional expectation in (2.20) becomes $\mathcal{L}(\hat{f}) = \mathbb{E}[|\hat{f}(\boldsymbol{x}) - f^*(\boldsymbol{x})|^2]$. Analogously to the linear case from Ali et al. (2019), this turns out to be the *Bayes risk* for out-of-sample predictors. Viewing $\boldsymbol{\beta}$ as a random signal in the teacher model allows us to get a sharper bound of the generalization error in Theorem 2.11 below.

Under Assumption 2.10, let n_1 be the smallest integer such that for all $n \geq n_1$,

$$\sup_{i \in [n]} |\langle \boldsymbol{x}, \boldsymbol{x}_i \rangle| \leq \frac{1}{\sqrt{2}}, \quad \left\| (\boldsymbol{X}^\top \boldsymbol{x})^{\odot(\ell+1)} \right\|_2 \leq \frac{\sigma_{>\ell}^2}{4\|\sigma\|_4^2}.$$

The following approximation holds for the test error between a random feature predictor and the corresponding kernel predictor in ridge regressions.

Theorem 2.11 (Generalization error approximation)**.** *Suppose Assumptions 2.1, 2.2, 2.9, and 2.10 hold. Then, with probability at least $1 - \log^{-1}(N)$, for any $N/\log^2 N \geq C_1(1+\lambda^2)n$, $n \geq \max\{n_0, n_1\}$, the difference between test errors of RFRR and KRR satisfies*

$$\left| \mathcal{L}(\hat{f}_\lambda^{(\mathsf{RF})}(\boldsymbol{x})) - \mathcal{L}(\hat{f}_\lambda^{(\mathsf{K})}(\boldsymbol{x})) \right| \leq C_2(1+\lambda)\log(N)\sqrt{\frac{n}{N}}, \tag{2.21}$$

for any $\lambda \geq 0$, where constant C_1 depends only on σ, and positive constant C_2 depends only on σ, τ and σ_ε.

When the width $N/\log^2(N) \gg n$, the right-hand side of (2.21) is vanishing. In other words, RFRR has the same generalization error as KRR for ultra-wide neural networks. Notice that Theorem 2.11 covers the ridge-less regression case when $\lambda = 0$.

2.4 Approximation of KRR by a Polynomial KRR

In this subsection, we study a polynomial kernel ridge regression (PKRR) induced by the polynomial kernel $\boldsymbol{K}_\ell$ in (2.4). We define an inner-product kernel by

$$\boldsymbol{K}_\ell(\boldsymbol{x}, \boldsymbol{z}) := \begin{cases} \|\sigma\|_2^2, & \text{if } \boldsymbol{x} = \boldsymbol{z} \\ \sum_{k=0}^{\ell} \zeta_k^2(\sigma)(\boldsymbol{x}^\top \boldsymbol{z})^k, & \text{otherwise,} \end{cases}$$

for any $\boldsymbol{x}, \boldsymbol{z} \in \mathbb{R}^d$. The parameter ℓ defined by Assumption 2.3, is determined by orthogonality among different data points in the training set. In practice, it is hard to implement the expected kernel $\boldsymbol{K}$, whereas this truncated kernel $\boldsymbol{K}_\ell$ with finite many parameters is a simpler model for implementation and theoretical analysis. Similarly, with (2.8) and (2.9), the predictor for kernel regression with respect to $\boldsymbol{K}_\ell$ is denoted by

$$\hat{f}_\lambda^{(\ell)}(\boldsymbol{x}) := \boldsymbol{K}_\ell(\boldsymbol{x}, \boldsymbol{X})(\boldsymbol{K}_\ell + \lambda \,\mathrm{Id})^{-1}\boldsymbol{y}, \tag{2.22}$$

where, by an abuse of notation, we use $\boldsymbol{K}_\ell$ to denote the $n \times n$ polynomial kernel matrix $\boldsymbol{K}_\ell(\boldsymbol{X}, \boldsymbol{X})$. For simplicity, denote $\boldsymbol{K}_{\ell,\lambda} := \boldsymbol{K}_\ell + \lambda \,\mathrm{Id}$ for any $\lambda \geq 0$.

Based on Proposition 2.4, we show that the performances of KRR with kernel $\boldsymbol{K}$ can be approached by the performances of $\hat{f}_\lambda^{(\ell)}$. Denote the training error, the CV, GCV and test error for $\boldsymbol{K}_\ell$ as $E_{\text{train}}^{(\ell,\lambda)}$, $\mathrm{CV}_n^{(\ell,\lambda)}$, $\mathrm{GCV}_n^{(\ell,\lambda)}$, and $\mathcal{L}(\hat{f}_\lambda^{(\ell)}(\boldsymbol{x}))$ respectively. By replacing $\boldsymbol{K}$ by $\boldsymbol{K}_\ell$, we can define these estimators of the PKRR similarly with (2.11), (2.13), and (2.16). Denote $\tilde{\boldsymbol{X}} = [\boldsymbol{X}, \boldsymbol{x}] \in \mathbb{R}^{d \times (n+1)}$ the concatenation of training and test data points. Denote

$$\Delta_\ell = \left\| \left(\boldsymbol{X}^\top \boldsymbol{X} \right)^{\odot \ell+1} - \mathrm{Id} \right\|_F$$

and $\tilde{\Delta}_\ell = \left\| \left(\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}} \right)^{\odot \ell+1} - \mathrm{Id} \right\|_F$. Under (2.1) and (2.19), we have $\Delta_\ell, \tilde{\Delta}_\ell = o_n(1)$.

Theorem 2.12. *Suppose Assumptions 2.2 and 2.3 hold. Then for $n \geq n_0$,*

$$\left| E_{train}^{(\ell,\lambda)} - E_{train}^{(\mathsf{K},\lambda)} \right| \leq \frac{C_1 \lambda^2 \|\boldsymbol{y}\|^2}{n} \Delta_\ell, \tag{2.23}$$

$$\left| GCV_n^{(\ell,\lambda)} - GCV_n^{(\mathsf{K},\lambda)} \right| \leq \frac{C_1(1+\lambda^4)\,\|\boldsymbol{y}\|^2}{n} \Delta_\ell, \tag{2.24}$$

$$\left| CV_n^{(\ell,\lambda)} - CV_n^{(\mathsf{K},\lambda)} \right| \leq \frac{C_1(1+\lambda^4)\,\|\boldsymbol{y}\|^2}{n} \Delta_\ell, \tag{2.25}$$

where $C_1 > 0$ depend only on σ. Furthermore, with Assumptions 2.9 and 2.10, for $n \geq \max\{n_0, n_1\}$, the generalization errors of KRR and PKRR satisfy that

$$\left| \mathcal{L}(\hat{f}_\lambda^{(\ell)}(\boldsymbol{x})) - \mathcal{L}(\hat{f}_\lambda^{(\mathsf{K})}(\boldsymbol{x})) \right| \leq C_2(1+\lambda)\tilde{\Delta}_\ell, \tag{2.26}$$

for some constant $C_2 > 0$ only depending on $\sigma, \tau, \sigma_\varepsilon$.

Based on the definition of ℓ in (2.3), if the training labels satisfy $\|\boldsymbol{y}\|^2 = O(n)$, the left-hand sides of (2.23)-(2.26) are all vanishing as $n \to \infty$. Combing the concentration between RFRR and KRR in Section 2.3, we can now conclude that, in terms of training/test errors and cross-validations, the performance of the RFRR is close to the performance of PKRR defined in (2.22) with high probability as long as $N/\log^2 N \gg n$ and $n \to \infty$. Therefore, the behaviors of the RFRR generated by ultra-wide neural

networks can be characterized by a much simpler PKRR induced by the expected kernel $\boldsymbol{K}$. For (2.26), we can actually verify the estimators $\hat{f}_\lambda^{(\mathsf{K})}(\boldsymbol{x})$ and $\hat{f}_\lambda^{(\mathsf{RF})}(\boldsymbol{x})$ are polynomials of $\boldsymbol{x}$ with degree at most ℓ, which is analogous to the second part of (Donhauser et al., 2021, Theorem C.2). Similar results on neural tangent feature regression are proved by Montanari and Zhong (2022) for uniform spherical distributed data. Due to this simplification, we can further obtain a lower bound of the generalization error of RFRR in the next subsection.

2.5 Polynomial Approximation Barrier for RFRR

The *polynomial approximation barrier* refers to the case when an estimator $\hat{f}_\lambda$ cannot learn any polynomial with a degree larger than a certain threshold (Donhauser et al., 2021). This phenomenon has been shown in both RFRR and KRR (Mei and Montanari, 2019; Ghorbani et al., 2021; Mei et al., 2022; Donhauser et al., 2021) under specific data distribution assumptions, e.g., uniform distributions on the unit sphere or hypercubes (or more general distributions with hypercontractivity assumptions and proper eigenvalue decays) and anisotropic distributions with covariance structures (Loureiro et al., 2021; Gerace et al., 2022).

Define $P_{>\ell} : L^2(\mathbb{R}, \Gamma) \to L^2(\mathbb{R}, \Gamma)$ as the projection onto the span of Hermite polynomials defined in (B.1) with degrees at least $\ell + 1$. Specifically, recalling $\boldsymbol{\beta} \sim \mathcal{N}(0, \mathrm{Id})$ and $\|\boldsymbol{x}\| = 1$, we can get $(P_{>\ell} f^*)(\boldsymbol{x}) = \sum_{k\geq \ell+1} \zeta_k(\tau) h_k(\boldsymbol{\beta}^\top \boldsymbol{x})$, where $\zeta_k(\tau)$ is defined by (B.2). Denote

$$\|P_{>\ell} f^*\|_2^2 = \mathbb{E}_{\boldsymbol{x},\boldsymbol{\beta}} \left(P_{>\ell} f^*(\boldsymbol{x})\right)^2 = \sum_{k\geq \ell+1} \zeta_k^2(\tau).$$

In the following theorem, we prove that the polynomial approximation barrier for RFRR is related to the ℓ-orthonormal properties of the training data. Theorem 2.7 and Theorem 2.11 verify that the RFRR achieves the same errors as KRR, as long as N is sufficiently large. Meanwhile, Theorem 2.12 shows KRR can be further approximated by a simpler polynomial kernel model, whose degree ℓ is determined by the ℓ-orthonormal property in (2.3). Combing these together, RFRR induced by an ultra-wide neural network is asymptotically equivalent to an ℓ-degree PKRR, which naturally implies that RFRR is unable to learn any function with higher-degree terms consistently.

Theorem 2.13 (Lower bound of the generalization error for RFRR)**.** *Under the assumptions of Theorem 2.11, with probability at least* $1 - \log^{-1}(N)$*, when* $N/\log^2 N \gg n$*,*

$$\begin{aligned}
&\mathcal{L}(\hat{f}_\lambda^{(\mathsf{RF})}) \\
&\geq \|P_{>\ell} f^*\|_2^2 + \sigma_{\boldsymbol{\varepsilon}}^2 \mathbb{E}_{\boldsymbol{x}} \left[\boldsymbol{K}_{m,\ell}^\top \boldsymbol{K}_{\lambda,\ell}^{-2} \boldsymbol{K}_{m,\ell}\right] - o_n(1) \\
&\geq \|P_{>\ell} f^*\|_2^2 - o_n(1), \qquad (2.27)
\end{aligned}$$

where $\boldsymbol{K}_{m,\ell} := \boldsymbol{K}_\ell(\boldsymbol{X}, \boldsymbol{x}), \boldsymbol{K}_{\ell,\lambda} := \lambda \,\mathrm{Id} + \boldsymbol{K}_\ell(\boldsymbol{X}, \boldsymbol{X})$.

In Theorem 2.13, we specifically consider a test data point with the ℓ-orthonormal property. This simplifies the teacher model in Assumption 2.9 since $f^*(\boldsymbol{x})$ has the same in distribution as $\tau(\xi)$ for $\xi \sim \mathcal{N}(0, 1)$. Therefore, Theorem 2.13 reveals that RFRR predictor $\hat{f}_\lambda^{(\mathsf{RF})}$ cannot learn the higher degree terms in the Hermite expansion of target function τ. This threshold ℓ is determined by the ℓ-orthonormal property of $\boldsymbol{X}$ in (2.3). The more orthogonal the data points in $\boldsymbol{X}$ are, the lower degree of Hermite polynomials this RFRR predictor can learn consistently.

Remark 2.14 (The variance term)*.* The second term in the first lower bound of (2.27) is related to the variance term in the generalization error of PKRR. This term can be further simplified based on some additional assumptions on the data distribution. Specifically, (Liang et al., 2020, Theorem 2) validated that for sub-Gaussian data,

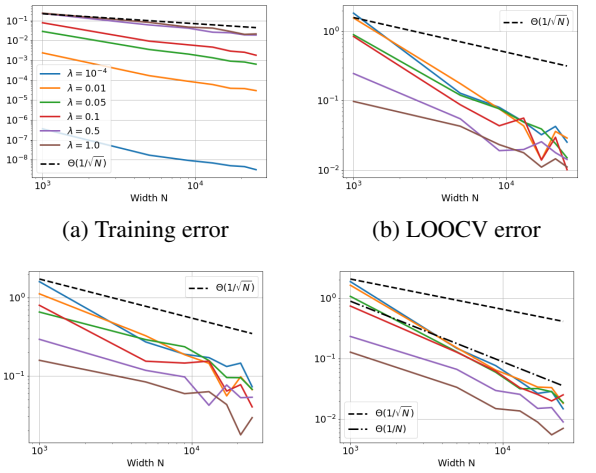
$$\mathrm{Tr}\, \boldsymbol{K}_{\ell,\lambda}^{-1} \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{K}_\ell(\boldsymbol{X}, \boldsymbol{x}) \boldsymbol{K}_\ell(\boldsymbol{x}, \boldsymbol{X})] \boldsymbol{K}_{\ell,\lambda}^{-1} \lesssim \frac{d^\alpha}{n} + \frac{n}{d^{\alpha+1}}$$

with high probability, when $d^\alpha \log d \lesssim n \lesssim d^{\alpha+1}$. Hence, this bound is vanishing in this polynomial regime (see also (Bartlett et al., 2021, Secion 4)). In contrast, under the critical regime $n \asymp d^\alpha$, this variance term, in KRR of any inner-product kernel for uniform spherical distribution, is provably non-degenerate, determined by the Marchenko-Pastur distribution, and may even result in a peak in the prediction curve (Misiakiewicz, 2022; Hu and Lu, 2022a; Xiao and Pennington, 2022).

Remark 2.15 (Comparison to previous work with random dataset)*.* The lower bounds in Theorem 2.13 exhibit the limitation of the RFRR and KRR: (2.27) implies RFRR estimator cannot learn any higher degree polynomials. This is useful when we aim to show that some estimator is superior to this RFRR estimator (Ba et al., 2022; Damian et al., 2022). Compared with the results of Ghorbani et al. (2021); Mei et al. (2022), our results cover more general training datasets for RFRR, though it is not optimal in some specific circumstances (see Remark 2.5), and we only address the single-neuron student-teacher model. Since we study RFRR on a general dataset without any data distribution assumptions, we cannot obtain a more precise characterization of the generalization error as the results by Mei et al. (2022). On the other hand, Donhauser et al. (2021) exhibited a lower bound $\|P_{>(2\lfloor 2\alpha \rfloor)} f^*\|_2^2$ on the generalization error for kernel ridge regression with a general rotational invariant kernel (which is $\|P_{>(2\lfloor \alpha \rfloor)} f^*\|_2^2$ when data $\boldsymbol{x}$ has unit length), where the dataset $\boldsymbol{X} \in \mathbb{R}^{d\times n}$ is random and satisfies $n \asymp d^\alpha$. Under more general assumptions on the dataset $\boldsymbol{X}$, we obtain a similar lower bound $\|P_{>(2\lfloor \alpha \rfloor)} f^*\|_2^2$ from (2.27) for both RFRR and its corresponding KRR.

3 SIMULATIONS

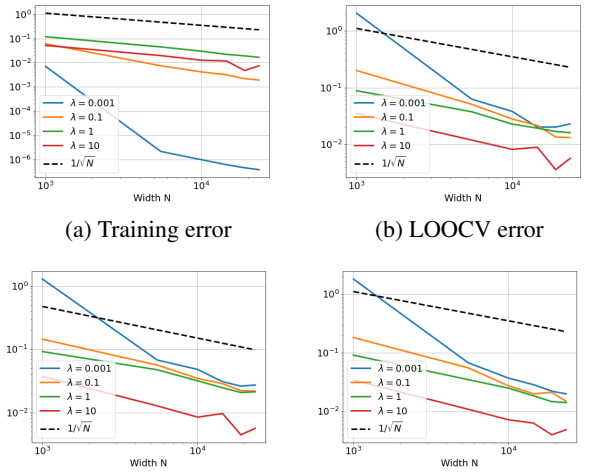
In Figure 1, we empirically verify the concentration bounds we derived in Theorems 2.7, 2.8, 2.11 and 2.12 using i.i.d.



(a) Training error (b) LOOCV error

(c) GCV error (d) Test error

Figure 1: Differences between KRR and RFRR with various ridge parameters λ for (a) training errors, (b) LOOCV errors, (c) GCV errors, and (d) generalization errors. Data $\boldsymbol{X}$ is i.i.d. sampled from uniform distribution $\text{Unif}(\mathbb{S}^{d-1})$ with $d = n = 500$ and training label noise $\sigma_\varepsilon = 0.6$. We repeat each experiment with 7 trials to average. The target function τ is Softplus.



(a) Training error (b) LOOCV error

(c) GCV error (d) Test error

Figure 2: Differences between KRR and RFRR with various ridge parameters λ for (a) training errors, (b) LOOCV errors, (c) GCV errors, and (d) generalization errors. Data points in $\boldsymbol{X}$ are randomly selected from CIFAR-10 with $d = 800$, $n = 1000$ training samples, and without label noise. We repeat each experiment with 5 trials. The target function τ is the ReLU function.

random data $\boldsymbol{X}$, where each data point is sampled from $\text{Unif}(\mathbb{S}^{d-1})$ with $d = n = 500$. As the width N increases, we observe that the differences for training errors, LOOCV, GCV, and generalization errors between RFFR and KRR are all convergent with a rate of at least $1/\sqrt{N}$. The activation function is a polynomial $p(x) := h_0(x) + \frac{1}{\sqrt{6}}h_1(x) + \frac{1}{3}h_2(x) + \frac{1}{6}h_3(x) + \frac{2}{3}h_4(x) + \frac{1}{2}h_5(x)$. For KRR, we utilize the polynomial KRR $\boldsymbol{K}_\ell$ defined by (2.4) with $\ell = 2$ for an approximation of the original $\boldsymbol{K}$. Additional simulations on the synthetic datasets are presented in Appendix A.

Analogously, we investigate the concentrations between RFRR and KRR on real-world data in Figure 2. We randomly select $d = 800$ features for each data vector and $n = 1000$ data points in the CIFAR-10 dataset. After normalizing the data points, we compare the performances of RFRR and KRR induced by the activation function $p(x)$. We observe that our theoretical concentration bound $1/\sqrt{N}$ derived from Section 2 is almost optimal in Figure 2. We expect to further explore which real-world datasets will empirically satisfy the ℓ-orthonormal property defined in Assumption 2.1 as a future direction.

4 CONCLUSION

In this paper, we studied the behavior of random feature ridge regression in the overparameterized regime ($N \gg n$) with a deterministic dataset under an ℓ-orthonormal assumption. In our analysis, we proposed refined matrix concentration inequalities with relaxed assumptions and a convenient Hermite polynomial expansion of the nonlinear activation function. These approaches allow us to go beyond the linear regime (Wang and Zhu, 2021), leading us to study any polynomial kernel approximation of RFRR and obtain new results for general deterministic datasets.

Our analysis has highlighted the impact of the degree of orthogonality among different input data points on the performance of RFRR in terms of training and generalization errors and cross-validation. In addition, Hermite polynomial expansion of σ is a universal way to precisely analyze RFRR induced by any two-layer neural networks with Gaussian random weights. As one-dimensional polynomials, they are easier to implement in practice compared to other orthogonal polynomial expansion approaches (Misiakiewicz, 2022; Hu and Lu, 2022a; Xiao and Pennington, 2022; Ghorbani et al., 2021; Mei et al., 2022) that depend on both data and weight distributions for RFRR. We anticipate that our approach can also be applied to analyze other random kernel matrices, including the empirical NTK, from more general multi-layer neural networks with general deterministic datasets.

Acknowledgements

Z.W. is partially supported by NSF DMS-2055340 and NSF DMS-2154099. Y.Z. is partially supported by NSF-Simons Research Collaborations on the Mathematical and Scientific Foundations of Deep Learning. This work was done in part while both authors were visiting the Simons Institute for the Theory of Computing during the summer of 2022. Z.W. would like to thank Denny Wu for his valuable suggestions and comments. Both authors thank Konstantin Donhauser and Yiqiao Zhong for their helpful discussions.

References

Adlam, B. and Pennington, J. (2020). The neural tangent kernel in high dimensions: Triple descent and a multi-scale theory of generalization. In *International Conference on Machine Learning*, pages 74–84. PMLR.

Ali, A., Kolter, J. Z., and Tibshirani, R. J. (2019). A continuous-time view of early stopping for least squares regression. In *The 22nd international conference on artificial intelligence and statistics*, pages 1370–1378. PMLR.

Ali, H. T., Liao, Z., and Couillet, R. (2021). Random matrices in service of ml footprint: ternary random features with no performance loss. In *International Conference on Learning Representations*.

Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics surveys*, 4:40–79.

Arora, S., Du, S. S., Hu, W., Li, Z., Salakhutdinov, R., and Wang, R. (2019). On exact computation with an infinitely wide neural net. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 8141–8150.

Avron, H., Kapralov, M., Musco, C., Musco, C., Velingker, A., and Zandieh, A. (2017). Random fourier features for kernel ridge regression: Approximation bounds and statistical guarantees. In *International Conference on Machine Learning*, pages 253–262. PMLR.

Ba, J., Erdogdu, M. A., Suzuki, T., Wang, Z., Wu, D., and Yang, G. (2022). High-dimensional asymptotics of feature learning: How one gradient step improves the representation. *arXiv preprint arXiv:2205.01445*.

Bach, F. (2013). Sharp analysis of low-rank kernel matrix approximations. In *Conference on Learning Theory*, pages 185–209. PMLR.

Bach, F. (2017). On the equivalence between kernel quadrature rules and random feature expansions. *The Journal of Machine Learning Research*, 18(1):714–751.

Bai, Z. and Silverstein, J. W. (2010). *Spectral analysis of large dimensional random matrices*, volume 20. Springer.

Bartlett, P. L., Montanari, A., and Rakhlin, A. (2021). Deep learning: a statistical viewpoint. *Acta numerica*, 30:87–201.

Benaych-Georges, F. and Knowles, A. (2017). Lectures on the local semicircle law for wigner matrices. In *Advanced topics in random matrices*, pages 1–90. Panor. Synthèses, 53, Soc. Math. France, Paris.

Benigni, L. and Péché, S. (2021). Eigenvalue distribution of some nonlinear models of random matrices. *Electronic Journal of Probability*, 26:1–37.

Benigni, L. and Péché, S. (2022). Largest eigenvalues of the conjugate kernel of single-layered neural networks. *arXiv preprint arXiv:2201.04753*.

Casella, G., Robert, C. P., and Wells, M. T. (2004). Generalized accept-reject sampling schemes. *Lecture Notes-Monograph Series*, pages 342–347.

Chen, Z., Schaeffer, H., and Ward, R. (2022). Concentration of random feature matrices in high-dimensions. In *Proceedings of Mathematical and Scientific Machine Learning*, volume 190 of *Proceedings of Machine Learning Research*, pages 287–302. PMLR.

Cho, Y. and Saul, L. K. (2009). Kernel methods for deep learning. In *Advances in Neural Information Processing Systems*, pages 342–350.

Chouard, C. (2022). Quantitative deterministic equivalent of sample covariance matrices with a general dependence structure. *arXiv preprint arXiv:2211.13044*.

Damian, A., Lee, J., and Soltanolkotabi, M. (2022). Neural networks can learn representations with gradient descent. In *Conference on Learning Theory*, pages 5413–5452. PMLR.

Daniely, A., Frostig, R., and Singer, Y. (2016). Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity. In *Advances In Neural Information Processing Systems*, pages 2253–2261.

Dhifallah, O. and Lu, Y. M. (2020). A precise performance analysis of learning with random features. *arXiv preprint arXiv:2008.11904*.

Donhauser, K., Wu, M., and Yang, F. (2021). How rotational invariance of common kernels prevents generalization in high dimensions. In *International Conference on Machine Learning*, pages 2804–2814. PMLR.

Du, S. S., Zhai, X., Poczos, B., and Singh, A. (2019). Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*.

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

Fan, Z. and Wang, Z. (2020). Spectra of the conjugate kernel and neural tangent kernel for linear-width neural networks. In *Advances in Neural Information Processing Systems*, volume 33, pages 7710–7721. Curran Associates, Inc.

Frei, S., Vardi, G., Bartlett, P. L., Srebro, N., and Hu, W. (2022). Implicit bias in leaky relu networks trained on high-dimensional data. *arXiv preprint arXiv:2210.07082*.

Gerace, F., Krzakala, F., Loureiro, B., Stephan, L., and Zdeborová, L. (2022). Gaussian universality of linear classifiers with random labels in high-dimension. *arXiv preprint arXiv:2205.13303*.

Gerace, F., Loureiro, B., Krzakala, F., Mézard, M., and Zdeborová, L. (2020). Generalisation error in learning with random features and the hidden manifold model. In *International Conference on Machine Learning*, pages 3452–3462. PMLR.

Ghorbani, B., Mei, S., Misiakiewicz, T., and Montanari, A. (2021). Linearized two-layers neural networks in high dimension. *The Annals of Statistics*, 49(2):1029–1054.

Goldt, S., Loureiro, B., Reeves, G., Krzakala, F., Mézard, M., and Zdeborová, L. (2022). The gaussian equivalence of generative models for learning with shallow neural networks. In *Mathematical and Scientific Machine Learning*, pages 426–471. PMLR.

Goldt, S., Mézard, M., Krzakala, F., and Zdeborová, L. (2020). Modeling the influence of data structure on learning in neural networks: The hidden manifold model. *Physical Review X*, 10(4):041044.

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. (2022). Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

Horn, R. A. and Johnson, C. R. (2012). *Matrix analysis*. Cambridge university press.

Hsu, D., Kakade, S. M., and Zhang, T. (2012). Random design analysis of ridge regression. In *Conference on learning theory*, pages 9–1. JMLR Workshop and Conference Proceedings.

Hu, H. and Lu, Y. M. (2022a). Sharp asymptotics of kernel ridge regression beyond the linear regime. *arXiv preprint arXiv:2205.06798*.

Hu, H. and Lu, Y. M. (2022b). Universality laws for high-dimensional learning with random features. *IEEE Transactions on Information Theory*.

Hu, W., Xiao, L., Adlam, B., and Pennington, J. (2020). The surprising simplicity of the early-time learning dynamics of neural networks. In *Advances in Neural Information Processing Systems*, volume 33, pages 17116–17128. Curran Associates, Inc.

Jacot, A., Gabriel, F., and Hongler, C. (2018). Neural tangent kernel: convergence and generalization in neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 8580–8589.

Jacot, A., Simsek, B., Spadaro, F., Hongler, C., and Gabriel, F. (2020a). Implicit regularization of random feature models. In *International Conference on Machine Learning*, pages 4631–4640. PMLR.

Jacot, A., Simsek, B., Spadaro, F., Hongler, C., and Gabriel, F. (2020b). Kernel alignment risk estimator: Risk prediction from training data. *Advances in Neural Information Processing Systems*, 33:15568–15578.

Lee, J., Bahri, Y., Novak, R., Schoenholz, S. S., Pennington, J., and Sohl-Dickstein, J. (2018). Deep neural networks as Gaussian processes. In *International Conference on Learning Representations*.

Liang, T. and Rakhlin, A. (2020). Just interpolate: Kernel "ridgeless" regression can generalize. *The Annals of Statistics*, 48(3):1329–1347.

Liang, T., Rakhlin, A., and Zhai, X. (2020). On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels. In *Conference on Learning Theory*, pages 2683–2711. PMLR.

Liao, Z. and Couillet, R. (2018). On the spectrum of random features maps of high dimensional data. In *International Conference on Machine Learning*, pages 3063–3071. PMLR.

Liao, Z. and Couillet, R. (2019). On inner-product kernels of high dimensional data. In *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 579–583. IEEE.

Liao, Z., Couillet, R., and Mahoney, M. W. (2020). A random matrix analysis of random fourier features: beyond the gaussian kernel, a precise phase transition, and the corresponding double descent. In *34th Conference on Neural Information Processing Systems*.

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

Liu, F., Liao, Z., and Suykens, J. (2021). Kernel regression in high dimensions: Refined analysis beyond double descent. In *International Conference on Artificial Intelligence and Statistics*, pages 649–657. PMLR.

Liu, S. and Dobriban, E. (2019). Ridge regression: Structure, cross-validation, and sketching. In *International Conference on Learning Representations*.

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

Loureiro, B., Gerbelot, C., Cui, H., Goldt, S., Krzakala, F., Mezard, M., and Zdeborová, L. (2021). Learning curves of generic features maps for realistic datasets with a teacher-student model. *Advances in Neural Information Processing Systems*, 34:18137–18151.

Lu, Y. M. and Yau, H.-T. (2022). An equivalence principle for the spectrum of random inner-product kernel matrices. *arXiv preprint arXiv:2205.06308*.

Matthews, A. G. d. G., Hron, J., Rowland, M., Turner, R. E., and Ghahramani, Z. (2018). Gaussian process behaviour in wide deep neural networks. In *International Conference on Learning Representations*.

Meanti, G., Carratino, L., De Vito, E., and Rosasco, L. (2022). Efficient hyperparameter tuning for large scale kernel ridge regression. In *International Conference on Artificial Intelligence and Statistics*, pages 6554–6572. PMLR.

Mei, S., Misiakiewicz, T., and Montanari, A. (2022). Generalization error of random feature and kernel methods: hypercontractivity and kernel matrix concentration. *Applied and Computational Harmonic Analysis*, 59:3–84.

Mei, S. and Montanari, A. (2019). The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*.

Miolane, L. and Montanari, A. (2021). The distribution of the lasso: Uniform control over sparse balls and adaptive parameter tuning. *The Annals of Statistics*, 49(4):2313–2335.

Misiakiewicz, T. (2022). Spectrum of inner-product kernel matrices in the polynomial regime and multiple descent phenomenon in kernel ridge regression. *arXiv preprint arXiv:2204.10425*.

Montanari, A. and Zhong, Y. (2022). The interpolation phase transition in neural networks: Memorization and generalization under lazy training. *The Annals of Statistics*, 50(5):2816–2847.

Murray, M., Jin, H., Bowman, B., and Montufar, G. (2022). Characterizing the spectrum of the NTK via a power series expansion. *arXiv preprint arXiv:2211.07844*.

Nguyen, Q. and Mondelli, M. (2020). Global convergence of deep networks with one wide layer followed by pyramidal topology. In *34th Conference on Neural Information Processing Systems*, volume 33.

Oymak, S. and Soltanolkotabi, M. (2020). Toward moderate overparameterization: Global convergence guarantees for training shallow neural networks. *IEEE Journal on Selected Areas in Information Theory*, 1(1):84–105.

Patil, P., Rinaldo, A., and Tibshirani, R. (2022). Estimating functionals of the out-of-sample error distribution in high-dimensional ridge regression. In *International Conference on Artificial Intelligence and Statistics*, pages 6087–6120. PMLR.

Pennington, J. and Worah, P. (2017). Nonlinear random matrix theory for deep learning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Poole, B., Lahiri, S., Raghu, M., Sohl-Dickstein, J., and Ganguli, S. (2016). Exponential expressivity in deep neural networks through transient chaos. In *Advances in Neural Information Processing Systems*, pages 3360–3368.

Rahimi, A. and Recht, B. (2007). Random features for large-scale kernel machines. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, pages 1177–1184.

Rahimi, A. and Recht, B. (2008). Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. *Advances in neural information processing systems*, 21.

Rudi, A. and Rosasco, L. (2017). Generalization properties of learning with random features. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Sahraee-Ardakan, M., Emami, M., Pandit, P., Rangan, S., and Fletcher, A. K. (2022). Kernel methods and multilayer perceptrons learn linear models in high dimensions. *arXiv preprint arXiv:2201.08082*.

Schoenholz, S. S., Gilmer, J., Ganguli, S., and Sohl-Dickstein, J. (2017). Deep information propagation. In *International Conference on Learning Representations*.

Seddik, M. E. A., Louart, C., Tamaazousti, M., and Couillet, R. (2020). Random matrix theory proves that deep learning representations of gan-data behave as gaussian mixtures. In *International Conference on Machine Learning*, pages 8573–8582. PMLR.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.

Wang, Z. and Zhu, Y. (2021). Deformed semicircle law and concentration of nonlinear random matrices for ultrawide neural networks. *arXiv preprint arXiv:2109.09304*.

Wei, A., Hu, W., and Steinhardt, J. (2022). More than a toy: Random matrix models predict how real-world neural representations generalize. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 23549–23588. PMLR.

Xiao, L. and Pennington, J. (2022). Precise learning curves and higher-order scaling limits for dot product kernel regression. *arXiv preprint arXiv:2205.14846*.

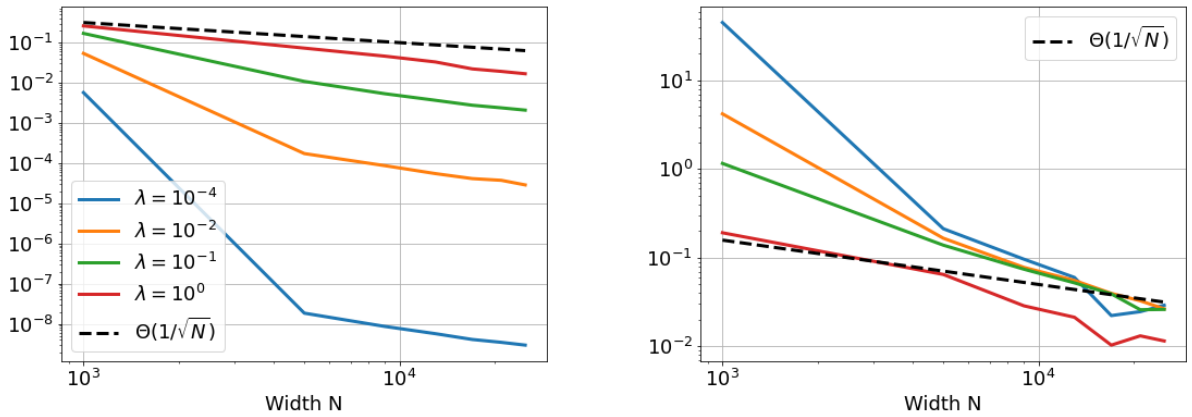
Xu, J., Maleki, A., Rad, K. R., and Hsu, D. (2021). Consistent risk estimation in moderately high-dimensional linear regression. *IEEE Transactions on Information Theory*, 67(9):5997–6030.

A ADDITIONAL SIMULATIONS

As a complementary, Figure 3 shows the convergence rates for the differences in training errors, LOOCV errors, GCV errors, and generalization errors between RFRR and KRR. In this experiment, the data points are i.i.d. sampled from $\mathrm{Unif}(\mathbb{S}^{d-1})$ with $d = 500$ and training samples $n = 1000$. The activation function is a degree-5 polynomial $p(x) = h_0(x) + \frac{1}{\sqrt{6}}h_1(x) + \frac{1}{3}h_2(x) + \frac{1}{6}h_3(x) + \frac{2}{3}h_4(x) + \frac{1}{2}h_5(x)$, where Hermite polynomials are defined in Definition B.1. As an approximation of the kernel $\boldsymbol{K}$ generated by $\sigma(x) = p(x)$, we can consider $\boldsymbol{K}_2$ defined by

$$\boldsymbol{K}_2 = \mathbf{1}\mathbf{1}^\top + \frac{1}{6}\boldsymbol{X}^\top\boldsymbol{X} + \frac{1}{9}(\boldsymbol{X}^\top\boldsymbol{X})^{\odot 2} + \frac{26}{36}\,\mathrm{Id}\,.$$

We employ this simple kernel $\boldsymbol{K}_2$ to compute the performances of KRR and compare them with the performances of RFRR generated by σ and (1.2). In this simulation, we consider a teacher model defined by Assumption 2.9 where τ is the Softplus function. Similarly with Figures 1 and 2, these results of the simulation match with our theorems in Section 2.



(a) Training error

(b) LOOCV error

(c) GCV error

(d) Test error

Figure 3: Differences between KRR and RFRR with various ridge parameters λ in terms of (a) training errors, (b) LOOCV errors, (c) GCV errors, and (d) generalization errors. Here, data $\boldsymbol{X}$ is sampled from $\mathrm{Unif}(\mathbb{S}^{d-1})$ with $d = 500$, $n = 1000$ and training label noise $\sigma_\varepsilon = 0.3$. We repeat each experiment with 5 trials to average. The target function τ is Softplus function.

B ADDITIONAL NOTATIONS AND DEFINITIONS

We denote Id as the identity matrix. Let $\boldsymbol{K}_\lambda = \boldsymbol{K} + \lambda\,\mathrm{Id}$ where $\boldsymbol{K}$ is defined by (1.3) and $\lambda \geq 0$ is the ridge parameter. Denote $\boldsymbol{K}_{N,\lambda} = \boldsymbol{K}_N + \lambda\,\mathrm{Id}$ where $\boldsymbol{K}_N$ is (1.2). Conventionally, let $\|\cdot\|$ be the ℓ_2-norm for vectors and $\ell_2 \to \ell_2$

operator norm for matrices. Let $\preccurlyeq$ be the Loewner order for positive semi-definite matrices. For any matrix $\boldsymbol{A} \in \mathbb{R}_{n\times n}$, $[\boldsymbol{A}]_{i,j}$ denotes the (i,j) entry of $\boldsymbol{A}$, and $[\boldsymbol{A}]_{[i,:]}$ denotes the i-th row of $\boldsymbol{A}$ for any $i,j \in [n]$. Recall that the constant $\lambda_0 = \frac{1}{2}\sigma^2_{>\ell} > 0$. In the following proofs in Appendix C, all the constants are universal and do not depend on n, d, and N.

The following normalized Hermite polynomials are necessary for expanding σ and approximating $\boldsymbol{K}$ by a polynomial kernel in Section 2.2 under Gaussian distributions.

Definition B.1 (Normalized Hermite polynomial)**.** The k-th normalized Hermite polynomial is given by

$$h_k(x) = \frac{1}{\sqrt{k!}}(-1)^k e^{x^2/2}\frac{d^k}{dx^k}e^{-x^2/2}. \tag{B.1}$$

These polynomials $\{h_k\}_{k=0}^{\infty}$ form an orthogonal basis of $L^2(\mathbb{R},\Gamma)$, where Γ denotes the standard Gaussian distribution. For any $\sigma_1, \sigma_2 \in L^2(\mathbb{R},\Gamma)$, the inner product, with respect to the standard Gaussian measure, is defined by

$$\langle \sigma_1, \sigma_2\rangle_\Gamma = \int_{-\infty}^{\infty} \sigma_1(x)\sigma_2(x)\frac{e^{-x^2/2}}{\sqrt{2\pi}}dx.$$

Based on the definition, every function $\sigma \in L^2(\mathbb{R},\Gamma)$ can be expanded as $\sigma(x) = \sum_{k=0}^{\infty}\zeta_k(\sigma)h_k(x)$, where $\zeta_k(\sigma)$ is the k-th Hermite coefficient given by

$$\zeta_k(\sigma) = \int_{-\infty}^{\infty}\sigma(x)h_k(x)\frac{e^{-x^2/2}}{\sqrt{2\pi}}dx, \tag{B.2}$$

and $\|\sigma\|_2^2 = \sum_{k=0}^{\infty}\zeta_k^2(\sigma)$. Moreover, we have $\langle h_k, h_j\rangle_\Gamma = \mathbb{E}[h_k(\xi)h_j(\xi)] = \delta_{j,k}$ for any $\xi \sim \mathcal{N}(0,1)$ and $k,j \in \mathbb{N}$. For more properties of Hermite polynomials, see Oymak and Soltanolkotabi (2020); Nguyen and Mondelli (2020).

C PROOFS OF MAIN RESULTS IN SECTION 2

C.1 Proof of Proposition 2.4

By the Hermite polynomial expansion of σ, for $i,j \in [n]$, we have

$$\boldsymbol{K}_{ij} = \mathbb{E}_{\boldsymbol{w}}[\sigma(\boldsymbol{w}_i^\top \boldsymbol{x}_i)\sigma(\boldsymbol{w}_i^\top \boldsymbol{x}_j)] = \sum_{k=0}^{\infty}\xi_k^2(\sigma)\langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle^k.$$

Thus, we can expand this kernel as

$$\boldsymbol{K} = \sum_{k=0}^{\infty}\xi_k^2(\sigma)\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{\odot k}, \quad \boldsymbol{K} - \boldsymbol{K}_\ell = \sum_{k=\ell+1}^{\infty}\xi_k^2(\sigma)\left(\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{\odot k} - \mathrm{Id}\right).$$

Then by Cauchy's inequality, for $n \geq n_0$,

$$\begin{aligned}
\|\boldsymbol{K} - \boldsymbol{K}_\ell\|^2 \leq \|\boldsymbol{K} - \boldsymbol{K}_\ell\|_F^2 &= \sum_{i\neq j}\left(\sum_{k=\ell+1}^{\infty}\xi_k^2(\sigma)\langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle^k\right)^2 \\
&\leq \sum_{i\neq j}\left(\sum_{k=\ell+1}^{\infty}\xi_k^4(\sigma)\right)\left(\sum_{k=\ell+1}^{\infty}\langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle^{2k}\right) \\
&\leq \|\sigma\|_4^4\sum_{i\neq j}\frac{\langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle^{2\ell+2}}{1-\max|\langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle|^2} \\
&\leq 2\|\sigma\|_4^4\left\|\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{\odot\ell+1} - \mathrm{Id}\right\|_F^2.
\end{aligned}$$

Therefore, from (2.5),

$$\|\boldsymbol{K} - \boldsymbol{K}_\ell\| \leq \sqrt{2}\|\sigma\|_4^2\left\|\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{\odot\ell+1} - \mathrm{Id}\right\|_F \leq \frac{1}{2\sqrt{2}}\sigma^2_{>\ell}.$$

Since

$$\left(\boldsymbol{X}^\top \boldsymbol{X}\right)_{ij}^{\odot k} = \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle^k = \langle \boldsymbol{x}_i \otimes \cdots \otimes \boldsymbol{x}_i, \boldsymbol{x}_j \otimes \cdots \otimes \boldsymbol{x}_j \rangle,$$

where $\boldsymbol{x}_i \otimes \cdots \otimes \boldsymbol{x}_i$ is the k-th tensor product of $\boldsymbol{x}_i$, $\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{\odot k}$ is positive semidefinite. Then

$$\lambda_{\min}(\boldsymbol{K}) \geq \lambda_{\min}(\boldsymbol{K}_\ell) - \|\boldsymbol{K} - \boldsymbol{K}_\ell\| \geq \sigma_{>\ell}^2 - \frac{1}{2\sqrt{2}}\sigma_{>\ell}^2 > \frac{1}{2}\sigma_{>\ell}^2 \equiv \lambda_0,$$

Notice that $\lambda_0 > 0$ because $\sigma_{>\ell}^2 > 0$ from Assumption 2.3.

C.2 Concentration Inequality for Normalized Random Kernel Matrices

Now we introduce the concentration inequality for $\boldsymbol{K}_N$ in a normalized version, which is the cornerstone for proving Theorem 2.7. Similar concentration results were also obtained in Theorem 3.2 of Montanari and Zhong (2022) for the neural tangent kernel (NTK), where the data matrix $\boldsymbol{X}$ is assumed to be uniformly random, and the activation function is assumed to have a polynomial growth rate, while we make no distribution assumption on $\boldsymbol{X}$ and only assume $\|\sigma\|_4$ is finite. To consider a normalized version of the kernel matrices, we need to consider $\boldsymbol{K}_\lambda^{-1}$. Under Assumption 2.3, we use Proposition 2.4 to make sure $\boldsymbol{K}_\lambda$ is invertible when $\lambda = 0$.

Proposition C.1 (Normalized random kernel matrix concentration)**.** Suppose that $\sigma \in L^4(\mathbb{R}, \Gamma)$. Then, under the same assumptions of Theorem 2.7, there exists some positive constants $C_1, C_2 > 0$ depending on σ, such that for any N satisfying $N/\log^2(N) > C_1 n$ and any $\lambda \geq 0, n \geq n_0$, we have

$$\left\| \boldsymbol{K}_\lambda^{-\frac{1}{2}} \left(\boldsymbol{K}_N - \boldsymbol{K}\right) \boldsymbol{K}_\lambda^{-\frac{1}{2}} \right\| \leq C_2 \log(N) \sqrt{\frac{n}{N}}, \tag{C.1}$$

with probability at least $1 - N^{-2}$, where $\boldsymbol{K}_\lambda = \boldsymbol{K} + \lambda \,\mathrm{Id}$.

Proof. Denote $\tilde{\sigma}(x) := \sigma(x)\mathbf{1}_{|x| \leq B}$, where B is a parameter to be decided later. Define

$$\begin{aligned}
\boldsymbol{K}_N &= \frac{1}{N}\sum_{i=1}^N \sigma(\boldsymbol{w}_i^\top \boldsymbol{X})^\top \sigma(\boldsymbol{w}_i^\top \boldsymbol{X}), & \boldsymbol{K} &= \mathbb{E}_{\boldsymbol{w}}[\sigma(\boldsymbol{w}^\top \boldsymbol{X})^\top \sigma(\boldsymbol{w}^\top \boldsymbol{X})], \\
\tilde{\boldsymbol{K}}_N &= \frac{1}{N}\sum_{i=1}^N \tilde{\sigma}(\boldsymbol{w}_i^\top \boldsymbol{X})^\top \tilde{\sigma}(\boldsymbol{w}_i^\top \boldsymbol{X}), & \tilde{\boldsymbol{K}} &= \mathbb{E}_{\boldsymbol{w}}[\tilde{\sigma}(\boldsymbol{w}^\top \boldsymbol{X})^\top \tilde{\sigma}(\boldsymbol{w}^\top \boldsymbol{X})].
\end{aligned}$$

For simplicity, we denote $\tilde{\boldsymbol{K}}_\lambda := \tilde{\boldsymbol{K}} + \lambda \,\mathrm{Id}$. Define

$$\boldsymbol{H}_i := \frac{1}{N} \tilde{\boldsymbol{K}}_\lambda^{-1/2} \tilde{\sigma}(\boldsymbol{w}_i^\top \boldsymbol{X})^\top \tilde{\sigma}(\boldsymbol{w}_i^\top \boldsymbol{X}) \tilde{\boldsymbol{K}}_\lambda^{-1/2}.$$

Notice that Proposition 2.4 implies that $\left\|\boldsymbol{K}_\lambda^{-1}\right\| \leq \lambda_0^{-1}$ for $\lambda \geq 0$. Firstly, based on the truncated function $\tilde{\sigma}(x)$, we have that for some universal constant $c > 0$,

$$\mathbb{P}\left(\boldsymbol{K}_N \neq \tilde{\boldsymbol{K}}_N\right) \leq \mathbb{P}\left(\max_{i \in [N], k \in [n]} |\boldsymbol{w}_i^\top \boldsymbol{x}_k| > B\right) \leq Nn\mathbb{P}\left(|\xi| > B\right) \leq cNn \exp\left(-B^2/2\right), \tag{C.2}$$

where $\xi \sim \mathcal{N}(0,1)$. Define the event by $A_i := \{\boldsymbol{w} : |\boldsymbol{w}^\top \boldsymbol{x}_i| \leq B\}$ for $i \in [n]$. Entry-wisely, we have

$$\begin{aligned}
\left|[\boldsymbol{K} - \tilde{\boldsymbol{K}}]_{i,j}\right| &= \left|\mathbb{E}_{\boldsymbol{w}}[\sigma(\boldsymbol{w}^\top \boldsymbol{x}_i)\sigma(\boldsymbol{w}^\top \boldsymbol{x}_j)\mathbf{1}_{A_j^c \cap A_j^c}]\right| \\
&\leq \mathbb{E}_{\boldsymbol{w}}[\sigma(\boldsymbol{w}^\top \boldsymbol{x}_i)^4]^{1/2} \mathbb{E}[\mathbf{1}_{A_j^c \cap A_j^c}]^{1/2} \\
&\leq \sqrt{2}\mathbb{E}[\sigma(\xi)^4]^{1/2} \mathbb{P}\left(A_i^c\right)^{1/2} \leq C_0 \exp\left(-B^2/4\right),
\end{aligned}$$

for some constant $C_0 > 0$ which only depends on $\|\sigma\|_4$. Therefore, $\|\boldsymbol{K} - \tilde{\boldsymbol{K}}\| \leq \|\boldsymbol{K} - \tilde{\boldsymbol{K}}\|_F \leq C_0 n \exp\left(-B^2/4\right)$ and

$$\left\| \boldsymbol{K}_\lambda^{-1/2} \left(\boldsymbol{K} - \tilde{\boldsymbol{K}}\right) \boldsymbol{K}_\lambda^{-1/2} \right\| \leq \frac{C_0}{\lambda_0} n \exp\left(-B^2/4\right). \tag{C.3}$$

For

$$B \geq \sqrt{4\log\left(\frac{2C_0 n}{\lambda_0}\right)}, \tag{C.4}$$

the above equation also implies that $\left\|\boldsymbol{K}-\tilde{\boldsymbol{K}}\right\| \leq \frac{\lambda_0}{2}$, and

$$\begin{aligned}\left\|\tilde{\boldsymbol{K}}_\lambda^{1/2}\boldsymbol{K}_\lambda^{-1/2}\right\|^2 &= \left\|\boldsymbol{K}_\lambda^{-1/2}\tilde{\boldsymbol{K}}_\lambda\boldsymbol{K}_\lambda^{-1/2}\right\| \\ &\leq \left\|\boldsymbol{K}_\lambda^{-1/2}\left(\boldsymbol{K}-\tilde{\boldsymbol{K}}\right)\boldsymbol{K}_\lambda^{-1/2}\right\| + \left\|\boldsymbol{K}_\lambda^{-1/2}\boldsymbol{K}_\lambda\boldsymbol{K}_\lambda^{-1/2}\right\| \leq \frac{3}{2}.\end{aligned} \tag{C.5}$$

Therefore, the smallest eigenvalues of $\boldsymbol{K}_\lambda$ and $\tilde{\boldsymbol{K}}_\lambda$ satisfy

$$\lambda_{\min}(\tilde{\boldsymbol{K}}_\lambda) \geq \lambda_{\min}(\boldsymbol{K}_\lambda) - \left\|\boldsymbol{K}-\tilde{\boldsymbol{K}}\right\| \geq \frac{\lambda_0}{2} > 0. \tag{C.6}$$

It suffices to analyze $\left\|\boldsymbol{K}_\lambda^{-1/2}(\tilde{\boldsymbol{K}}_N-\boldsymbol{K})K_\lambda^{-1/2}\right\|$ because of (C.2) and the following equation:

$$\mathbb{P}\left(\left\|\boldsymbol{K}_\lambda^{-1/2}(\boldsymbol{K}_N-\boldsymbol{K})K_\lambda^{-1/2}\right\| \geq t\right) \leq \mathbb{P}\left(\left\|\boldsymbol{K}_\lambda^{-1/2}(\tilde{\boldsymbol{K}}_N-\boldsymbol{K})\boldsymbol{K}_\lambda^{-1/2}\right\| \geq t\right) + \mathbb{P}\left(\boldsymbol{K}_N \neq \tilde{\boldsymbol{K}}_N\right). \tag{C.7}$$

Meanwhile, by (C.3), (C.4), and (C.5), we know that

$$\begin{aligned}\left\|\boldsymbol{K}_\lambda^{-1/2}(\tilde{\boldsymbol{K}}_N-\boldsymbol{K})\boldsymbol{K}_\lambda^{-1/2}\right\| \leq\ & \left\|\boldsymbol{K}_\lambda^{-1/2}(\tilde{\boldsymbol{K}}_N-\tilde{\boldsymbol{K}})\boldsymbol{K}_\lambda^{-1/2}\right\| + \left\|\boldsymbol{K}_\lambda^{-1/2}(\tilde{\boldsymbol{K}}-\boldsymbol{K})K_\lambda^{-1/2}\right\| \\ \leq\ & \left\|\boldsymbol{K}_\lambda^{-1/2}\tilde{\boldsymbol{K}}_\lambda^{1/2}\right\|^2\left\|\tilde{\boldsymbol{K}}_\lambda^{-1/2}(\tilde{\boldsymbol{K}}_N-\tilde{\boldsymbol{K}})\tilde{\boldsymbol{K}}_\lambda^{-1/2}\right\| + \left\|\boldsymbol{K}_\lambda^{-1/2}(\tilde{\boldsymbol{K}}-\boldsymbol{K})K_\lambda^{-1/2}\right\| \\ \leq\ & \frac{3}{2}\left\|\tilde{\boldsymbol{K}}_\lambda^{-1/2}(\tilde{\boldsymbol{K}}_N-\tilde{\boldsymbol{K}})\tilde{\boldsymbol{K}}_\lambda^{-1/2}\right\| + \frac{C_0 n}{\lambda_0}\exp\left(-B^2/4\right)\end{aligned} \tag{C.8}$$

Hence, we only need to prove the concentration inequality for $\tilde{\boldsymbol{K}}_\lambda^{-1/2}(\tilde{\boldsymbol{K}}_N-\tilde{\boldsymbol{K}})\tilde{\boldsymbol{K}}_\lambda^{-1/2} = \sum_{i=1}^N \boldsymbol{H}_i - \mathbb{E}\boldsymbol{H}_i$. In terms of the definition of $\tilde{\sigma}$ and (C.6), we know that, almost surely,

$$\|\boldsymbol{H}_i - \mathbb{E}\boldsymbol{H}_i\| \leq 2\|\boldsymbol{H}_i\| \leq \frac{4}{\lambda_0 N}\left\|\tilde{\sigma}(\boldsymbol{w}_i^\top \boldsymbol{X})\right\|^2 \leq \frac{4B^2 n}{\lambda_0 N},$$

where we take expectation with respect to $\boldsymbol{w}_i$. Analogously, applying (C.6), we have

$$\begin{aligned}\boldsymbol{H}_i^2 &= \frac{1}{N^2}\tilde{\boldsymbol{K}}_\lambda^{-1/2}\tilde{\sigma}(\boldsymbol{w}_i^\top \boldsymbol{X})^\top\tilde{\sigma}(\boldsymbol{w}_i^\top \boldsymbol{X})\tilde{\boldsymbol{K}}_\lambda^{-1}\tilde{\sigma}(\boldsymbol{w}_i^\top \boldsymbol{X})^\top\tilde{\sigma}(\boldsymbol{w}_i^\top \boldsymbol{X})\tilde{\boldsymbol{K}}_\lambda^{-1/2} \\ &\preccurlyeq \frac{2\left\|\sigma(\boldsymbol{w}_i^\top \boldsymbol{X})\right\|^2}{\lambda_0 N^2}\tilde{\boldsymbol{K}}_\lambda^{-1/2}\tilde{\sigma}(\boldsymbol{w}_i^\top \boldsymbol{X})^\top\tilde{\sigma}(\boldsymbol{w}_i^\top \boldsymbol{X})\tilde{\boldsymbol{K}}_\lambda^{-1/2} \\ &\preccurlyeq \frac{2B^2 n}{\lambda_0 N^2}\tilde{\boldsymbol{K}}_\lambda^{-1/2}\tilde{\sigma}(\boldsymbol{w}_i^\top \boldsymbol{X})^\top\tilde{\sigma}(\boldsymbol{w}_i^\top \boldsymbol{X})\tilde{\boldsymbol{K}}_\lambda^{-1/2}.\end{aligned}$$

Notice that $\mathbb{E}[\tilde{\sigma}(\boldsymbol{w}_i^\top \boldsymbol{X})^\top\tilde{\sigma}(\boldsymbol{w}_i^\top \boldsymbol{X})] = \tilde{\boldsymbol{K}}$. Hence, $\mathbb{E}[\tilde{\boldsymbol{K}}_\lambda^{-1/2}\tilde{\sigma}(\boldsymbol{w}_i^\top \boldsymbol{X})^\top\tilde{\sigma}(\boldsymbol{w}_i^\top \boldsymbol{X})\tilde{\boldsymbol{K}}_\lambda^{-1/2}] = \frac{1}{1+\lambda}\,\mathrm{Id}$, and

$$\mathbb{E}[(\boldsymbol{H}_i - \mathbb{E}[\boldsymbol{H}_i])^2] \preccurlyeq \mathbb{E}\boldsymbol{H}_i^2 \preccurlyeq \frac{2B^2 n}{\lambda_0 N^2}\,\mathrm{Id}\,.$$

Thus, applying Theorem 5.4.1 of (Vershynin, 2018), we obtain

$$\mathbb{P}\left(\left\|\sum_{i=1}^N \boldsymbol{H}_i - \mathbb{E}\boldsymbol{H}_i\right\| > t\right) \leq 2n\exp\left(-\frac{t^2/2}{v+at/3}\right), \tag{C.9}$$

where $v \leq \frac{2B^2 n}{\lambda_0 N}$ and $a = \frac{4B^2 n}{\lambda_0 N}$. Take $t = B^2\sqrt{n/N}$ and $B = C'\sqrt{\log N}$. Then for $N \geq B^4 n$, by taking constant $C' > 0$ sufficiently large, (C.4) holds and the right hand side of (C.2) is no great than $\frac{1}{2}N^{-2}$. Moreover, (C.9) implies that there exists an absolute constant $C'' > 0$ such that

$$\mathbb{P}\left(\left\|\tilde{\boldsymbol{K}}_\lambda^{-1/2}(\tilde{\boldsymbol{K}}_N-\tilde{\boldsymbol{K}})\tilde{\boldsymbol{K}}_\lambda^{-1/2}\right\| > C''\log(N)\sqrt{\frac{n}{N}}\right) \leq \frac{1}{2}N^{-2}, \tag{C.10}$$

for sufficiently large C'. Here both $C', C'' > 0$ are determined by λ_0. Notice that for all large N, the second term of (C.8) can be also bounded by $C''' \log(N)\sqrt{\frac{n}{N}}$ for some constant $C''' > 0$ only depending on $\|\sigma\|_4$ and λ_0. Combing (C.2), (C.7), (C.8), and (C.10), we can conclude that there exists some large constants $C_1, C_2 > 0$, such that with probability at least $1 - N^{-2}$, when $N/\log^2(N) > C_1 n$, and $n \geq n_0$,

$$\left\| \boldsymbol{K}_\lambda^{-1/2} \left(\boldsymbol{K}_N - \boldsymbol{K} \right) \boldsymbol{K}_\lambda^{-1/2} \right\| \leq C_2 \log(N) \sqrt{\frac{n}{N}},$$

as desired, where C_1, C_2 are constants depending only on $\|\sigma\|_4$ and λ_0. □

C.3 Proof of Theorem 2.7

We first prove the following corollary from Proposition C.1.

Corollary C.2. Following the notations of Proposition C.1, let us denote $t = C_1 \log(N)\sqrt{\frac{n}{N}}$. When $t \in (0,1)$, under the same assumptions as Proposition C.1, with probability at least $1 - N^{-2}$, when $n \geq n_0$, the following holds:

$$\begin{aligned} \left\| (\boldsymbol{K}_N + \lambda \operatorname{Id})^{-1/2} \left(\boldsymbol{K} - \boldsymbol{K}_N \right) (\boldsymbol{K}_N + \lambda \operatorname{Id})^{-1/2} \right\| \leq & t, \\ \left\| \boldsymbol{K}_\lambda^{-1/2} (\boldsymbol{K}_N + \lambda \operatorname{Id})^{1/2} \right\| \leq & \sqrt{1+t}, \\ \left\| \boldsymbol{K}_\lambda^{1/2} (\boldsymbol{K}_N + \lambda \operatorname{Id})^{-1/2} \right\| \leq & (1-t)^{-1/2}, \end{aligned}$$

and the smallest eigenvalue $\lambda_{\min}(\boldsymbol{K}_N) \geq (1-t)\lambda_0$.

Proof. Based on Proposition C.1, under the event in (C.1), we can deduce that

$$\begin{aligned} (\boldsymbol{K}_N - \boldsymbol{K}) &\preccurlyeq t\boldsymbol{K}_\lambda, \\ (\boldsymbol{K} - \boldsymbol{K}_N) &\preccurlyeq t\boldsymbol{K}_\lambda, \\ (\boldsymbol{K} - \boldsymbol{K}_N) &\preccurlyeq \frac{t}{1-t}(\boldsymbol{K}_N + \lambda \operatorname{Id}), \\ (\boldsymbol{K}_N + \lambda \operatorname{Id}) &\preccurlyeq (1+t)\boldsymbol{K}_\lambda, \\ \boldsymbol{K}_\lambda &\preccurlyeq \frac{1}{1-t}(\boldsymbol{K}_N + \lambda \operatorname{Id}), \end{aligned} \tag{C.11}$$

with probability at least $1 - N^{-2}$. These imply the results of the Corollary C.2, where the bound of $\lambda_{\min}(\boldsymbol{K}_N)$ is due to Proposition 2.4 and (C.11). □

Now we are ready to prove Theorem 2.7.

Proof of Theorem 2.7. From the definitions of training errors in (2.10) and (2.11), Proposition 2.4 and Corollary C.2 implies that both $\boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X})$ and $\boldsymbol{K}_N(\boldsymbol{X}, \boldsymbol{X})$ are invertible with probability at least $1 - N^{-2}$ when $t \in (0, 3/4)$. Thus, we have when $t \in (0, 3/4)$, with probability at least $1 - N^{-2}$, when $n \geq n_0$,

$$\begin{aligned} & \left| E_{\text{train}}^{(\mathsf{RF},\lambda)} - E_{\text{train}}^{(\mathsf{K},\lambda)} \right| = \frac{\lambda^2}{n} \left| \operatorname{Tr}[(\boldsymbol{K} + \lambda \operatorname{Id})^{-2} \boldsymbol{y}\boldsymbol{y}^\top] - \operatorname{Tr}[(\boldsymbol{K}_N + \lambda \operatorname{Id})^{-2} \boldsymbol{y}\boldsymbol{y}^\top] \right| \\ = & \frac{\lambda^2}{n} \left| \boldsymbol{y}^\top \left[(\boldsymbol{K} + \lambda \operatorname{Id})^{-2} - (\boldsymbol{K}_N + \lambda \operatorname{Id})^{-2} \right] \boldsymbol{y} \right| \\ \leq & \frac{\lambda^2}{n} \| (\boldsymbol{K} + \lambda \operatorname{Id})^{-2} - (\boldsymbol{K}_N + \lambda \operatorname{Id})^{-2} \| \cdot \|\boldsymbol{y}\|^2 \\ \leq & \frac{\lambda^2 \|\boldsymbol{y}\|^2}{n} \| (\boldsymbol{K} + \lambda \operatorname{Id})^{-1} - (\boldsymbol{K}_N + \lambda \operatorname{Id})^{-1} \| \cdot (\| (\boldsymbol{K} + \lambda \operatorname{Id})^{-1} \| + \| (\boldsymbol{K}_N + \lambda \operatorname{Id})^{-1} \|) \\ \leq & \frac{5\lambda^2 \|\boldsymbol{y}\|^2}{\lambda_0 n} \| (\boldsymbol{K} + \lambda \operatorname{Id})^{-1} - (\boldsymbol{K}_N + \lambda \operatorname{Id})^{-1} \|, \end{aligned} \tag{C.12}$$

where in the last line, we employ the fact that $\left\| (\boldsymbol{K}_N + \lambda \operatorname{Id})^{-1} \right\| \leq 4\lambda_0^{-1}$ and $\left\| (\boldsymbol{K} + \lambda \operatorname{Id})^{-1} \right\| \leq \lambda_0^{-1}$ from Corollary C.2 and Proposition 2.4, respectively.

Take $C_2 = \sqrt{2}C_1$ in Proposition C.1. For any N satisfying $N/\log^2(N) > 2C_1^2 n$, we can make $0 < t < 3/4$, where $t = C_1 \log(N)\sqrt{\frac{n}{N}}$. From this, considering the identity $A^{-1} - B^{-1} = B^{-1}(B-A)A^{-1}$, and applying Proposition C.1 and Corollary C.2, we obtain that

$$\begin{aligned}
&\left\|(\boldsymbol{K}+\lambda\,\mathrm{Id})^{-1} - (\boldsymbol{K}_N+\lambda\,\mathrm{Id})^{-1}\right\| \\
=&\left\|(\boldsymbol{K}_\lambda)^{-1/2}(\boldsymbol{K}_\lambda)^{-1/2}\left(\boldsymbol{K}_N-\boldsymbol{K}\right)(\boldsymbol{K}_\lambda)^{-1/2}(\boldsymbol{K}_\lambda)^{1/2}(\boldsymbol{K}_N+\lambda\,\mathrm{Id})^{-1/2}(\boldsymbol{K}_N+\lambda\,\mathrm{Id})^{-1/2}\right\| \\
\le&\frac{1}{2\lambda_0}\left\|(\boldsymbol{K}_\lambda)^{-1/2}\left(\boldsymbol{K}_N-\boldsymbol{K}\right)(\boldsymbol{K}_\lambda)^{-1/2}\right\|\left\|(\boldsymbol{K}_\lambda)^{1/2}(\boldsymbol{K}_N+\lambda\,\mathrm{Id})^{-1/2}\right\| \\
\le&\frac{t}{2\lambda_0\sqrt{1-t}} \le \frac{t}{\lambda_0}.
\end{aligned} \tag{C.13}$$

Hence, from (C.12), we get

$$\left|E_{\text{train}}^{(\mathsf{RF},\lambda)} - E_{\text{train}}^{(\mathsf{K},\lambda)}\right| \le \frac{5\lambda^2\left\|\boldsymbol{y}\right\|^2 t}{\lambda_0^2 n},$$

which finishes the proof of Theorem 2.7. □

C.4 Proof of Theorem 2.8

We start with the following estimate on the *normalized* trace $\operatorname{tr}\boldsymbol{K}_\lambda^{-1}$.

Lemma C.3. *Under Assumption 2.2, we have* $\operatorname{tr}\boldsymbol{K}_\lambda = \lambda + \|\sigma\|_2^2$ *and when* $n \ge n_0$*,*

$$(\lambda + \|\sigma\|_2^2)^{-1} \le \operatorname{tr}\boldsymbol{K}_\lambda^{-1} \le \lambda_0^{-1}.$$

Proof. By definition of $\boldsymbol{K}$, we know $\operatorname{Tr}[\boldsymbol{K}] = n\mathbb{E}_{\boldsymbol{w}}[\sigma(\boldsymbol{w}^\top\boldsymbol{x})^2] = n\mathbb{E}[\sigma(\xi)^2] = n\left\|\sigma\right\|_2^2$ for $\xi \sim \mathcal{N}(0,1)$. Hence, $\operatorname{Tr}\boldsymbol{K}_\lambda = n\left(\lambda + \|\sigma\|_2^2\right)$. Denote $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_n \ge 0$ as the eigenvalues of $\boldsymbol{K}$. Then, by Cauchy–Schwartz inequality, we have

$$n = \sum_{i=1}^n \frac{1}{\sqrt{\lambda_i+\lambda}}\sqrt{\lambda_i+\lambda} \le \left(\operatorname{Tr}\boldsymbol{K}_\lambda^{-1}\right)^{1/2}\left(\operatorname{Tr}\boldsymbol{K}_\lambda\right)^{1/2}.$$

Therefore, we can get $(\lambda + \|\sigma\|_2^2)^{-1} \le \operatorname{tr}\boldsymbol{K}_\lambda^{-1}$. Meanwhile, based on Proposition 2.4, $\operatorname{tr}\boldsymbol{K}_\lambda^{-1} \le \left\|\boldsymbol{K}_\lambda^{-1}\right\| \le \lambda_{\min}^{-1}(\boldsymbol{K}) \le \lambda_0^{-1}$. Notice that $\lambda_0 = \frac{1}{2}\sigma_{>\ell}^2 \le \|\sigma\|_2^2$. □

Recall that $\boldsymbol{K}_{N,\lambda} = \boldsymbol{K}_N + \lambda\,\mathrm{Id}$. The following lemma for the approximation of $\boldsymbol{K}_\lambda$ with $\boldsymbol{K}_{N,\lambda}$ holds.

Lemma C.4. *Under the assumptions of Proposition C.1, for sufficiently large constant* $C > 0$*, when* $N/\log^2(N) > C(1+\lambda^2)n$*, we have, with probability at least* $1 - N^{-2}$*, when* $n \ge n_0$*,*

$$\left|\operatorname{tr}\boldsymbol{K}_\lambda^{-1} - \operatorname{tr}\boldsymbol{K}_{N,\lambda}^{-1}\right| \le C_0\log(N)\sqrt{\frac{n}{N}}, \tag{C.14}$$

and

$$\frac{1}{2(\lambda+\|\sigma\|_2^2)} \le \operatorname{tr}\boldsymbol{K}_{N,\lambda}^{-1} \le \frac{3}{2\lambda_0},$$

where constants $C, C_0 > 0$ *only depends on* λ_0 *and* $\|\sigma\|_4$*.*

Proof. From (C.13) and Lemma C.3, by taking $t = C_1\log(N)\sqrt{\frac{n}{N}} \in (0,1)$, we have

$$\begin{aligned}
\left|\operatorname{tr}\boldsymbol{K}_\lambda^{-1} - \operatorname{tr}\boldsymbol{K}_{N,\lambda}^{-1}\right| \le \left\|\boldsymbol{K}_\lambda^{-1} - \boldsymbol{K}_{N,\lambda}^{-1}\right\| \le \frac{t}{\lambda_0(1-t)}, \\
(\lambda+\|\sigma\|_2^2)^{-1} - \frac{t}{\lambda_0(1-t)} \le \operatorname{tr}\boldsymbol{K}_{N,\lambda}^{-1} \le \lambda_0^{-1} + \frac{t}{\lambda_0(1-t)}.
\end{aligned} \tag{C.15}$$

Considering sufficiently large constant $C > 0$ with $N/\log^2(N) > C(1+\lambda^2)n$, we can ensure that t is sufficiently small and satisfies $0 \le t \le \min\{1/2, \lambda_0/4(\lambda + \|\sigma\|_2^2)\}$. Then,

$$\frac{t}{\lambda_0(1-t)} \le \frac{1}{2(\lambda + \|\sigma\|_2^2)} \le \frac{1}{2\lambda_0}. \tag{C.16}$$

Hence, taking $C_0 = 2C_1/\lambda_0$, we can conclude (C.14). The second statement follows from (C.15) and (C.16) directly. □

Lemma C.5. *Based on the definitions of LOOCVs of KRR and RFRR in* (2.13)*, we have shortcut formulae* (2.14) *and* (2.15) *for KRR and RFRR respectively: for any $\lambda \ge 0$,*

$$\begin{aligned} CV_n^{(\mathsf{K},\lambda)} &= \frac{1}{n}\boldsymbol{y}^\top \boldsymbol{K}_\lambda^{-1}\boldsymbol{D}^{-2}\boldsymbol{K}_\lambda^{-1}\boldsymbol{y}, \\ CV_n^{(\mathsf{RF},\lambda)} &= \frac{1}{n}\boldsymbol{y}^\top \boldsymbol{K}_{N,\lambda}^{-1}\boldsymbol{D}_N^{-2}\boldsymbol{K}_{N,\lambda}^{-1}\boldsymbol{y}, \end{aligned}$$

where $\boldsymbol{D}$ and $\boldsymbol{D}_N$ are diagonal matrices with diagonals $[\boldsymbol{D}]_{ii} = [\boldsymbol{K}_\lambda^{-1}]_{ii}$ and $[\boldsymbol{D}_N]_{ii} = [\boldsymbol{K}_{N,\lambda}^{-1}]_{ii}$, for $i \in [n]$, respectively. When $n \ge n_0$, we have

$$(\lambda + \|\sigma\|_2^2)^{-1} \le \|\boldsymbol{D}\| \le \lambda_0^{-1}. \tag{C.17}$$

Additionally, for a constant $C > 0$ depending on $\lambda_0, \|\sigma\|_2$, when $N/\log^2(N) > C(1+\lambda^2)n$,

$$\frac{1}{2}(\lambda + \|\sigma\|_2^2)^{-1} \le \|\boldsymbol{D}_N\| \le 2\lambda_0^{-1}, \tag{C.18}$$

$$\left\|\boldsymbol{D}^{-2} - \boldsymbol{D}_N^{-2}\right\| \le C_0(1+\lambda^4)\log(N)\sqrt{\frac{n}{N}}, \tag{C.19}$$

with probability at least $1 - N^{-2}$, for some constant $C_0 > 0$ which only depends on λ_0 and $\|\sigma\|_2$.

Proof. For $i \in [n]$, denote $\boldsymbol{y}^{-i} \in \mathbb{R}^{n-1}$ by the vector $\boldsymbol{y}$ with the i-th entry removed, $\boldsymbol{X}^{-i}$ by the data $\boldsymbol{X}$ with the i-th colmun removed, and $\boldsymbol{K}_{-i,\lambda}$ by the matrix $\boldsymbol{K}_\lambda$ with both the i-th row and column removed. Based on Schur complement and resolvent identities (Benaych-Georges and Knowles, 2017, Lemma 3.5), we have for any $i, j \in [n]$ with $j \ne i$, the (i,j) entry of $\boldsymbol{K}_\lambda^{-1}$ is given by

$$[\boldsymbol{K}_\lambda^{-1}]_{i,j} = -[\boldsymbol{K}_\lambda^{-1}]_{ii}\sum_{k\ne i}[\boldsymbol{K}]_{i,k}[\boldsymbol{K}_{-i,\lambda}^{-1}]_{k,j}. \tag{C.20}$$

Thus, from definition (2.13), we can exploit (C.20) to obtain

$$\begin{aligned} & y_i - \hat{f}_{\lambda,-i}^{(\mathsf{K})}(\boldsymbol{x}_i) = y_i - \boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{X}^{-i})\boldsymbol{K}_{-i,\lambda}^{-1}\boldsymbol{y}^{-i} \\ =\ & y_i + \frac{[\boldsymbol{K}_\lambda^{-1}]_{[i,\ne i]}\boldsymbol{y}^{-i}}{[\boldsymbol{K}_\lambda^{-1}]_{ii}} + \left(\frac{[\boldsymbol{K}_\lambda^{-1}]_{ii}}{[\boldsymbol{K}_\lambda^{-1}]_{ii}}y_i - y_i\right) = \frac{[\boldsymbol{K}_\lambda^{-1}]_{[i,:]}\boldsymbol{y}}{[\boldsymbol{K}_\lambda^{-1}]_{ii}}, \end{aligned}$$

for any $i \in [n]$, where $[\boldsymbol{K}_\lambda^{-1}]_{[i,\ne i]}$ is the i-th row of $\boldsymbol{K}_\lambda^{-1}$ with the i-th entry removed, and $[\boldsymbol{K}_\lambda^{-1}]_{[i,:]}$ is the i-th row of $\boldsymbol{K}_\lambda^{-1}$. Hence, in matrix form, we can get

$$\mathrm{CV}_n^{(\mathsf{K},\lambda)} = \frac{1}{n}\sum_{i=1}^n \frac{\boldsymbol{y}^\top[\boldsymbol{K}_\lambda^{-1}]_{[i,:]}^\top[\boldsymbol{K}_\lambda^{-1}]_{[i,:]}\boldsymbol{y}}{[\boldsymbol{K}_\lambda^{-1}]_{ii}} = \frac{1}{n}\boldsymbol{y}^\top\boldsymbol{K}_\lambda^{-1}\boldsymbol{D}^{-2}\boldsymbol{K}_\lambda^{-1}\boldsymbol{y}.$$

Going through the same procedure, we can verify (2.15) as well.

Secondly, applying Theorem A.4 of (Bai and Silverstein, 2010), we have

$$[\boldsymbol{D}]_{ii} = [\boldsymbol{K}_\lambda^{-1}]_{ii} = \frac{1}{[\boldsymbol{K}_\lambda]_{ii} - \boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{X}^{-i})\boldsymbol{K}_{-i,\lambda}^{-1}\boldsymbol{K}(\boldsymbol{X}^{-i}, \boldsymbol{x}_i)},$$

for any $i \in [n]$. Recall that, in the proof of Lemma C.3, we have shown $[\boldsymbol{K}_\lambda]_{ii} = \lambda + \|\sigma\|_2^2$ for all $i \in [n]$. Therefore, we have

$$(\lambda + \|\sigma\|_2^2)^{-1} \le [\boldsymbol{K}_\lambda^{-1}]_{ii} \le \|\boldsymbol{K}_\lambda^{-1}\| \le \lambda_0^{-1},\ \forall i \in [n].$$

which verifies the result (C.17).

Meanwhile, from the proof of Lemma C.4, for sufficiently large constants C, C_1 depending only on $\lambda_0, \|\sigma\|_2$, with $t = C_1 \log(N)\sqrt{n/N} \le \min\{1/2, \lambda_0/4(\lambda + \|\sigma\|_2^2)\}$ and $N/\log^2(N) > C(1+\lambda^2)n$, we have

$$\|\boldsymbol{D} - \boldsymbol{D}_N\| \le \left\|\boldsymbol{K}_\lambda^{-1} - \boldsymbol{K}_{N,\lambda}^{-1}\right\| \le \frac{t}{\lambda_0(1-t)}, \tag{C.21}$$

with probability at least $1 - N^{-2}$. Therefore, we can verify (C.18) as follows

$$\frac{1}{2}(\lambda + \|\sigma\|_2^2)^{-1} \le (\lambda + \|\sigma\|_2^2)^{-1} - \|\boldsymbol{D} - \boldsymbol{D}_N\| \le \|\boldsymbol{D}_N\| \le \lambda_0^{-1} + \|\boldsymbol{D} - \boldsymbol{D}_N\| \le 2\lambda_0^{-1}.$$

Finally, combining (C.17), (C.18) and (C.21) together, we can obtain that

$$\begin{aligned}\left\|\boldsymbol{D}^{-2} - \boldsymbol{D}_N^{-2}\right\| &\le \|\boldsymbol{D}\|^{-2} \|\boldsymbol{D}_N\|^{-2} (\|\boldsymbol{D}_N\| + \|\boldsymbol{D}\|) \|\boldsymbol{D} - \boldsymbol{D}_N\| \\ &\le 12\frac{(\lambda + \|\sigma\|_2^2)^4}{\lambda_0^3} t \le C_0(1+\lambda^4)\log(N)\sqrt{\frac{n}{N}}.\end{aligned}$$

This finally completes the proof of this lemma. □

Proof of Theorem 2.8. We start with (2.17). Recall (2.10), (2.11), and $\boldsymbol{K}_{N,\lambda} = \boldsymbol{K}_N + \lambda \,\mathrm{Id}$. Using the expression (2.16), we have

$$\begin{aligned}|\mathrm{GCV}_n^{(\mathsf{K},\lambda)} - \mathrm{GCV}_n^{(\mathsf{RF},\lambda)}| \le& \frac{1}{\lambda^2}\left|\left(\left(\operatorname{tr}\boldsymbol{K}_\lambda^{-1}\right)^{-2} - \left(\operatorname{tr}\boldsymbol{K}_{N,\lambda}^{-1}\right)^{-2}\right) E_{\text{train}}^{(\mathsf{K},\lambda)}\right| + \frac{1}{\lambda^2\left(\operatorname{tr}\boldsymbol{K}_\lambda^{-1}\right)^2}\left|E_{\text{train}}^{(\mathsf{K},\lambda)} - E_{\text{train}}^{(\mathsf{RF},\lambda)}\right| \\ \le& \frac{1}{n}\left\|\boldsymbol{K}_\lambda^{-1}\boldsymbol{y}\right\|^2 \left|\left(\operatorname{tr}\boldsymbol{K}_\lambda^{-1}\right)^{-2} - \left(\operatorname{tr}\boldsymbol{K}_{N,\lambda}^{-1}\right)^{-2}\right| \qquad \text{(C.22)} \\ &+ C_2(\lambda + \|\sigma\|_2^2)^2 \frac{\log N \|\boldsymbol{y}\|^2}{\sqrt{nN}}, \qquad \text{(C.23)}\end{aligned}$$

where (C.23) is due to (2.12) and Lemma C.3, and C_2 is a constant depending on $\|\sigma\|_4$ and λ_0.

For the first term (C.22), when $N/\log^2 N \ge C(1+\lambda^2)n$ for a sufficiently large C, together with Lemmas C.3 and C.4, we can show that with probability at least $1 - N^{-2}$,

$$\begin{aligned}&\frac{1}{n}\left\|\boldsymbol{K}_\lambda^{-1}\boldsymbol{y}\right\|^2 \left|\left(\operatorname{tr}\boldsymbol{K}_\lambda^{-1}\right)^{-2} - \left(\operatorname{tr}\boldsymbol{K}_{N,\lambda}^{-1}\right)^{-2}\right| \\ \le\ & \left(\operatorname{tr}\boldsymbol{K}_\lambda^{-1}\right)^{-2}\left(\operatorname{tr}\boldsymbol{K}_{N,\lambda}^{-1}\right)^{-2}\left|\operatorname{tr}(\boldsymbol{K}_\lambda^{-1} - \boldsymbol{K}_{N,\lambda}^{-1})\right| \operatorname{tr}\left(\boldsymbol{K}_\lambda^{-1} + \boldsymbol{K}_{N,\lambda}^{-1}\right)\frac{1}{n}\|\boldsymbol{K}_\lambda^{-2}\|\,\|\boldsymbol{y}\|^2 \\ \le\ & \frac{20(\lambda + \|\sigma\|_2^2)^4}{\lambda_0^3 n}\|\boldsymbol{y}\|^2 C_0\log(N)\sqrt{\frac{n}{N}} \le \frac{C_1(1+\lambda^4)\|\boldsymbol{y}\|^2\log(N)}{\sqrt{nN}},\end{aligned}$$

for some constant C_1 which only relies on $\|\sigma\|_2$ and λ_0. Hence, the bounds of (C.22) and (C.23) imply

$$|\mathrm{GCV}_n^{(\mathsf{K},\lambda)} - \mathrm{GCV}_n^{(\mathsf{RF},\lambda)}| \le \frac{c(1+\lambda^4)\log N\|\boldsymbol{y}\|^2}{\sqrt{nN}}$$

for a constant c depending on λ_0, $\|\sigma\|_2$, and $\|\sigma\|_4$. This proves (2.17) for the GCV concentration.

Now we consider the second result (2.18) for LOOCV. Similar to the analysis of GCV, with the help of the shortcut formulae (2.14) and (2.15), we can get

$$\begin{aligned}\left|\mathrm{CV}_n^{(\mathsf{K},\lambda)} - \mathrm{CV}_n^{(\mathsf{RF},\lambda)}\right| \le\ & \frac{\|\boldsymbol{y}\|^2}{n}\left\|(\boldsymbol{K}_\lambda^{-1} - \boldsymbol{K}_{N,\lambda}^{-1})\boldsymbol{D}^{-2}\boldsymbol{K}_\lambda^{-1}\right\| + \frac{\|\boldsymbol{y}\|^2}{n}\left\|\boldsymbol{K}_{N,\lambda}^{-1}(\boldsymbol{D}^{-2} - \boldsymbol{D}_N^{-2})\boldsymbol{K}_\lambda^{-1}\right\| \\ &+ \frac{\|\boldsymbol{y}\|^2}{n}\left\|\boldsymbol{K}_{N,\lambda}^{-1}\boldsymbol{D}_N^{-2}(\boldsymbol{K}_\lambda^{-1} - \boldsymbol{K}_{N,\lambda}^{-1})\right\| \le \frac{c(1+\lambda^4)\log N\|\boldsymbol{y}\|^2}{\sqrt{nN}}, \qquad \text{(C.24)}\end{aligned}$$

with probability at least $1 - N^{-2}$. Here, we exploit (C.17), (C.18) and (C.19) in Lemma C.5. This verifies the second result (2.18) for LOOCV. □

C.5 Proof of Theorem 2.11

For simplicity, we denote $\boldsymbol{K}_{N,\lambda} = (\boldsymbol{K}_N + \lambda \mathrm{Id})$, $\boldsymbol{K}_\lambda = (\boldsymbol{K} + \lambda \mathrm{Id})$, $\boldsymbol{K}_{m,N} := \boldsymbol{K}_N(\boldsymbol{X}, \boldsymbol{x}) \in \mathbb{R}^n$ and $\boldsymbol{K}_m := \boldsymbol{K}(\boldsymbol{X}, \boldsymbol{x}) \in \mathbb{R}^n$. Define

$$\boldsymbol{K}_N^{(2)} := \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{K}_N(\boldsymbol{X}, \boldsymbol{x})\boldsymbol{K}_N(\boldsymbol{x}, \boldsymbol{X})], \;\; \boldsymbol{K}^{(2)} := \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{K}(\boldsymbol{X}, \boldsymbol{x})\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{X})],$$

where we take expectation with respect to test data $\boldsymbol{x}$ defined in Assumption 2.10. Recalling Assumption 2.9, we have

$$f^*(\boldsymbol{x}) = \tau(\langle \boldsymbol{\beta}, \boldsymbol{x} \rangle), \quad \boldsymbol{y} = \tau(\boldsymbol{X}^\top \boldsymbol{\beta}) + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\beta} \sim \mathcal{N}(0, \mathrm{Id})$. Denote the kernel given by τ as $\boldsymbol{\Psi} := \mathbb{E}_{\boldsymbol{\beta}}[\tau(\boldsymbol{X}^\top \boldsymbol{\beta})\tau(\boldsymbol{\beta}^\top \boldsymbol{X})]$ and the vector by $\boldsymbol{u} := \mathbb{E}_{\boldsymbol{\beta}}[\tau(\boldsymbol{X}^\top \boldsymbol{\beta}) f^*(\boldsymbol{x})] \in \mathbb{R}^n$.

We begin with the following lemmas about the bounds and concentrations with respect to $\boldsymbol{K}_N^{(2)}$, $\boldsymbol{K}^{(2)}$, $\boldsymbol{K}_{N,\lambda}$, $\boldsymbol{K}_{m,N}$, and $\boldsymbol{K}_m$.

Lemma C.6. *There exist some constant $C > 0$ depending on $\lambda_0, \|\sigma\|_2^2$ such that, with probability at least $1 - N^{-1}$, when $n \geq \max\{n_0, n_1\}$,*

$$\boldsymbol{K}_{m,N}^\top \boldsymbol{K}_{N,\lambda}^{-1} \boldsymbol{K}_{m,N} < \|\sigma\|_2^2 + \|\sigma\|_4^2 + \lambda,$$

when $N/\log^2(N) > C(1+\lambda^2)n$. Moreover, we have

$$\boldsymbol{K}_m^\top \boldsymbol{K}_\lambda^{-1} \boldsymbol{K}_m < \|\sigma\|_2^2 + \lambda.$$

Proof. Consider an enlarged block matrix $\tilde{\boldsymbol{K}} \in \mathbb{R}^{(n+1)\times(n+1)}$ defined by

$$\tilde{\boldsymbol{K}} := \begin{pmatrix} \boldsymbol{K} & \boldsymbol{K}_m \\ \boldsymbol{K}_m^\top & \boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}) \end{pmatrix}, \tag{C.25}$$

where $\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}) = \mathbb{E}[\sigma(\boldsymbol{w}^\top \boldsymbol{x})(\sigma(\boldsymbol{w}^\top \boldsymbol{x})] = \|\sigma\|_2^2$. Let $\tilde{\boldsymbol{K}}_\lambda := \tilde{\boldsymbol{K}} + \lambda \mathrm{Id}$. Analogously to (2.4), let us define

$$\tilde{\boldsymbol{K}}_\ell := \sum_{k=0}^{\ell} \zeta_k^2(\sigma)(\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}})^{\odot k} + \sigma_{>\ell}^2 \mathrm{Id} \in \mathbb{R}^{(n+1)\times(n+1)}, \tag{C.26}$$

where $\tilde{\boldsymbol{X}} = [\boldsymbol{X}, \boldsymbol{x}] \in \mathbb{R}^{d\times(n+1)}$ is the concatenation of training and test data points. By Assumption 2.10, analogously to the proof of Proposition 2.4, we have

$$\begin{aligned} \left\| \tilde{\boldsymbol{K}}_\ell - \tilde{\boldsymbol{K}} \right\|^2 &\leq \left\| \tilde{\boldsymbol{K}}_\ell - \tilde{\boldsymbol{K}} \right\|_F^2 \\ &\leq 2\|\sigma\|_4^4 \left\| \left(\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}} \right)^{\odot \ell+1} - \mathrm{Id} \right\|_F^2 \\ &= 2\|\sigma\|_4^4 \left\| \left(\boldsymbol{X}^\top \boldsymbol{X} \right)^{\odot \ell+1} - \mathrm{Id} \right\|_F^2 + 4\|\sigma\|_4^4 \left\| (\boldsymbol{X}^\top \boldsymbol{x})^{\odot(\ell+1)} \right\|_2^2 \leq \frac{3}{8}\lambda_0^2. \end{aligned} \tag{C.27}$$

Since $\lambda_{\min}(\tilde{\boldsymbol{K}}) \geq \lambda_0 > 0$, we have $\lambda_{\min}(\tilde{\boldsymbol{K}}_\lambda) \geq \frac{1}{4}\lambda_0$ and $\tilde{\boldsymbol{K}}_\lambda$ is positive definite for any $\lambda \geq 0$. By Theorem 7.7.7 of Horn and Johnson (2012), since both $\boldsymbol{K}_\lambda$ and $\tilde{\boldsymbol{K}}_\lambda$ are positive definite, the Schur complement of $\tilde{\boldsymbol{K}}_\lambda$ given by $\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}) + \lambda - \boldsymbol{K}_m^\top \boldsymbol{K}_\lambda^{-1} \boldsymbol{K}_m$ is positive, which concludes our second result in this lemma.

Similarly, consider the block matrix $\tilde{\boldsymbol{K}}_N \in \mathbb{R}^{(n+1)\times(n+1)}$ defined by

$$\tilde{\boldsymbol{K}}_N := \boldsymbol{K}_N(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{X}}) = \begin{pmatrix} \boldsymbol{K}_N & \boldsymbol{K}_{m,N} \\ \boldsymbol{K}_{m,N}^\top & \boldsymbol{K}_N(\boldsymbol{x}, \boldsymbol{x}) \end{pmatrix}.$$

Let $\tilde{\boldsymbol{K}}_{N,\lambda} := \lambda \mathrm{Id} + \tilde{\boldsymbol{K}}_N$. Combing Assumption 2.10 and (C.27), we can easily ensure Proposition C.1 and Corollary C.2 still hold for $\tilde{\boldsymbol{K}}_\lambda$ and $\tilde{\boldsymbol{K}}_{N,\lambda}$. Therefore, with probability at least $1 - N^{-2}$, $\lambda_{\min}(\tilde{\boldsymbol{K}}_{N,\lambda}) \geq \frac{1}{2}\lambda_0$, for sufficiently large constant $C > 0$ with $N/\log^2(N) > C(1+\lambda^2)n$, which implies that $\tilde{\boldsymbol{K}}_{N,\lambda}$ is positive definite with probability $1 - N^{-2}$.

Again, from Theorem 7.7.7 of Horn and Johnson (2012), we can get $\boldsymbol{K}_N(\boldsymbol{x},\boldsymbol{x}) + \lambda - \boldsymbol{K}_{m,N}^\top \boldsymbol{K}_{N,\lambda}^{-1} \boldsymbol{K}_{m,N} > 0$ with probability at least $1 - N^{-2}$. Thus,

$$0 \le \boldsymbol{K}_{m,N}^\top \boldsymbol{K}_{N,\lambda}^{-1} \boldsymbol{K}_{m,N} < \boldsymbol{K}_N(\boldsymbol{x},\boldsymbol{x}) + \lambda.$$

Notice that $\boldsymbol{K}_N(\boldsymbol{x},\boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^N \sigma(\boldsymbol{w}_i^\top \boldsymbol{x})^2$, $\mathbb{E}[\boldsymbol{K}_N(\boldsymbol{x},\boldsymbol{x})] = \|\sigma\|_2^2$, and

$$\mathbb{E}\left(\boldsymbol{K}_N(\boldsymbol{x},\boldsymbol{x}) - \|\sigma\|_2^2\right)^2 = \frac{1}{N}\mathrm{Var}(\sigma(\boldsymbol{w}^\top \boldsymbol{x})^2) = \frac{\|\sigma\|_4^4 - \|\sigma\|_2^4}{N} \le \frac{\|\sigma\|_4^4}{N},$$

where $\boldsymbol{w} \sim \mathcal{N}(0, \mathrm{Id})$. Therefore, by Markov inequality, we conclude that, with probability at least $1 - 1/N$, $\boldsymbol{K}_N(\boldsymbol{x},\boldsymbol{x}) \le \|\sigma\|_2^2 + \|\sigma\|_4^2$. Therefore, with probability at least $1 - N^{-1}$, we have $\boldsymbol{K}_{m,N}^\top \boldsymbol{K}_{N,\lambda}^{-1} \boldsymbol{K}_{m,N} < \|\sigma\|_2^2 + \|\sigma\|_4^2 + \lambda$. □

Lemma C.7. *Suppose that, for any $0 \le k \le \ell$, if $\zeta_k(\sigma) \neq 0$, then $\zeta_k(\tau) \neq 0$. Then, there exists a universal constant $C > 0$ only depending on σ and τ such that for any $\lambda \ge 0$ and $n \ge n_0$,*

$$\left\| \boldsymbol{K}_\lambda^{-1/2} \boldsymbol{\Psi} \boldsymbol{K}_\lambda^{-1/2} \right\| \le C,$$

Proof. Analogously to (2.4), we define a truncation version of the kernel $\boldsymbol{\Psi}$ by

$$\boldsymbol{\Psi}_\ell := \sum_{k=0}^{\ell} \zeta_k^2(\tau) \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{\odot k} + \tau_{>\ell}^2 \,\mathrm{Id}\,. \tag{C.28}$$

Define $\boldsymbol{K}_{\ell,\lambda} := \boldsymbol{K}_\ell + \lambda \,\mathrm{Id}$. By the assumption and definition of $\boldsymbol{K}_\ell$, there exists some constant $C > 0$ such that $\boldsymbol{\Psi}_\ell \preccurlyeq C\boldsymbol{K}_{\ell,\lambda}$ for any $\lambda \ge 0$. Here this constant C only relies on the Hermite coefficients $\zeta_k(\tau)$, λ_0 and $\zeta_k(\sigma)$ for $0 \le k \le \ell$. Next, applying Proposition 2.4 for nonlinear function τ, we have

$$\|\boldsymbol{\Psi} - \boldsymbol{\Psi}_\ell\| \le \sqrt{2}\|\tau\|_4^2 \left\| \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{\odot \ell+1} - \mathrm{Id} \right\|_F \le \frac{\sqrt{2}\sigma_{>\ell}^2 \|\tau\|_4^2}{4\|\sigma\|_4^2}. \tag{C.29}$$

Proposition 2.4 also indicates that $\|\boldsymbol{K} - \boldsymbol{K}_\ell\| \le \frac{1}{2}\lambda_0$. This implies that

$$\boldsymbol{K}_\lambda^{-1/2} \boldsymbol{K}_{\ell,\lambda} \boldsymbol{K}_\lambda^{-1/2} \preccurlyeq \frac{3}{2}\,\mathrm{Id},$$

for any $\lambda \ge 0$. Then, for any $\lambda \ge 0$, we can estimate its contribution by

$$\begin{aligned}
\left\| \boldsymbol{K}_\lambda^{-1/2} \boldsymbol{\Psi} \boldsymbol{K}_\lambda^{-1/2} \right\| \le\ & \left\| \boldsymbol{K}_\lambda^{-1/2} (\boldsymbol{\Psi} - \boldsymbol{\Psi}_\ell) \boldsymbol{K}_\lambda^{-1/2} \right\| + \left\| \boldsymbol{K}_\lambda^{-1/2} \boldsymbol{\Psi}_\ell \boldsymbol{K}_\lambda^{-1/2} \right\| \\
\le\ & \lambda_0^{-1} \|\boldsymbol{\Psi} - \boldsymbol{\Psi}_\ell\| + \left\| \boldsymbol{K}_\lambda^{-1/2} \boldsymbol{K}_{\ell,\lambda}^{1/2} \boldsymbol{K}_{\ell,\lambda}^{-1/2} \boldsymbol{\Psi}_\ell \boldsymbol{K}_{\ell,\lambda}^{-1/2} \boldsymbol{K}_{\ell,\lambda}^{1/2} \boldsymbol{K}_\lambda^{-1/2} \right\| \\
\le\ & \frac{\sqrt{2}\sigma_{>\ell}^2 \|\tau\|_4^2}{4\lambda_0 \|\sigma\|_4^2} + \left\| \boldsymbol{K}_{\ell,\lambda}^{1/2} \boldsymbol{K}_\lambda^{-1/2} \right\|^2 \left\| \boldsymbol{K}_{\ell,\lambda}^{-1/2} \boldsymbol{\Psi}_\ell \boldsymbol{K}_{\ell,\lambda}^{-1/2} \right\| \\
\le\ & \frac{\sqrt{2}\sigma_{>\ell}^2 \|\tau\|_4^2}{4\lambda_0 \|\sigma\|_4^2} + \frac{3}{2}C.
\end{aligned}$$

Therefore, there is a constant depending only on σ, τ as the upper bound for $\left\| \boldsymbol{K}_\lambda^{-1/2} \boldsymbol{\Psi} \boldsymbol{K}_\lambda^{-1/2} \right\|$. This completes the proof of this lemma. □

Lemma C.8. *There exists a constant $C > 0$ depending only on σ, τ such that for any $\lambda \ge 0$ and $n \ge \max\{n_0, n_1\}$,*

$$\left\| \boldsymbol{K}_\lambda^{-\frac{1}{2}} \boldsymbol{u} \right\| = \left\| \mathbb{E}_{\boldsymbol{\beta}}[\tau(\boldsymbol{\beta}^\top \boldsymbol{x})\tau(\boldsymbol{\beta}^\top \boldsymbol{X})] \boldsymbol{K}_\lambda^{-\frac{1}{2}} \right\| \le C(1 + \lambda^{1/2}),$$

Proof. Denote $\boldsymbol{K}_{\tau,m} := \mathbb{E}_{\boldsymbol{\beta}}[\tau(\boldsymbol{\beta}^\top \boldsymbol{x})\tau(\boldsymbol{\beta}^\top \boldsymbol{X})]^\top$ and $\boldsymbol{\Psi}_\lambda := \boldsymbol{\Psi} + \lambda \,\mathrm{Id}$. Analogously to (C.25) and (C.26), we can consider

$$\tilde{\boldsymbol{\Psi}}_\lambda = \mathbb{E}_{\boldsymbol{\beta}}[\tau(\boldsymbol{\beta}^\top \tilde{\boldsymbol{X}})^\top \tau(\boldsymbol{\beta}^\top \tilde{\boldsymbol{X}})] + \lambda \,\mathrm{Id} = \begin{pmatrix} \boldsymbol{\Psi}_\lambda & \boldsymbol{K}_{\tau,m} \\ \boldsymbol{K}_{\tau,m}^\top & \mathbb{E}_{\boldsymbol{\beta}}[\tau(\boldsymbol{\beta}^\top \boldsymbol{x})^2] + \lambda \end{pmatrix}$$

where $\tilde{\boldsymbol{X}} = [\boldsymbol{X}, \boldsymbol{x}]$. For any $\lambda \geq 0$, both $\tilde{\boldsymbol{\Psi}}$ and $\boldsymbol{\Psi}$ are positive definite because of (C.29) and (C.28). Following the proof of Lemma C.6, we can similarly derive that the Schur complement $\mathbb{E}_{\boldsymbol{\beta}}[\tau(\boldsymbol{\beta}^\top \boldsymbol{x})^2] + \lambda - \boldsymbol{K}_{\tau,m}^\top \boldsymbol{\Psi}_\lambda^{-1} \boldsymbol{K}_{\tau,m}$ is positive, where $\mathbb{E}_{\boldsymbol{\beta}}[\tau(\boldsymbol{\beta}^\top \boldsymbol{x})^2] = \|\tau\|_2^2$. Therefore, we have

$$
\begin{aligned}
& \left\| \mathbb{E}_{\boldsymbol{\beta}}[\tau(\boldsymbol{\beta}^\top \boldsymbol{x}) \tau(\boldsymbol{\beta}^\top \boldsymbol{X})] \boldsymbol{K}_\lambda^{-\frac{1}{2}} \right\|^2 = \boldsymbol{K}_{\tau,m}^\top \boldsymbol{K}_\lambda^{-1} \boldsymbol{K}_{\tau,m} \\
= \; & \boldsymbol{K}_{\tau,m}^\top \boldsymbol{\Psi}_\lambda^{-\frac{1}{2}} \boldsymbol{\Psi}_\lambda^{\frac{1}{2}} \boldsymbol{K}_\lambda^{-1} \boldsymbol{\Psi}_\lambda^{\frac{1}{2}} \boldsymbol{\Psi}_\lambda^{-\frac{1}{2}} \boldsymbol{K}_{\tau,m} \\
\leq \; & \boldsymbol{K}_{\tau,m}^\top \boldsymbol{\Psi}_\lambda^{-1} \boldsymbol{K}_{\tau,m} \cdot \left\| \boldsymbol{\Psi}_\lambda^{\frac{1}{2}} \boldsymbol{K}_\lambda^{-1} \boldsymbol{\Psi}_\lambda^{\frac{1}{2}} \right\| \leq (\lambda + \|\tau\|_2^2) \left\| \boldsymbol{\Psi}_\lambda^{\frac{1}{2}} \boldsymbol{K}_\lambda^{-1} \boldsymbol{\Psi}_\lambda^{\frac{1}{2}} \right\|.
\end{aligned}
$$

Additionally, following the same proof of Lemma C.7, we can also obtain $\left\| \boldsymbol{\Psi}_\lambda^{\frac{1}{2}} \boldsymbol{K}_\lambda^{-1} \boldsymbol{\Psi}_\lambda^{\frac{1}{2}} \right\| \leq C$, for some constant $C > 0$ which only depends on σ, τ. This concludes the proof. □

The following lemma is analogous to Lemma 5 in Montanari and Zhong (2022), which addresses the concentrations of $\boldsymbol{K}_N^{(2)}$ and $f^*(\boldsymbol{x}) \boldsymbol{K}_N(\boldsymbol{X}, \boldsymbol{x})$ respectively.

Lemma C.9. *Suppose that the assumptions of Theorem 2.11 hold. For any $\lambda \geq 0$, define*

$$
\begin{aligned}
\delta_1 :=& \mathbb{E}_{\boldsymbol{\beta},\varepsilon} \left[\boldsymbol{y}^\top \boldsymbol{K}_\lambda^{-1} \left(\boldsymbol{K}_N^{(2)} - \boldsymbol{K}^{(2)} \right) \boldsymbol{K}_\lambda^{-1} \boldsymbol{y} \right], \\
\delta_2 :=& \mathbb{E}_{\boldsymbol{\beta},\boldsymbol{x}} \left[f^*(\boldsymbol{x}) \left(\boldsymbol{K}_N(\boldsymbol{x}, \boldsymbol{X}) - \boldsymbol{K}(\boldsymbol{x}, \boldsymbol{X}) \right) \boldsymbol{K}_\lambda^{-1} f^*(\boldsymbol{X}) \right].
\end{aligned}
$$

Then, for sufficiently large n, with probability at least $1 - 1/\log^2(N)$, when $n \geq \max\{n_0, n_1\}$, there exists some constant $C > 0$ depending only on σ, τ such that

$$
|\delta_1| \leq C(1+\lambda) \log(N) \sqrt{\frac{n}{N}}, \tag{C.30}
$$

$$
|\delta_2| \leq C(1+\lambda) \log(N) \sqrt{\frac{n}{N}}. \tag{C.31}
$$

Proof. Let $\boldsymbol{v} := \boldsymbol{K}_\lambda^{-1} \boldsymbol{y}$ and $\tilde{g}(\boldsymbol{x}) := \boldsymbol{K}_m^\top \boldsymbol{v}$. Notice that $\delta_1 = \delta_{1,1} + \delta_{1,2}$, where

$$
\begin{aligned}
\delta_{1,1} :=& \; \mathbb{E}_{\boldsymbol{\beta},\varepsilon}[\boldsymbol{v}^\top \mathbb{E}_{\boldsymbol{x}} \left[(\boldsymbol{K}_{m,N} - \boldsymbol{K}_m)(\boldsymbol{K}_{m,N} - \boldsymbol{K}_m)^\top \right] \boldsymbol{v}], \\
\delta_{1,2} :=& \; 2\mathbb{E}_{\boldsymbol{\beta},\varepsilon}[\boldsymbol{v}^\top \mathbb{E}_{\boldsymbol{x}} \left[(\boldsymbol{K}_{m,N} - \boldsymbol{K}_m) \boldsymbol{K}_m^\top \right] \boldsymbol{v}] = 2\mathbb{E}_{\boldsymbol{\beta},\varepsilon,\boldsymbol{x}}[\boldsymbol{v}^\top (\boldsymbol{K}_{m,N} - \boldsymbol{K}_m) \tilde{g}(\boldsymbol{x})],
\end{aligned}
$$

Taking expectation with respect to $\boldsymbol{W}$, we can obtain

$$
\begin{aligned}
0 \leq \mathbb{E}[\delta_{1,1}] =& \; \frac{1}{N^2} \sum_{i,j=1}^N \mathbb{E}_{\boldsymbol{\beta},\varepsilon} \left[\boldsymbol{v}^\top \mathbb{E}_{\boldsymbol{W},\boldsymbol{x}} \left[\left(\sigma(\boldsymbol{w}_i^\top \boldsymbol{x}) \sigma(\boldsymbol{w}_i^\top \boldsymbol{X})^\top - \boldsymbol{K}_m \right) \left(\sigma(\boldsymbol{w}_j^\top \boldsymbol{x}) \sigma(\boldsymbol{w}_j^\top \boldsymbol{X}) - \boldsymbol{K}_m^\top \right) \right] \boldsymbol{v} \right] \\
=& \; \frac{1}{N^2} \sum_{i=1}^N \mathbb{E}_{\boldsymbol{\beta},\varepsilon} \left[\boldsymbol{v}^\top \mathbb{E}_{\boldsymbol{W},\boldsymbol{x}} \left[\left(\sigma(\boldsymbol{w}_i^\top \boldsymbol{x}) \sigma(\boldsymbol{w}_i^\top \boldsymbol{X})^\top - \boldsymbol{K}_m \right) \left(\sigma(\boldsymbol{w}_i^\top \boldsymbol{x}) \sigma(\boldsymbol{w}_i^\top \boldsymbol{X}) - \boldsymbol{K}_m^\top \right) \right] \boldsymbol{v} \right] \\
=& \; \frac{1}{N^2} \sum_{i=1}^N \mathbb{E}_{\boldsymbol{\beta},\varepsilon} \left[\boldsymbol{v}^\top \mathbb{E}_{\boldsymbol{W},\boldsymbol{x}} \left[\sigma(\boldsymbol{w}_i^\top \boldsymbol{x})^2 \sigma(\boldsymbol{w}_i^\top \boldsymbol{X})^\top \sigma(\boldsymbol{w}_i^\top \boldsymbol{X}) - \boldsymbol{K}_m \boldsymbol{K}_m^\top \right] \boldsymbol{v} \right] \\
\leq & \; \frac{1}{N} \mathbb{E}_{\boldsymbol{\beta},\varepsilon} \left[\boldsymbol{v}^\top \mathbb{E}_{\boldsymbol{w},\boldsymbol{x}} \left[\sigma(\boldsymbol{w}^\top \boldsymbol{x})^2 \sigma(\boldsymbol{w}^\top \boldsymbol{X})^\top \sigma(\boldsymbol{w}^\top \boldsymbol{X}) \right] \boldsymbol{v} \right],
\end{aligned} \tag{C.32}
$$

where in the last line we apply the fact $\boldsymbol{K}_m = \mathbb{E}_{\boldsymbol{w}_i}[\sigma(\boldsymbol{w}_i^\top \boldsymbol{x}) \sigma(\boldsymbol{w}_i^\top \boldsymbol{X})^\top] \in \mathbb{R}^n$ for any i-th row of $\boldsymbol{W}$.

Furthermore, by applying Lemma C.7 and Cauchy–Schwartz inequality, we have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\beta},\varepsilon}\left[\boldsymbol{v}^\top \mathbb{E}_{\boldsymbol{w},\boldsymbol{x}}\left[\sigma(\boldsymbol{w}^\top\boldsymbol{x})^2\sigma(\boldsymbol{w}^\top\boldsymbol{X})^\top\sigma(\boldsymbol{w}^\top\boldsymbol{X})\right]\boldsymbol{v}\right]\\
=\;& \mathrm{Tr}\left(\boldsymbol{K}_\lambda^{-1}\mathbb{E}_{\boldsymbol{w},\boldsymbol{x}}\left[\sigma(\boldsymbol{w}^\top\boldsymbol{x})^2\sigma(\boldsymbol{w}^\top\boldsymbol{X})^\top\sigma(\boldsymbol{w}^\top\boldsymbol{X})\right]\boldsymbol{K}_\lambda^{-1}\left(\boldsymbol{\Psi}+\sigma_\varepsilon^2\,\mathrm{Id}\right)\right)\\
=\;& \mathbb{E}_{\boldsymbol{w},\boldsymbol{x}}\left[\sigma(\boldsymbol{w}^\top\boldsymbol{x})^2\sigma(\boldsymbol{w}^\top\boldsymbol{X})\boldsymbol{K}_\lambda^{-1}\left(\boldsymbol{\Psi}+\sigma_\varepsilon^2\,\mathrm{Id}\right)\boldsymbol{K}_\lambda^{-1}\sigma(\boldsymbol{w}^\top\boldsymbol{X})^\top\right]\\
\le\;& \|\sigma\|_4^2\,\mathbb{E}_{\boldsymbol{w},\boldsymbol{x}}\left[\left\|\sigma(\boldsymbol{w}^\top\boldsymbol{X})\right\|^4\left\|\boldsymbol{K}_\lambda^{-1}\left(\boldsymbol{\Psi}+\sigma_\varepsilon^2\,\mathrm{Id}\right)\boldsymbol{K}_\lambda^{-1}\right\|^2\right]^{\frac{1}{2}}\\
\le\;& \|\sigma\|_4^2\cdot\frac{1}{\lambda_0}\left(C+\frac{\sigma_\varepsilon^2}{\lambda_0}\right)\mathbb{E}_{\boldsymbol{w}}\left[\left\|\sigma(\boldsymbol{w}^\top\boldsymbol{X})\right\|^4\right]^{\frac{1}{2}}\\
=\;& \|\sigma\|_4^2\cdot\frac{1}{\lambda_0}\left(C+\frac{\sigma_\varepsilon^2}{\lambda_0}\right)\mathbb{E}_{\boldsymbol{w}}\left[\left(\sum_{i=1}^n\sigma(\boldsymbol{w}^\top\boldsymbol{x}_i)^2\right)^2\right]^{\frac{1}{2}}\\
\le\;& \|\sigma\|_4^4\cdot\frac{1}{\lambda_0}\left(C+\frac{\sigma_\varepsilon^2}{\lambda_0}\right)\cdot n
\end{aligned}\tag{C.33}$$

where $\boldsymbol{w}\sim\mathcal{N}(0,\mathrm{Id})$ is independent of $\boldsymbol{x}$ and $\|\sigma\|_4^4=\mathbb{E}_{\boldsymbol{w},\boldsymbol{x}}[\sigma(\boldsymbol{w}^\top\boldsymbol{x})^4]$. Therefore, combining (C.32) and (C.33), we can conclude that

$$\mathbb{E}[|\delta_{1,1}|]\le C_{1,1}\frac{n}{N},$$

for some constant $C_{1,1}>0$ which only relies on σ and σ_ε. Then, Markov inequality deduces that for sufficiently large n,

$$\mathbb{P}\left(|\delta_{1,1}|>4C_{1,1}\log^2(N)\frac{n}{N}\right)\le\frac{1}{4\log^2(N)}.\tag{C.34}$$

Next, we consider δ_2. Let $\boldsymbol{z}_1,\boldsymbol{z}_2$ be two i.i.d. copies of $\boldsymbol{x}$, and $\boldsymbol{\beta}_1,\boldsymbol{\beta}_2$ be two i.i.d. copies of $\boldsymbol{\beta}$. Let $\boldsymbol{u}_i:=\boldsymbol{K}_\lambda^{-1}\tau(\boldsymbol{\beta}_i^\top\boldsymbol{X})^\top$ and $g_i(\boldsymbol{x}):=\tau(\boldsymbol{\beta}_i^\top\boldsymbol{x})$ for $i=1,2$. Notice that $\mathbb{E}_{\boldsymbol{w}_i}[\sigma(\boldsymbol{w}_i^\top\boldsymbol{z}_1)\sigma(\boldsymbol{X}^\top\boldsymbol{w}_i)]=\boldsymbol{K}(\boldsymbol{X},\boldsymbol{z}_1)$. Then, taking expectation with respect to $\boldsymbol{W}$, we can obtain

$$\begin{aligned}
&\mathbb{E}[\delta_2^2]=\mathbb{E}_{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2}\left[\boldsymbol{u}_1^\top\mathbb{E}_{\boldsymbol{W},\boldsymbol{z}_1,\boldsymbol{z}_2}[(\boldsymbol{K}_N(\boldsymbol{X},\boldsymbol{z}_1)-\boldsymbol{K}(\boldsymbol{X},\boldsymbol{z}_1))\,g_1(\boldsymbol{z}_1)g_2(\boldsymbol{z}_2)\,(\boldsymbol{K}_N(\boldsymbol{z}_2,\boldsymbol{X})-\boldsymbol{K}(\boldsymbol{z}_2,\boldsymbol{X}))]\boldsymbol{u}_2\right]\\
=\;&\frac{1}{N^2}\sum_{i,j=1}^N\mathbb{E}\left[\boldsymbol{u}_1^\top\left(\sigma(\boldsymbol{w}_i^\top\boldsymbol{z}_1)\sigma(\boldsymbol{X}^\top\boldsymbol{w}_i)-\boldsymbol{K}(\boldsymbol{X},\boldsymbol{z}_1)\right)g_1(\boldsymbol{z}_1)g_2(\boldsymbol{z}_2)\left(\sigma(\boldsymbol{w}_j^\top\boldsymbol{z}_2)\sigma(\boldsymbol{w}_j^\top\boldsymbol{X})-\boldsymbol{K}(\boldsymbol{z}_2,\boldsymbol{X})\right)\boldsymbol{u}_2\right]\\
=\;&\frac{1}{N^2}\sum_{i=1}^N\mathbb{E}\left[g_1(\boldsymbol{z}_1)g_2(\boldsymbol{z}_2)\sigma(\boldsymbol{w}_i^\top\boldsymbol{z}_1)\sigma(\boldsymbol{w}_i^\top\boldsymbol{z}_2)\boldsymbol{u}_1^\top\left(\sigma(\boldsymbol{X}^\top\boldsymbol{w}_i)\sigma(\boldsymbol{w}_i^\top\boldsymbol{X})-\boldsymbol{K}(\boldsymbol{X},\boldsymbol{z}_1)\boldsymbol{K}(\boldsymbol{z}_2,\boldsymbol{X})\right)\boldsymbol{u}_2\right]\\
\le\;&\frac{1}{N}\mathbb{E}_{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\boldsymbol{w},\boldsymbol{z}_1,\boldsymbol{z}_2}\left[g_1(\boldsymbol{z}_1)g_2(\boldsymbol{z}_2)\sigma(\boldsymbol{w}^\top\boldsymbol{z}_1)\sigma(\boldsymbol{w}^\top\boldsymbol{z}_2)\cdot\boldsymbol{u}_1^\top\sigma(\boldsymbol{X}^\top\boldsymbol{w})\sigma(\boldsymbol{w}^\top\boldsymbol{X})\boldsymbol{u}_2\right],
\end{aligned}$$

where in the last line, we apply the following bound:

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\boldsymbol{z}_1,\boldsymbol{z}_2}\left[g_1(\boldsymbol{z}_1)g_2(\boldsymbol{z}_2)\sigma(\boldsymbol{w}_i^\top\boldsymbol{z}_1)\sigma(\boldsymbol{w}_i^\top\boldsymbol{z}_2)\boldsymbol{u}_1^\top\boldsymbol{K}(\boldsymbol{X},\boldsymbol{z}_1)\boldsymbol{K}(\boldsymbol{z}_2,\boldsymbol{X})\boldsymbol{u}_2\right]\\
=&\left(\mathbb{E}_{\boldsymbol{\beta},\boldsymbol{x}}\left[\tau(\boldsymbol{\beta}^\top\boldsymbol{x})\sigma(\boldsymbol{w}_i^\top\boldsymbol{x})\tau(\boldsymbol{\beta}^\top\boldsymbol{X})\boldsymbol{K}_\lambda^{-1}\boldsymbol{K}(\boldsymbol{X},\boldsymbol{x})\right]\right)^2\ge 0.
\end{aligned}$$

Let $\boldsymbol{v}_i:=\boldsymbol{K}_\lambda^{-1/2}\mathbb{E}_{\boldsymbol{\beta}_i}[\tau(\boldsymbol{\beta}_i^\top\boldsymbol{z}_i)\tau(\boldsymbol{\beta}_i^\top\boldsymbol{X})]^\top$ for $i=1,2$. Then, Lemma C.8 shows that $\|\boldsymbol{v}_i\|\le C(1+\lambda)$ for some universal

constant C. Thus, similarly with the derivation of (C.33), we can deduce that

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\boldsymbol{w},\boldsymbol{z}_1,\boldsymbol{z}_2}\left[g_1(\boldsymbol{z}_1)g_2(\boldsymbol{z}_2)\sigma(\boldsymbol{w}^\top\boldsymbol{z}_1)\sigma(\boldsymbol{w}^\top\boldsymbol{z}_2)\cdot\boldsymbol{u}_1^\top\sigma(\boldsymbol{X}^\top\boldsymbol{w})\sigma(\boldsymbol{w}^\top\boldsymbol{X})\boldsymbol{u}_2\right]\\
=\;&\mathbb{E}_{\boldsymbol{w},\boldsymbol{z}_1,\boldsymbol{z}_2}\left[\sigma(\boldsymbol{w}^\top\boldsymbol{z}_1)\sigma(\boldsymbol{w}^\top\boldsymbol{z}_2)\cdot\boldsymbol{v}_1\boldsymbol{K}_\lambda^{-1/2}\sigma(\boldsymbol{X}^\top\boldsymbol{w})\sigma(\boldsymbol{w}^\top\boldsymbol{X})\boldsymbol{K}_\lambda^{-1/2}\boldsymbol{v}_2\right]\\
\le\;&\mathbb{E}_{\boldsymbol{w},\boldsymbol{z}_1,\boldsymbol{z}_2}\left[\sigma(\boldsymbol{w}^\top\boldsymbol{z}_1)^2\sigma(\boldsymbol{w}^\top\boldsymbol{z}_2)^2\right]^{\frac12}\mathbb{E}_{\boldsymbol{w},\boldsymbol{z}_1,\boldsymbol{z}_2}\left[\|\boldsymbol{v}_1\|^2\|\boldsymbol{v}_2\|^2\left\|\boldsymbol{K}_\lambda^{-1/2}\sigma(\boldsymbol{X}^\top\boldsymbol{w})\sigma(\boldsymbol{w}^\top\boldsymbol{X})\boldsymbol{K}_\lambda^{-1/2}\right\|^2\right]^{\frac12}\\
\le\;&C^2(1+\lambda)^2\|\sigma\|_4^2\,\mathbb{E}_{\boldsymbol{w}}\left[\left(\sigma(\boldsymbol{w}^\top\boldsymbol{X})\boldsymbol{K}_\lambda^{-1}\sigma(\boldsymbol{X}^\top\boldsymbol{w})\right)^2\right]^{\frac12}\\
\le\;&\frac{C^2(1+\lambda)^2\|\sigma\|_4^2}{\lambda_0}\mathbb{E}_{\boldsymbol{w}}\left[\left(\sum_{i=1}^n\sigma(\boldsymbol{w}^\top\boldsymbol{x}_i)^2\right)^2\right]^{\frac12}\\
\le\;&\frac{C^2(1+\lambda)^2\|\sigma\|_4^4\,n}{\lambda_0},
\end{aligned}$$

where the last line is analogous to (C.33). Therefore, $\mathbb{E}[\delta_2^2]\le C_2(1+\lambda)^2\frac{n}{N}$. This indicates, for any $t>0$,

$$\mathbb{P}\left(|\delta_2|>2t(1+\lambda)\right)\le\frac{C_2n}{Nt^2}.$$

Hence, by taking $t=\log(N)\sqrt{C_2n/N}$, we can conclude the bound of δ_2 in (C.31) with probability at least $1-\frac14\log^{-2}(N)$.

The analysis of $\delta_{1,2}$ is similar to the analysis for δ_2. By definition, we have

$$\begin{aligned}
\delta_{1,2}=\;&2\,\mathrm{Tr}\left(\mathbb{E}_{\boldsymbol{x}}\left[(\boldsymbol{K}_{m,N}-\boldsymbol{K}_m)\boldsymbol{K}_m^\top\right]\boldsymbol{K}_\lambda^{-1}(\boldsymbol{\Psi}+\sigma_\varepsilon^2\,\mathrm{Id})\boldsymbol{K}_\lambda^{-1}\right)\\
=\;&2\mathbb{E}_{\boldsymbol{x}}\left[\boldsymbol{K}_m^\top\boldsymbol{K}_\lambda^{-1}(\boldsymbol{\Psi}+\sigma_\varepsilon^2\,\mathrm{Id})\boldsymbol{K}_\lambda^{-1}(\boldsymbol{K}_{m,N}-\boldsymbol{K}_m)\right].
\end{aligned}$$

Then, consider $\boldsymbol{z}_1,\boldsymbol{z}_2$ as i.i.d. copies of $\boldsymbol{x}$. Let $\boldsymbol{A}:=\boldsymbol{K}_\lambda^{-1}(\boldsymbol{\Psi}+\sigma_\varepsilon^2\,\mathrm{Id})\boldsymbol{K}_\lambda^{-1}$ and

$$\boldsymbol{K}_{m,i}:=\mathbb{E}_{\boldsymbol{w}}[\sigma(\boldsymbol{w}^\top\boldsymbol{z}_i)\sigma(\boldsymbol{w}^\top\boldsymbol{X})]^\top\in\mathbb{R}^n,$$

for $i=1,2$. Then, we have

$$\begin{aligned}
\mathbb{E}[\delta_{1,2}^2]=&\frac{4}{N^2}\sum_{i,j=1}^N\mathbb{E}\left[\boldsymbol{K}_{m,1}^\top\boldsymbol{A}\left(\sigma(\boldsymbol{w}_i^\top\boldsymbol{z}_1)\sigma(\boldsymbol{X}^\top\boldsymbol{w}_i)-\boldsymbol{K}_{m,1}\right)\left(\sigma(\boldsymbol{w}_j^\top\boldsymbol{z}_2)\sigma(\boldsymbol{w}_j^\top\boldsymbol{X})-\boldsymbol{K}_{m,2}^\top\right)\boldsymbol{A}\boldsymbol{K}_{m,2}\right]\\
=\;&\frac{4}{N^2}\sum_{i=1}^N\mathbb{E}\left[\boldsymbol{K}_{m,1}^\top\boldsymbol{A}\left(\sigma(\boldsymbol{w}_i^\top\boldsymbol{z}_1)\sigma(\boldsymbol{X}^\top\boldsymbol{w}_i)-\boldsymbol{K}_{m,1}\right)\left(\sigma(\boldsymbol{w}_i^\top\boldsymbol{z}_2)\sigma(\boldsymbol{w}_i^\top\boldsymbol{X})-\boldsymbol{K}_{m,2}^\top\right)\boldsymbol{A}\boldsymbol{K}_{m,2}\right]\\
=\;&\frac{4}{N^2}\sum_{i=1}^N\mathbb{E}\left[\boldsymbol{K}_{m,1}^\top\boldsymbol{A}\left(\sigma(\boldsymbol{w}_i^\top\boldsymbol{z}_1)\sigma(\boldsymbol{w}_i^\top\boldsymbol{z}_2)\sigma(\boldsymbol{X}^\top\boldsymbol{w}_i)\sigma(\boldsymbol{w}_i^\top\boldsymbol{X})-\boldsymbol{K}_{m,1}\boldsymbol{K}_{m,2}^\top\right)\boldsymbol{A}\boldsymbol{K}_{m,2}\right]\\
\overset{(i)}{\le}\;&\frac{4}{N}\mathbb{E}_{\boldsymbol{w},\boldsymbol{z}_1,\boldsymbol{z}_2}\left[\sigma(\boldsymbol{w}^\top\boldsymbol{z}_1)\sigma(\boldsymbol{w}^\top\boldsymbol{z}_2)\boldsymbol{K}_{m,1}^\top\boldsymbol{A}\sigma(\boldsymbol{X}^\top\boldsymbol{w})\sigma(\boldsymbol{w}^\top\boldsymbol{X})\boldsymbol{A}\boldsymbol{K}_{m,2}\right]\\
\le\;&\frac{4}{N}\mathbb{E}\left[\sigma(\boldsymbol{w}^\top\boldsymbol{z}_1)^2\sigma(\boldsymbol{w}^\top\boldsymbol{z}_2)^2\right]^{\frac12}\mathbb{E}\left[\left\|\boldsymbol{K}_{m,1}^\top\boldsymbol{K}_\lambda^{-\frac12}\right\|^2\left\|\boldsymbol{K}_{m,2}^\top\boldsymbol{K}_\lambda^{-\frac12}\right\|^2\left\|\boldsymbol{K}_\lambda^{\frac12}\boldsymbol{A}\sigma(\boldsymbol{X}^\top\boldsymbol{w})\sigma(\boldsymbol{w}^\top\boldsymbol{X})\boldsymbol{A}\boldsymbol{K}_\lambda^{\frac12}\right\|^2\right]\\
\overset{(ii)}{\le}\;&\frac{4C^2(1+\lambda)^2\|\sigma\|_4^2}{\lambda_0N}\mathbb{E}\left[\left(\sum_{i=1}^n\sigma(\boldsymbol{w}^\top\boldsymbol{x}_i)^2\right)^2\right]^{1/2}\le C_{1,2}(1+\lambda)^2\frac{n}{N},
\end{aligned}$$

for some constant $C_{1,2}>0$, where (i) is because of positiveness of $\boldsymbol{A}$ and (ii) is due to Lemmas C.7 and C.8. Thus, Markov inequality allows us to obtain for a constant $C>0$,

$$\mathbb{P}\left(|\delta_{1,2}|>C(1+\lambda)\log(N)\sqrt{\frac{n}{N}}\right)\le\frac{1}{4\log^2(N)},$$

Together with (C.34), we can conclude the bound for δ_1 in (C.30). □

Based on the above lemmas, we are now ready to prove Theorem 2.11 for the concentrations of the generalization errors between RFRR and KRR.

Proof of Theorem 2.11. Recall $\boldsymbol{K} = \boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X})$ and $\boldsymbol{K}_N = \boldsymbol{K}_N(\boldsymbol{X}, \boldsymbol{X})$. Hence, we can further decompose the test errors (2.20) for both RFRR and KRR in the following way:

$$\begin{aligned}
\mathcal{L}(\hat{f}_\lambda^{(\mathsf{K})}) = \ & \mathbb{E}[|f^*(\boldsymbol{x})|^2] + \operatorname{Tr}\left[(\boldsymbol{K}+\lambda\,\mathrm{Id})^{-1}\mathbb{E}[\boldsymbol{y}\boldsymbol{y}^\top](\boldsymbol{K}+\lambda\,\mathrm{Id})^{-1}\mathbb{E}[\boldsymbol{K}(\boldsymbol{X},\boldsymbol{x})\boldsymbol{K}(\boldsymbol{x},\boldsymbol{X})]\right] \\
& - 2\operatorname{Tr}\left[(\boldsymbol{K}+\lambda\,\mathrm{Id})^{-1}\mathbb{E}[\boldsymbol{y}f^*(\boldsymbol{x})\boldsymbol{K}(\boldsymbol{x},\boldsymbol{X})]\right], \\
\mathcal{L}(\hat{f}_\lambda^{(\mathsf{RF})}) = \ & \mathbb{E}[|f^*(\boldsymbol{x})|^2] + \operatorname{Tr}\left[(\boldsymbol{K}_N+\lambda\,\mathrm{Id})^{-1}\mathbb{E}[\boldsymbol{y}\boldsymbol{y}^\top](\boldsymbol{K}_N+\lambda\,\mathrm{Id})^{-1}\mathbb{E}[\boldsymbol{K}_N(\boldsymbol{X},\boldsymbol{x})\boldsymbol{K}_N(\boldsymbol{x},\boldsymbol{X})]\right] \\
& - 2\operatorname{Tr}\left[(\boldsymbol{K}_N+\lambda\,\mathrm{Id})^{-1}\mathbb{E}[\boldsymbol{y}f^*(\boldsymbol{x})\boldsymbol{K}_N(\boldsymbol{x},\boldsymbol{X})]\right],
\end{aligned}$$

where we are taking expectations with respect to $\boldsymbol{x}$, $\boldsymbol{\beta}$, and $\boldsymbol{\varepsilon}$. Let us denote

$$\begin{aligned}
E_1 := \ & \operatorname{Tr}\left[(\boldsymbol{K}_N+\lambda\,\mathrm{Id})^{-1}\mathbb{E}[\boldsymbol{y}\boldsymbol{y}^\top](\boldsymbol{K}_N+\lambda\,\mathrm{Id})^{-1}\mathbb{E}[\boldsymbol{K}_N(\boldsymbol{X},\boldsymbol{x})\boldsymbol{K}_N(\boldsymbol{x},\boldsymbol{X})]\right], \\
\bar{E}_1 := \ & \operatorname{Tr}\left[(\boldsymbol{K}+\lambda\,\mathrm{Id})^{-1}\mathbb{E}[\boldsymbol{y}\boldsymbol{y}^\top](\boldsymbol{K}+\lambda\,\mathrm{Id})^{-1}\mathbb{E}[\boldsymbol{K}(\boldsymbol{X},\boldsymbol{x})\boldsymbol{K}(\boldsymbol{x},\boldsymbol{X})]\right], \\
E_2 := \ & \operatorname{Tr}\left[(\boldsymbol{K}_N+\lambda\,\mathrm{Id})^{-1}\mathbb{E}[\boldsymbol{y}f^*(\boldsymbol{x})\boldsymbol{K}_N(\boldsymbol{x},\boldsymbol{X})]\right], \\
\bar{E}_2 := \ & \operatorname{Tr}\left[(\boldsymbol{K}+\lambda\,\mathrm{Id})^{-1}\mathbb{E}[\boldsymbol{y}f^*(\boldsymbol{x})\boldsymbol{K}(\boldsymbol{x},\boldsymbol{X})]\right].
\end{aligned}$$

Therefore, by taking the expectation with respect to $\boldsymbol{\beta}$ and ε, we have

$$\begin{aligned}
E_1 = \ & \mathbb{E}_{\boldsymbol{x}}\left[\boldsymbol{K}_N(\boldsymbol{x},\boldsymbol{X})(\boldsymbol{K}_N+\lambda\,\mathrm{Id})^{-1}\left(\boldsymbol{\Psi}+\sigma_\varepsilon^2\,\mathrm{Id}\right)(\boldsymbol{K}_N+\lambda\,\mathrm{Id})^{-1}\boldsymbol{K}_N(\boldsymbol{X},\boldsymbol{x})\right], \\
\bar{E}_1 = \ & \mathbb{E}_{\boldsymbol{x}}\left[\boldsymbol{K}(\boldsymbol{x},\boldsymbol{X})(\boldsymbol{K}+\lambda\,\mathrm{Id})^{-1}\left(\boldsymbol{\Psi}+\sigma_\varepsilon^2\,\mathrm{Id}\right)(\boldsymbol{K}+\lambda\,\mathrm{Id})^{-1}\boldsymbol{K}(\boldsymbol{X},\boldsymbol{x})\right], \\
E_2 = \ & \operatorname{Tr}\left[(\boldsymbol{K}_N+\lambda\,\mathrm{Id})^{-1}\mathbb{E}[\boldsymbol{u}\boldsymbol{K}_N(\boldsymbol{x},\boldsymbol{X})]\right], \\
\bar{E}_2 = \ & \operatorname{Tr}\left[(\boldsymbol{K}+\lambda\,\mathrm{Id})^{-1}\mathbb{E}[\boldsymbol{u}\boldsymbol{K}(\boldsymbol{x},\boldsymbol{X})]\right],
\end{aligned}$$

where $\boldsymbol{\Psi} = \mathbb{E}_{\boldsymbol{\beta}}[\tau(\boldsymbol{X}^\top\boldsymbol{\beta})\tau(\boldsymbol{\beta}^\top\boldsymbol{X})]$ and $\boldsymbol{u} = \mathbb{E}_{\boldsymbol{\beta}}[\tau(\boldsymbol{X}^\top\boldsymbol{\beta})f^*(\boldsymbol{x})] \in \mathbb{R}^n$. We can further get the decomposition: $E_1 - \bar{E}_1 = J_{1,1} + J_{1,2} + J_{1,3}$ and $E_2 - \bar{E}_2 = J_{2,1} + J_{2,2}$, where

$$\begin{aligned}
J_{1,1} := \ & \mathbb{E}_{\boldsymbol{x}}\left[\boldsymbol{K}_{m,N}^\top\left(\boldsymbol{K}_{N,\lambda}^{-1} - \boldsymbol{K}_\lambda^{-1}\right)\left(\boldsymbol{\Psi}+\sigma_\varepsilon^2\,\mathrm{Id}\right)\boldsymbol{K}_{N,\lambda}^{-1}\boldsymbol{K}_{m,N}\right], \\
J_{1,2} := \ & \mathbb{E}_{\boldsymbol{x}}\left[\boldsymbol{K}_{m,N}^\top\boldsymbol{K}_\lambda^{-1}\left(\boldsymbol{\Psi}+\sigma_\varepsilon^2\,\mathrm{Id}\right)\left(\boldsymbol{K}_{N,\lambda}^{-1} - \boldsymbol{K}_\lambda^{-1}\right)\boldsymbol{K}_{m,N}\right], \\
J_{1,3} := \ & \mathbb{E}_{\boldsymbol{\beta},\boldsymbol{\varepsilon}}\left[\boldsymbol{y}^\top\boldsymbol{K}_\lambda^{-1}\left(\boldsymbol{K}_N^{(2)} - \boldsymbol{K}^{(2)}\right)\boldsymbol{K}_\lambda^{-1}\boldsymbol{y}\right], \\
J_{2,1} := \ & \mathbb{E}_{\boldsymbol{x}}\left[\boldsymbol{K}_N(\boldsymbol{x},\boldsymbol{X})\left(\boldsymbol{K}_{N,\lambda}^{-1} - \boldsymbol{K}_\lambda^{-1}\right)\boldsymbol{u}\right], \\
J_{2,2} := \ & \mathbb{E}_{\boldsymbol{x}}\left[\left(\boldsymbol{K}_N(\boldsymbol{x},\boldsymbol{X}) - \boldsymbol{K}(\boldsymbol{x},\boldsymbol{X})\right)\boldsymbol{K}_\lambda^{-1}\boldsymbol{u}\right].
\end{aligned}$$

Recall that $\boldsymbol{\Psi} = \mathbb{E}_{\boldsymbol{\beta}}[\tau(\boldsymbol{X}^\top\boldsymbol{\beta})\tau(\boldsymbol{\beta}^\top\boldsymbol{X})]$, $\boldsymbol{u} = \mathbb{E}_{\boldsymbol{\beta}}[\tau(\boldsymbol{X}^\top\boldsymbol{\beta})f^*(\boldsymbol{x})] \in \mathbb{R}^n$ and $\boldsymbol{\Psi}_\lambda = \boldsymbol{\Psi} + \lambda\,\mathrm{Id}$. Notice that

$$\begin{aligned}
\boldsymbol{K}_{N,\lambda}^{-1} - \boldsymbol{K}_\lambda^{-1} = \ & \boldsymbol{K}_{N,\lambda}^{-1}\left(\boldsymbol{K} - \boldsymbol{K}_N\right)\boldsymbol{K}_\lambda^{-1} \\
= \ & \boldsymbol{K}_{N,\lambda}^{-\frac{1}{2}}\boldsymbol{K}_{N,\lambda}^{-\frac{1}{2}}\boldsymbol{K}_\lambda^{\frac{1}{2}}\boldsymbol{K}_\lambda^{-\frac{1}{2}}\left(\boldsymbol{K} - \boldsymbol{K}_N\right)\boldsymbol{K}_\lambda^{-\frac{1}{2}}\boldsymbol{K}_\lambda^{-\frac{1}{2}}.
\end{aligned}$$

Hence, we can apply Proposition C.1, Corollary C.2, Lemmas C.6, C.7 and C.8 to conclude that

$$
\begin{aligned}
|J_{1,1}| \le\ & \mathbb{E}_{\boldsymbol{x}}\left[\left\|\boldsymbol{K}_{m,N}^{\top}\boldsymbol{K}_{N,\lambda}^{-1/2}\right\|^2\right]\cdot\left\|\boldsymbol{K}_{N,\lambda}^{-1/2}\boldsymbol{K}_{\lambda}^{1/2}\right\|^2\left\|\boldsymbol{K}_{\lambda}^{-1/2}\left(\boldsymbol{K}-\boldsymbol{K}_N\right)\boldsymbol{K}_{\lambda}^{-1/2}\right\|\cdot\left\|\boldsymbol{K}_{\lambda}^{-1/2}\left(\boldsymbol{\Psi}+\sigma_\varepsilon^2\,\mathrm{Id}\right)\boldsymbol{K}_{\lambda}^{-1/2}\right\| \\
\le\ & C(1+\lambda)\log(N)\sqrt{\frac{n}{N}}, \\
|J_{1,2}| \le\ & \mathbb{E}_{\boldsymbol{x}}\left[\left\|\boldsymbol{K}_{m,N}^{\top}\boldsymbol{K}_{N,\lambda}^{-1/2}\right\|^2\right]\cdot\left\|\boldsymbol{K}_{N,\lambda}^{-1/2}\boldsymbol{K}_{\lambda}^{1/2}\right\|\left\|\boldsymbol{K}_{N,\lambda}^{1/2}\boldsymbol{K}_{\lambda}^{-1/2}\right\| \\
& \cdot\left\|\boldsymbol{K}_{\lambda}^{-1/2}\left(\boldsymbol{K}-\boldsymbol{K}_N\right)\boldsymbol{K}_{\lambda}^{-1/2}\right\|\left\|\boldsymbol{K}_{\lambda}^{-1/2}\left(\boldsymbol{\Psi}+\sigma_\varepsilon^2\,\mathrm{Id}\right)\boldsymbol{K}_{\lambda}^{-1/2}\right\| \\
\le\ & C(1+\lambda)\log(N)\sqrt{\frac{n}{N}}, \\
|J_{2,1}| \le\ & \mathbb{E}_{\boldsymbol{x}}\left[\left|\boldsymbol{K}_{m,N}^{\top}\boldsymbol{K}_{N,\lambda}^{-1/2}\boldsymbol{K}_{N,\lambda}^{-1/2}\boldsymbol{K}_{\lambda}^{1/2}\boldsymbol{K}_{\lambda}^{-1/2}(\boldsymbol{K}-\boldsymbol{K}_N)\boldsymbol{K}_{\lambda}^{-1/2}\boldsymbol{K}_{\lambda}^{-1/2}\boldsymbol{u}\right|\right] \\
\le\ & \mathbb{E}_{\boldsymbol{x}}\left[\left\|\boldsymbol{K}_{m,N}^{\top}\boldsymbol{K}_{N,\lambda}^{-1/2}\right\|\cdot\left\|\boldsymbol{K}_{N,\lambda}^{-1/2}\boldsymbol{K}_{\lambda}^{1/2}\right\|\cdot\left\|\boldsymbol{K}_{\lambda}^{-1/2}(\boldsymbol{K}-\boldsymbol{K}_N)\boldsymbol{K}_{\lambda}^{-1/2}\right\|\left\|\boldsymbol{K}_{\lambda}^{-1/2}\boldsymbol{u}\right\|\right] \\
\le\ & C(1+\lambda)\log(N)\sqrt{\frac{n}{N}},
\end{aligned}
$$

for some constant $C>0$ depending on the norms of τ and σ, λ_0 and σ_ε with probability at least $1-N^{-1}$.

Meanwhile, based on Lemma C.9, $|J_{1,3}|$ and $|J_{2,2}|$ are both less than $C\log(N)\sqrt{\frac{n}{N}}$ with probability at least $1-\log^{-2}(N)$, because $\delta_1=J_{1,3}$ and $\delta_2=J_{2,2}$. Hence, combing the controls of $J_{1,1}, J_{1,2}, J_{1,3}$ and $J_{2,1}, J_{2,2}$, we complete the proof of Theorem 2.11. □

C.6 Proof of Theorem 2.12

We first show (2.23). In the proof of Proposition 2.4, we know $\lambda_{\min}(\boldsymbol{K}_\ell)\ge 2\lambda_0$ and $\lambda_{\min}(\boldsymbol{K})\ge\lambda_0$. Similar to the proof of (C.12), using the closed form formula of the training error from (2.10) and Proposition 2.4, we have

$$
\begin{aligned}
\left|E_{\text{train}}^{(\ell,\lambda)}-E_{\text{train}}^{(\mathsf{K},\lambda)}\right| =& \frac{\lambda^2}{n}\left|\boldsymbol{y}^{\top}\left[(\boldsymbol{K}_\ell+\lambda\,\mathrm{Id})^{-2}-(\boldsymbol{K}+\lambda\,\mathrm{Id})^{-2}\right]\boldsymbol{y}\right| \\
\le& \frac{\lambda^2}{n}\|(\boldsymbol{K}_\ell+\lambda\,\mathrm{Id})^{-2}-(\boldsymbol{K}+\lambda\,\mathrm{Id})^{-2}\|\cdot\|\boldsymbol{y}\|^2 \\
\le& \frac{3\lambda^2\|\boldsymbol{y}\|^2}{2\lambda_0 n}\|(\boldsymbol{K}_\ell+\lambda\,\mathrm{Id})^{-1}-(\boldsymbol{K}+\lambda\,\mathrm{Id})^{-1}\| \\
\le& \frac{3\lambda^2\|\boldsymbol{y}\|^2}{2\lambda_0^3 n}\|\boldsymbol{K}-\boldsymbol{K}_\ell\| \le \frac{C\lambda^2\|\boldsymbol{y}\|^2\|\sigma\|_4^2}{\lambda_0^3 n}\left\|\left(\boldsymbol{X}^{\top}\boldsymbol{X}\right)^{\odot\ell+1}-\mathrm{Id}\right\|_F
\end{aligned}
\tag{C.35}
$$

for an absolute constant $C>0$, where in the third inequality, we use the estimate

$$
\|(\boldsymbol{K}_\ell+\lambda\,\mathrm{Id})^{-1}-(\boldsymbol{K}+\lambda\,\mathrm{Id})^{-1}\| = \left\|\boldsymbol{K}_\lambda^{-1}(\boldsymbol{K}-\boldsymbol{K}_\ell)\boldsymbol{K}_{\ell,\lambda}^{-1}\right\| \le \frac{1}{(\lambda+\lambda_0)\lambda_0}\|\boldsymbol{K}-\boldsymbol{K}_\ell\|. \tag{C.36}
$$

Next, we prove (2.24). With the same proof in Lemma C.3, we also have

$$
\left(\lambda+\|\sigma\|_2^2\right)^{-1} \le \operatorname{tr}\boldsymbol{K}_{\ell,\lambda}^{-1} \le \lambda_0^{-1}. \tag{C.37}
$$

From the definition of GCV in (2.16), we have

$$
\begin{aligned}
|\mathrm{GCV}_n^{(\mathsf{K},\lambda)}-\mathrm{GCV}_n^{(\ell,\lambda)}| \le& \lambda^{-2}\left|\left(\left(\operatorname{tr}\boldsymbol{K}_\lambda^{-1}\right)^{-2}-\left(\operatorname{tr}\boldsymbol{K}_{\ell,\lambda}^{-1}\right)^{-2}\right)E_{\text{train}}^{(\mathsf{K},\lambda)}\right| \\
&+\lambda^{-2}\left|\left(\operatorname{tr}\boldsymbol{K}_\lambda^{-1}\right)^{-2}\left(E_{\text{train}}^{(\mathsf{K},\lambda)}-E_{\text{train}}^{(\ell,\lambda)}\right)\right|.
\end{aligned}
\tag{C.38}
$$

Equipped with (C.37) and Lemma C.3, following every step in the proof of (2.17) in Section C.4, we can obtain a similar bound for (C.38) as follows:

$$\begin{aligned}
&\lambda^{-2}\left|\left(\left(\operatorname{tr}\boldsymbol{K}_{\lambda}^{-1}\right)^{-2}-\left(\operatorname{tr}\boldsymbol{K}_{\ell,\lambda}^{-1}\right)^{-2}\right)E_{\text{train}}^{(\mathsf{K},\lambda)}\right| \\
\leq&\left(\operatorname{tr}\boldsymbol{K}_{\lambda}^{-1}\right)^{-2}\left(\operatorname{tr}\boldsymbol{K}_{\ell,\lambda}^{-1}\right)^{-2}\left|\operatorname{tr}(\boldsymbol{K}_{\lambda}^{-1}-\boldsymbol{K}_{\ell,\lambda}^{-1})\right|\operatorname{tr}\left(\boldsymbol{K}_{\lambda}^{-1}+\boldsymbol{K}_{\ell,\lambda}^{-1}\right)\frac{1}{n}\|\boldsymbol{K}_{\lambda}^{-2}\|\,\|\boldsymbol{y}\|^{2} \\
\leq&\frac{8(\lambda+\|\sigma\|_2^2)^4}{\lambda_0^5 n}\|\boldsymbol{K}-\boldsymbol{K}_{\ell}\|\leq\frac{8\sqrt{2}(\lambda+\|\sigma\|_2^2)^4\|\sigma\|_4^2}{\lambda_0^5 n}\left\|\left(\boldsymbol{X}^{\top}\boldsymbol{X}\right)^{\odot\ell+1}-\mathrm{Id}\right\|_F.
\end{aligned}$$

Similarly, for the second term (C.23), we have from (C.35) and Lemma C.3,

$$\lambda^{-2}\left|\left(\operatorname{tr}\boldsymbol{K}_{\lambda}^{-1}\right)^{-2}\left(E_{\text{train}}^{(\mathsf{K},\lambda)}-E_{\text{train}}^{(\ell,\lambda)}\right)\right|\leq\frac{C(\lambda+\|\sigma\|_2^2)^2\|\sigma\|_4^2}{\lambda_0^3 n}\left\|\left(\boldsymbol{X}^{\top}\boldsymbol{X}\right)^{\odot\ell+1}-\mathrm{Id}\right\|_F,$$

which implies (2.24). Next, we verify (2.25). Recall (2.14) and (2.15). Analogously, we have

$$\mathrm{CV}_n^{(\ell,\lambda)}=\frac{1}{n}\boldsymbol{y}^{\top}\boldsymbol{K}_{\ell,\lambda}^{-1}\boldsymbol{D}_{\ell}^{-2}\boldsymbol{K}_{\ell,\lambda}^{-1}\boldsymbol{y},$$

where $\boldsymbol{D}_\ell$ is a diagonal matrix with diagonals $[\boldsymbol{D}_\ell]_{ii}=[\boldsymbol{K}_{\ell,\lambda}^{-1}]_{ii}$ for $i\in[n]$. Notice that $\|\boldsymbol{D}_\ell-\boldsymbol{D}\|$ has the same upper bound as (C.36), and any $[\boldsymbol{D}_\ell]_{ii}$ has the same lower and upper bounds as (C.37) for $i\in[n]$. Hence, repeatedly applying Proposition 2.4 and following (C.24), we can obtain

$$\left|\mathrm{CV}_n^{(\ell,\lambda)}-\mathrm{CV}_n^{(\mathsf{K},\lambda)}\right|\leq C_2(1+\lambda^4)\frac{\|\boldsymbol{y}\|^2}{n}\left\|\left(\boldsymbol{X}^{\top}\boldsymbol{X}\right)^{\odot\ell+1}-\mathrm{Id}\right\|_F,$$

for some constant $C_2>0$ which only relies on $\|\sigma\|_2,\|\sigma\|_4$, and λ_0. This concludes the bound in (2.25).

Finally, we can repeat the analysis in the proof of Theorem 2.11 and apply (C.36) to obtain (2.26). By taking expectation with respect to $\boldsymbol{\beta}$ and ε, we have $\left|\mathcal{L}(\hat{f}_\lambda^{(\ell)}(\boldsymbol{x}))-\mathcal{L}(\hat{f}_\lambda^{(\mathsf{K})}(\boldsymbol{x}))\right|\leq|E_1'-\bar{E}_1|+|E_2'-\bar{E}_2|$, where

$$\begin{aligned}
E_1'&:=\mathbb{E}_{\boldsymbol{x}}\left[\boldsymbol{K}_{\ell}(\boldsymbol{x},\boldsymbol{X})\boldsymbol{K}_{\ell,\lambda}^{-1}\left(\boldsymbol{\Psi}+\sigma_\varepsilon^2\,\mathrm{Id}\right)\boldsymbol{K}_{\ell,\lambda}^{-1}\boldsymbol{K}_{\ell}(\boldsymbol{X},\boldsymbol{x})\right],\\
\bar{E}_1&=\mathbb{E}_{\boldsymbol{x}}\left[\boldsymbol{K}(\boldsymbol{x},\boldsymbol{X})\boldsymbol{K}_{\lambda}^{-1}\left(\boldsymbol{\Psi}+\sigma_\varepsilon^2\,\mathrm{Id}\right)\boldsymbol{K}_{\lambda}^{-1}\boldsymbol{K}(\boldsymbol{X},\boldsymbol{x})\right],\\
E_2'&:=\operatorname{Tr}\left[\boldsymbol{K}_{\ell,\lambda}^{-1}\mathbb{E}[\boldsymbol{u}\boldsymbol{K}_{\ell}(\boldsymbol{x},\boldsymbol{X})]\right],\\
\bar{E}_2&=\operatorname{Tr}\left[\boldsymbol{K}_{\lambda}^{-1}\mathbb{E}[\boldsymbol{u}\boldsymbol{K}(\boldsymbol{x},\boldsymbol{X})]\right].
\end{aligned}$$

Denote $\boldsymbol{K}_{m,\ell}=\boldsymbol{K}_\ell(\boldsymbol{X},\boldsymbol{x})$ and $\boldsymbol{K}_{\ell,\lambda}=\lambda\,\mathrm{Id}+\boldsymbol{K}_\ell(\boldsymbol{X},\boldsymbol{X})$. Recall that $\boldsymbol{\Psi}=\mathbb{E}_{\boldsymbol{\beta}}[\tau(\boldsymbol{X}^{\top}\boldsymbol{\beta})\tau(\boldsymbol{\beta}^{\top}\boldsymbol{X})]$ and $\boldsymbol{u}=\mathbb{E}_{\boldsymbol{\beta}}[\tau(\boldsymbol{X}^{\top}\boldsymbol{\beta})f^*(\boldsymbol{x})]\in\mathbb{R}^n$. Because of the Assumption 2.10, similar to the proof of Proposition 2.4, we obtain

$$\|\boldsymbol{K}_{m,\ell}-\boldsymbol{K}_m\|\leq\|\boldsymbol{K}_{m,\ell}-\boldsymbol{K}_m\|_2\leq\sqrt{2}\|\sigma\|_4^2\left\|(\boldsymbol{X}^{\top}\boldsymbol{x})^{\odot(\ell+1)}\right\|_2\leq\frac{1}{\sqrt{2}}\lambda_0. \tag{C.39}$$

Moreover, analogously to Lemma C.6, we have

$$\left\|\boldsymbol{K}_{m,\ell}^{\top}\boldsymbol{K}_{\ell,\lambda}^{-1/2}\right\|\leq\|\sigma\|_2^2+\lambda. \tag{C.40}$$

Also, following the proofs of Lemma C.7 and Lemma C.8, we can check that

$$\left\|\boldsymbol{K}_{\ell,\lambda}^{-1/2}\boldsymbol{\Psi}\boldsymbol{K}_{\ell,\lambda}^{-1/2}\right\|,\ \left\|\boldsymbol{K}_{\ell,\lambda}^{-1/2}\boldsymbol{u}\right\|\leq C, \tag{C.41}$$

for some constant $C>0$ depending only on σ,τ. Therefore, because of Proposition 2.4, Lemmas C.6 and C.7, and (C.39),

(C.40) and (C.41), we can deduce that

$$\begin{aligned}
|E_1' - \bar{E}_1| \le\ & \left|(\boldsymbol{K}_{m,\ell} - \boldsymbol{K}_m)^\top \boldsymbol{K}_{\ell,\lambda}^{-1/2} \boldsymbol{K}_{\ell,\lambda}^{-1/2} \left(\boldsymbol{\Psi} + \sigma_\varepsilon^2 \operatorname{Id}\right) \boldsymbol{K}_{\ell,\lambda}^{-1/2} \boldsymbol{K}_{\ell,\lambda}^{-1/2} \boldsymbol{K}_{m,\ell}\right| \\
& + \left|\boldsymbol{K}_m^\top \boldsymbol{K}_\lambda^{-1/2} \boldsymbol{K}_\lambda^{-1/2} (\boldsymbol{K} - \boldsymbol{K}_\ell) \boldsymbol{K}_{\ell,\lambda}^{-1/2} \boldsymbol{K}_{\ell,\lambda}^{-1/2} \left(\boldsymbol{\Psi} + \sigma_\varepsilon^2 \operatorname{Id}\right) \boldsymbol{K}_{\ell,\lambda}^{-1/2} \boldsymbol{K}_{\ell,\lambda}^{-1/2} \boldsymbol{K}_{m,\ell}\right| \\
& + \left|\boldsymbol{K}_m^\top \boldsymbol{K}_\lambda^{-1/2} \boldsymbol{K}_\lambda^{-1/2} \left(\boldsymbol{\Psi} + \sigma_\varepsilon^2 \operatorname{Id}\right) \boldsymbol{K}_\lambda^{-1/2} \boldsymbol{K}_\lambda^{-1/2} (\boldsymbol{K} - \boldsymbol{K}_\ell) \boldsymbol{K}_{\ell,\lambda}^{-1/2} \boldsymbol{K}_{\ell,\lambda}^{-1/2} \boldsymbol{K}_{m,\ell}\right| \\
& + \left|\boldsymbol{K}_m^\top \boldsymbol{K}_\lambda^{-1/2} \boldsymbol{K}_\lambda^{-1/2} \left(\boldsymbol{\Psi} + \sigma_\varepsilon^2 \operatorname{Id}\right) \boldsymbol{K}_\lambda^{-1/2} \boldsymbol{K}_\lambda^{-1/2} (\boldsymbol{K}_{m,\ell} - \boldsymbol{K}_m)\right| \\
\le\ & C_1'(1+\lambda) \left\| \left(\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}}\right)^{\odot \ell+1} - \operatorname{Id} \right\|_F,
\end{aligned}$$

for some constant $C_1' > 0$. Similarly, due to Lemma C.8, (C.39), (C.40) and (C.41), we can obtain

$$\begin{aligned}
|E_2' - \bar{E}_2| \le\ & \mathbb{E}\left|(\boldsymbol{K}_{m,\ell} - \boldsymbol{K}_m)^\top \boldsymbol{K}_{\ell,\lambda}^{-1/2} \boldsymbol{K}_{\ell,\lambda}^{-1/2} \boldsymbol{u}\right| \\
& + \mathbb{E}\left|\boldsymbol{K}_m^\top \boldsymbol{K}_{\ell,\lambda}^{-1/2} \boldsymbol{K}_{\ell,\lambda}^{-1/2} (\boldsymbol{K} - \boldsymbol{K}_\ell) \boldsymbol{K}_\lambda^{-1/2} \boldsymbol{K}_\lambda^{-1/2} \boldsymbol{u}\right| \\
\le\ & C_2'(1+\lambda) \left\| \left(\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}}\right)^{\odot \ell+1} - \operatorname{Id} \right\|_F,
\end{aligned}$$

for some constant $C_2' > 0$. This completes the proof of (2.26).

C.7 Proof of Theorem 2.13

First, we state a more generic statement of the lower bound of the generalization error for RFRR. Instead of proving Theorem 2.13, we prove the following theorem in this section.

Theorem C.10. *Under the assumptions of Theorem 2.11, when $N/\log^2(N) \ge C_1(1+\lambda^2)n$ and $n \ge \max\{n_0, n_1\}$, with probability at least $1 - \log^{-1}(N)$,*

$$\mathcal{L}(\hat{f}_\lambda^{(\mathsf{RF})}) \ge\ \|P_{>\ell} f^*\|_2^2 - C_2(1+\lambda)\log(N)\sqrt{\frac{n}{N}} - C_2\sqrt{n}\left\|(\boldsymbol{X}^\top \boldsymbol{x})^{\odot(\ell+1)}\right\|_2,$$

and

$$\begin{aligned}
\mathcal{L}(\hat{f}_\lambda^{(\mathsf{RF})}) \ge\ & \|P_{>\ell} f^*\|_2^2 + \sigma_{\boldsymbol{\varepsilon}}^2 \mathbb{E}_{\boldsymbol{x}}\left[\boldsymbol{K}_{m,\ell}^\top \boldsymbol{K}_{\lambda,\ell}^{-2} \boldsymbol{K}_{m,\ell}\right] - C_2\sqrt{n}\left\|(\boldsymbol{X}^\top \boldsymbol{x})^{\odot(\ell+1)}\right\|_2 \\
& - C_2(1+\lambda)\left(\left\|\left(\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}}\right)^{\odot \ell+1} - \operatorname{Id}\right\|_F + \log(N)\sqrt{\frac{n}{N}}\right),
\end{aligned} \tag{C.42}$$

where C_1 depends only on σ, and $C_2 > 0$ depends only on σ, τ and σ_ε. In particular, when $N/\log^2 N \gg n$, with high probability,

$$\begin{aligned}
\mathcal{L}(\hat{f}_\lambda^{(\mathsf{RF})}) &\ge \|P_{>\ell} f^*\|_2^2 + \sigma_{\boldsymbol{\varepsilon}}^2 \mathbb{E}_{\boldsymbol{x}}\left[\boldsymbol{K}_{m,\ell}^\top \boldsymbol{K}_{\lambda,\ell}^{-2} \boldsymbol{K}_{m,\ell}\right] - o_n(1) \\
&\ge \|P_{>\ell} f^*\|_2^2 - o_n(1).
\end{aligned}$$

Proof. Since $\tau \in L^2(\mathbb{R}, \Gamma)$, we have the following Hermite expansion: $\tau(x) = \sum_{k=0}^\infty \zeta_k(\tau) h_k(x)$. Then

$$\begin{aligned}
f^*(\boldsymbol{x}) &= \tau(\boldsymbol{\beta}^\top \boldsymbol{x}) = \sum_{k=0}^\infty \zeta_k(\tau) h_k(\boldsymbol{\beta}^\top \boldsymbol{x}), \\
\left(P_{\le \ell} f^*\right)(\boldsymbol{x}) &= \sum_{k<\ell} \zeta_k(\tau) h_k(\boldsymbol{\beta}^\top \boldsymbol{x}), \quad \left(P_{>\ell} f^*\right)(\boldsymbol{x}) = \sum_{k \ge \ell+1} \zeta_k(\tau) h_k(\boldsymbol{\beta}^\top \boldsymbol{x}).
\end{aligned}$$

Similarly, we define

$$\begin{aligned}
f^*(\boldsymbol{X}) &= \tau(\boldsymbol{\beta}^\top \boldsymbol{X}) = \sum_{k=0}^\infty \zeta_k(\tau) h_k(\boldsymbol{\beta}^\top \boldsymbol{X}) \in \mathbb{R}^n, \\
\left(P_{\le \ell} f^*\right)(\boldsymbol{X}) &= \sum_{k \le \ell} \zeta_k(\tau) h_k(\boldsymbol{\beta}^\top \boldsymbol{X}), \quad \left(P_{>\ell} f^*\right)(\boldsymbol{X}) = \sum_{k \ge \ell+1} \zeta_k(\tau) h_k(\boldsymbol{\beta}^\top \boldsymbol{X}),
\end{aligned} \tag{C.43}$$

By the property of Hermite polynomials in (2.3), we know

$$\mathbb{E}_{\boldsymbol{\beta}}[h_j(\boldsymbol{\beta}^\top \boldsymbol{x}) h_k(\boldsymbol{\beta}^\top \boldsymbol{x}_i)] = \delta_{jk} \langle \boldsymbol{x}, \boldsymbol{x}_i \rangle^k.$$

This implies

$$\begin{aligned}
\|f^*\|_2^2 = \mathbb{E}_{\boldsymbol{\beta}}[f^*(\boldsymbol{x})^2] = \sum_{k=0}^{\infty} \zeta_k(\tau)^2 = \|\tau\|_2^2, \\
\|P_{\le \ell} f^*\|_2^2 = \sum_{k=0}^{\ell} \zeta_k^2(\tau), \qquad \|P_{>\ell} f^*\|_2^2 = \sum_{k=\ell+1}^{\infty} \zeta_k^2(\tau), \\
\mathbb{E}[P_{\le \ell} f^*(\boldsymbol{x}) P_{>\ell} f^*(\boldsymbol{x})] = 0.
\end{aligned} \tag{C.44}$$

From (2.9), the predictor of the KRR is given by

$$\hat{f}_\lambda^{(\mathsf{K})}(\boldsymbol{x}) := \boldsymbol{K}(\boldsymbol{x}, \boldsymbol{X})(\boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X}) + \lambda \,\mathrm{Id})^{-1} \left(f^*(\boldsymbol{X}) + \boldsymbol{\varepsilon}\right),$$

where

$$\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{X}) = \sum_{k=0}^{\infty} \zeta_k^2(\sigma) (\boldsymbol{x}^\top \boldsymbol{X})^{\odot k} \in \mathbb{R}^{1 \times n},$$

and from Assumption 2.10,

$$\|\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{X})\| \le \sqrt{2n}\, \|\sigma\|_4^2 . \tag{C.45}$$

Define

$$\begin{aligned}
P_{\le \ell} \hat{f}_\lambda^{(\mathsf{K})}(\boldsymbol{x}) &:= \boldsymbol{K}(\boldsymbol{x}, \boldsymbol{X})(\boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X}) + \lambda \,\mathrm{Id})^{-1} \left(P_{\le \ell} f^*(\boldsymbol{X})\right), \\
P_{>\ell} \hat{f}_\lambda^{(\mathsf{K})}(\boldsymbol{x}) &:= \boldsymbol{K}(\boldsymbol{x}, \boldsymbol{X})(\boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X}) + \lambda \,\mathrm{Id})^{-1} \left(P_{>\ell} f^*(\boldsymbol{X}) + \boldsymbol{\varepsilon}\right).
\end{aligned}$$

From the orthogonal relation in (2.3) and (C.43),

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\beta}, \boldsymbol{\varepsilon}}[P_{\le \ell} \hat{f}_\lambda^{(\mathsf{K})}(\boldsymbol{x}) P_{>\ell} \hat{f}_\lambda^{(\mathsf{K})}(\boldsymbol{x})] = \mathbf{0}, \\
&\mathbb{E}_{\boldsymbol{\beta}}[(P_{>\ell} f^*)(\boldsymbol{X})(P_{>\ell} f^*)(\boldsymbol{x})] = \sum_{k \ge \ell+1} \zeta_k^2(\tau)((\boldsymbol{x}^\top \boldsymbol{x}_1)^k, \ldots, (\boldsymbol{x}^\top \boldsymbol{x}_n)^k).
\end{aligned} \tag{C.46}$$

Then by the linearity of expectation, we have

$$\mathbb{E}_{\boldsymbol{\beta}, \boldsymbol{\varepsilon}}[\hat{f}_\lambda^{(\mathsf{K})}(\boldsymbol{x}) P_{>\ell} f^*(\boldsymbol{x})] = \sum_{k > \ell} \zeta_k^2(\tau) \boldsymbol{K}(\boldsymbol{x}, \boldsymbol{X})(\boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X}) + \lambda \,\mathrm{Id})^{-1}((\boldsymbol{x}^\top \boldsymbol{x}_1)^k, \ldots, (\boldsymbol{x}^\top \boldsymbol{x}_n)^k)^\top,$$

which implies

$$\begin{aligned}
\left|\mathbb{E}_{\boldsymbol{\beta}, \boldsymbol{\varepsilon}}[P_{>\ell} \hat{f}_\lambda^{(\mathsf{K})}(\boldsymbol{x}) P_{>\ell} f^*(\boldsymbol{x})]\right| &\le \|\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{X})\|\, \lambda_0^{-1} \sum_{k=\ell+1}^{\infty} \zeta_k^2(\tau) \left\|(\boldsymbol{X}^\top \boldsymbol{x})^{\odot k}\right\|_2 \\
&\le \sqrt{2n}\, \|\sigma\|_4^2\, \lambda_0^{-1} \|\tau\|_4^2 \left(\sum_{k=\ell+1}^{\infty} \left\|(\boldsymbol{X}^\top \boldsymbol{x})^{\odot k}\right\|_2^2 \right)^{1/2} \\
&\le 2\sqrt{2n}\, \|\sigma\|_4^2\, \lambda_0^{-1} \|\tau\|_4^2 \left\|(\boldsymbol{X}^\top \boldsymbol{x})^{\odot (\ell+1)}\right\|_2,
\end{aligned} \tag{C.47}$$

where the second inequality is due to (C.45), and the third inequality comes from Cauchy's inequality. Recall the general-

ization error of any predictor defined in (2.20). We have

$$\begin{aligned}
\mathcal{L}(\hat{f}_\lambda^{(\mathsf{K})}) =& \ \mathbb{E}\left(f^*(\boldsymbol{x}) - \hat{f}_\lambda^{(\mathsf{K})}(\boldsymbol{x})\right)^2 \\
=& \ \mathbb{E}\left(P_{\le \ell} f^*(\boldsymbol{x}) + P_{>\ell} f^*(\boldsymbol{x}) - P_{\le \ell}\hat{f}_\lambda^{(\mathsf{K})}(\boldsymbol{x}) - P_{>\ell}\hat{f}_\lambda^{(\mathsf{K})}(\boldsymbol{x})\right)^2 \\
=& \ \mathbb{E}\left(P_{\le \ell} f^*(\boldsymbol{x}) - P_{\le \ell}\hat{f}_\lambda^{(\mathsf{K})}(\boldsymbol{x})\right)^2 + \mathbb{E}\left(P_{>\ell} f^*(\boldsymbol{x}) - P_{>\ell}\hat{f}_\lambda^{(\mathsf{K})}(\boldsymbol{x})\right)^2 \\
& + 2\mathbb{E}\left[\left(P_{\le \ell} f^*(\boldsymbol{x}) - P_{\le \ell}\hat{f}_\lambda^{(\mathsf{K})}(\boldsymbol{x})\right)\left(P_{>\ell} f^*(\boldsymbol{x}) - P_{>\ell}\hat{f}_\lambda^{(\mathsf{K})}(\boldsymbol{x})\right)\right] \\
\ge& \ \mathbb{E}\left(P_{>\ell} f^*(\boldsymbol{x}) - P_{>\ell}\hat{f}_\lambda^{(\mathsf{K})}(\boldsymbol{x})\right)^2 \\
=& \ \|P_{>\ell} f^*\|_2^2 + \mathbb{E}[P_{>\ell}\hat{f}_\lambda^{(\mathsf{K})}(\boldsymbol{x})^2] - 2\mathbb{E}[P_{>\ell} f^*(\boldsymbol{x}) P_{>\ell}\hat{f}_\lambda^{(\mathsf{K})}(\boldsymbol{x})] \\
\ge& \ \|P_{>\ell} f^*\|_2^2 + \mathbb{E}[P_{>\ell}\hat{f}_\lambda^{(\mathsf{K})}(\boldsymbol{x})^2] - 4\sqrt{2n}\lambda_0^{-1}\|\sigma\|_4^2\|\tau\|_4^2\left\|(\boldsymbol{X}^\top\boldsymbol{x})^{\odot(\ell+1)}\right\|_2, \qquad \text{(C.48)}
\end{aligned}$$

where the first inequality is due to the orthogonal relations (C.44) and (C.46), and the second inequality is due to (C.47). Let $\boldsymbol{v} = \boldsymbol{K}_\lambda^{-1}\boldsymbol{K}(\boldsymbol{X}, \boldsymbol{x})$. The second term in (C.48) can be written as

$$\begin{aligned}
\mathbb{E}[P_{>\ell}\hat{f}_\lambda^{(\mathsf{K})}(\boldsymbol{x})^2] &= \mathbb{E}[(P_{>\ell} f^*(\boldsymbol{X}) + \boldsymbol{\varepsilon})^\top(\boldsymbol{K}_\lambda^{-1}\boldsymbol{K}(\boldsymbol{X}, \boldsymbol{x})\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{X})\boldsymbol{K}_\lambda^{-1})(P_{>\ell} f^*(\boldsymbol{X}) + \boldsymbol{\varepsilon})] \\
&= \mathbb{E}_{\boldsymbol{x}}\sum_{ij}\boldsymbol{v}_i\boldsymbol{v}_j\left(\mathbb{E}_{\boldsymbol{\beta}}[P_{>\ell} f^*(\boldsymbol{x}_i)P_{>\ell} f^*(\boldsymbol{x}_j)] + \delta_{ij}\sigma_{\boldsymbol{\varepsilon}}^2\right) \\
&= \sigma_{\boldsymbol{\varepsilon}}^2\mathbb{E}_{\boldsymbol{x}}\|\boldsymbol{v}\|^2 + \mathbb{E}[P_{>\ell} f^*(\boldsymbol{X})^\top(\boldsymbol{v}\boldsymbol{v}^\top)P_{>\ell} f^*(\boldsymbol{X})], \\
&\ge \sigma_{\boldsymbol{\varepsilon}}^2\mathbb{E}_{\boldsymbol{x}}\|\boldsymbol{v}\|^2 = \sigma_{\boldsymbol{\varepsilon}}^2\operatorname{Tr}\boldsymbol{K}_\lambda^{-1}\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{K}(\boldsymbol{X}, \boldsymbol{x})\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{X})]\boldsymbol{K}_\lambda^{-1}.
\end{aligned}$$

On the other hand, from the generalization error approximation bounds in (2.21), we obtain with probability at least $1 - \log^{-1}(N)$, when $N/\log^2(N) \ge C_1(1+\lambda^2)n$,

$$\begin{aligned}
\mathcal{L}(\hat{f}_\lambda^{(\mathsf{RF})}) &\ge \|P_{>\ell} f^*\|_2^2 + \sigma_{\boldsymbol{\varepsilon}}^2\operatorname{Tr}\boldsymbol{K}_\lambda^{-1}\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{K}(\boldsymbol{X}, \boldsymbol{x})\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{X})]\boldsymbol{K}_\lambda^{-1} \\
&\quad - C_2(1+\lambda)\log(N)\sqrt{\frac{n}{N}} - C_3\sqrt{n}\left\|(\boldsymbol{X}^\top\boldsymbol{x})^{\odot(\ell+1)}\right\|_2 \\
&\ge \|P_{>\ell} f^*\|_2^2 - C_2(1+\lambda)\log(N)\sqrt{\frac{n}{N}} - C_3\sqrt{n}\left\|(\boldsymbol{X}^\top\boldsymbol{x})^{\odot(\ell+1)}\right\|_2.
\end{aligned}$$

Since we can approximate $K(\boldsymbol{x}, \boldsymbol{X})$ with $K_\ell(\boldsymbol{x}, \boldsymbol{X})$, we can apply the proof of Theorem 2.12 to obtain that

$$\begin{aligned}
&\left|\operatorname{Tr}\boldsymbol{K}_\lambda^{-1}\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{K}(\boldsymbol{X}, \boldsymbol{x})\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{X})]\boldsymbol{K}_\lambda^{-1} - \mathbb{E}_{\boldsymbol{x}}\left[\boldsymbol{K}_{m,\ell}^\top\boldsymbol{K}_{\lambda,\ell}^{-2}\boldsymbol{K}_{m,\ell}\right]\right| \\
\le& \left|\mathbb{E}_{\boldsymbol{x}}\left[(\boldsymbol{K}_{m,\ell} - \boldsymbol{K}_m)^\top\boldsymbol{K}_{\lambda,\ell}^{-2}\boldsymbol{K}_{m,\ell}\right]\right| + \left|\mathbb{E}_{\boldsymbol{x}}\left[\boldsymbol{K}_m^\top\left(\boldsymbol{K}_{\lambda,\ell}^{-1} - \boldsymbol{K}_\lambda^{-1}\right)\boldsymbol{K}_{\lambda,\ell}^{-1}\boldsymbol{K}_{m,\ell}\right]\right| \\
&+ \left|\mathbb{E}_{\boldsymbol{x}}\left[\boldsymbol{K}_m^\top\boldsymbol{K}_m^{-2}\left(\boldsymbol{K}_{m,\ell} - \boldsymbol{K}_m\right)\right]\right| + \left|\mathbb{E}_{\boldsymbol{x}}\left[\boldsymbol{K}_m^\top\boldsymbol{K}_\lambda^{-1}\left(\boldsymbol{K}_{\lambda,\ell}^{-1} - \boldsymbol{K}_\lambda^{-1}\right)\boldsymbol{K}_{m,\ell}\right]\right| \\
\le& C(1+\lambda)\left\|\left(\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}}\right)^{\odot\ell+1} - \mathrm{Id}\right\|_F
\end{aligned}$$

for some constant $C > 0$ depending on $\sigma, \tau, \sigma_\varepsilon$, when in the last inequality, we exploit Proposition 2.4 and (C.39). Thus, we conclude that under the same assumptions of Theorem 2.11, with probability at least $1 - \log^{-1} N$,

$$\begin{aligned}
\mathcal{L}(\hat{f}_\lambda^{(\mathsf{RF})}) \ge& \|P_{>\ell} f^*\|_2^2 + \sigma_{\boldsymbol{\varepsilon}}^2\mathbb{E}_{\boldsymbol{x}}\left[\boldsymbol{K}_{m,\ell}^\top\boldsymbol{K}_{\lambda,\ell}^{-2}\boldsymbol{K}_{m,\ell}\right] \\
&- C(1+\lambda)\left\|\left(\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}}\right)^{\odot\ell+1} - \mathrm{Id}\right\|_F - C_2(1+\lambda)\log(N)\sqrt{\frac{n}{N}} - C_3\sqrt{n}\left\|(\boldsymbol{X}^\top\boldsymbol{x})^{\odot(\ell+1)}\right\|_2.
\end{aligned}$$

This completes the proof of the lower bound of (C.42). □