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 Gaussian initialization 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 \( f : \mathbb{R}^{d \times n} \rightarrow \mathbb{R}^n \) such that

\[
f(X) := \frac{1}{\sqrt{N}} \theta^\top \sigma(WX),
\]

where \( X \in \mathbb{R}^{d \times n} \) is the input data matrix, \( W \in \mathbb{R}^{N \times d} \) is the weight matrix for the first layer, \( \theta \in \mathbb{R}^N \) is the second layer weight, and \( \sigma \) 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 \( W \) with sufficiently large width \( N \). We will always view the input data \( X \) as a deterministic matrix (independent of the random weights in \( W \)) with certain assumptions. We fix the random matrix \( W \) and only train the second layer \( \theta \) via training data \( X \). This procedure is the same as the linear regression of random feature vectors \( \{\sigma(Wx_i) \in \mathbb{R}^N, i \in [n]\} \). The empirical CK matrix is defined by

\[
K_N := \frac{1}{N} \sigma(WX) \sigma(WX)^\top \in \mathbb{R}^{n \times n}.
\]

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

\[
K := \mathbb{E}K_N = \mathbb{E}_w[\sigma(w^\top X) \sigma(w^\top X)],
\]

under the spectral norm when width \( N \) is sufficiently large, where \( w \) is the standard normal random vector in \( \mathbb{R}^d \).
Random feature regression has already attracted as a random approximation of the reproducing kernel Hilbert space (RKHS) defined by population kernel function $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ such that

$$K(x_1, x_2) := \mathbb{E}_w[\sigma((w, x_1))\sigma((w, x_2))],$$  \hspace{1cm} (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 $K$ to represent both the $n \times n$ kernel matrix $K(X, X)$ depending on dataset $X$ and the kernel function in (1.4). Denote the output of the first layer by

$$\Phi := \sigma(WX) \in \mathbb{R}^{N \times n}. \hspace{1cm} (1.5)$$

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

Most of the recent results considered the RFRR with the data points $X$ 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); Fun 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 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 $N \approx n$, the concentration of the CMX kernel around its expectation fails, and the limiting spectrum of the CMX 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 CMX 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 $N \approx n$, the behavior of RFRR is determined by the limiting spectra of the CMX (Gerace et al., 2020; Mei and Montanari, 2019; Adam and Pennington, 2022; Chouard, 2022). Specifically, Louart et al. (2018); Liao et al. (2020); Hu and Lu (2022a); 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 $N \approx \sqrt{n \log n}$. 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 CMX 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 \( n \gg d \) 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 in High Dimensions. In general, the spectra of rotational invariant kernels have been analyzed by El Karoui (2010); Liao and Couillet (2019); Ali et al. (2021) when \( n \gg d \) 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).

## 2 MAIN RESULTS

**Notations** We use \( \text{tr}(A) = \frac{1}{\mathbb{N}} \sum_{i,j} A_{ij} \) as the normalized trace of a matrix \( A \in \mathbb{R}^{n \times n} \) and \( \text{Tr}(A) = \sum_{i,j} A_{ij} \). Denote vectors by lowercase boldface. \( \|A\| \) is the spectral norm for any matrix \( A \), \( \|A\|_F \) denotes the Frobenius norm, and \( \|x\|_2 \) is the \( \ell_2 \)-norm of any vector \( x \). 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 \( \mathbb{E}_w[\cdot] \) be the expectation with respect to the random vector \( w \).

### 2.1 Model Assumptions

Before stating our main results, we list the following assumptions for the random weights \( W \), the activation function \( \sigma \), and input data \( X \).

**Assumption 2.1.** The entries of weight matrix \( W \in \mathbb{R}^{N \times d} \) 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 \( \sigma \). 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 \( \Gamma \). Define the \( L^2(\Gamma) \) and \( L^4(\Gamma) \)-norm of \( \sigma \) by \( \|\sigma\|_2 = \left( \mathbb{E}[\sigma^2(\xi)] \right)^{1/2}, \|\sigma\|_4 = \left( \mathbb{E}[\sigma^4(\xi)] \right)^{1/4} \), 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 \( \sigma \) 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 \( X_n \in \mathbb{R}^{d_n \times n} \) with growing \( d_n \) as \( n \to \infty \), where all \( X_n \) satisfy the following assumption. Below we drop the dependence on \( n \) for the ease of notations. We treat \( X \) as a deterministic matrix under the following asymptotic condition.

**Assumption 2.3** (\( \ell \)-orthonormal dataset). Suppose that the input data \( X \in \mathbb{R}^{d \times n} \) satisfies \( \|x_i\| = 1, \forall i \in [n] \). Let \( \ell \in \mathbb{N} \) be the smallest integer such that
\[
\lim_{n \to \infty} \left\| \left( X^T X \right)^{\odot (\ell+1)} - I \right\|_F = 0. \tag{2.1}
\]
We further assume \( \sigma^2 > 0 : = \|\sigma\|_2^2 - \sum_{k=1}^{\ell} \zeta_k^2(\sigma) > 0 \).

Different from previous work that requires an upper bound on the maximal angle \( \varepsilon_n := \max_{i \neq j} |\langle x_i, x_j \rangle| \) (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,
\[
\left\| \left( X^T X \right)^{\odot (\ell+1)} - I \right\|_F \leq n \varepsilon_n^{\ell+1}, \tag{2.2}
\]
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 \( \ell \) 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 \( \{x_i\}_{i \in [n]} \) are i.i.d. uniformly distributed on \( S^{d-1} \) and \( n = \Theta(d^n) \) for \( \alpha \in \mathbb{R}_+ \), then \( \varepsilon_n = O \left( \frac{\log^{1/2} n}{d_n^{1/2}} \right) \) with high probability (see, for example, Vershynin (2018)), and we can take \( \ell = 2 \lceil \alpha \rceil \) and condition on the high probability event to make \( X \) 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 \( x_\alpha, x_\beta \) in \( X \), and any two Hermite polynomials \( h_j, h_k \), we have (Nguyen and Mondelli, 2020, Lemma D.2)
\[
\mathbb{E}_w [h_j(\langle w, x_\alpha \rangle)h_k(\langle w, x_\beta \rangle)] = \delta_{jk}(x_\alpha, x_\beta)^k. \tag{2.3}
\]
This relation also appears in Oymak and Soltanolkotabi (2020), which directly gives the following power expansion of the expected kernel \( K \) in (1.3): \( K = \)
Proposition 2.4. as an approximation of 

\[ \sum_{k=0}^{\infty} \xi_k^2(\sigma) \left( X^T X \right)^{\odot k} . \]

Hence, the kernel function \( K \) 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 \( d \gg n \). 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-\( \ell \) polynomial inner-product kernel by

\[ K_\ell := \sum_{k=0}^{\ell} \xi_k^2(\sigma) \left( X^T X \right)^{\odot k} + \sigma_{\ell+1}^2 \mathbb{I}, \tag{2.4} \]

Here \( \sigma_{\ell+1}^2 \) is an extra ridge parameter added to the polynomial kernel \( \sum_{k=0}^{\ell} \xi_k^2(\sigma) \left( X^T X \right)^{\odot k} \). 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 \( \left( X^T X \right)^{\odot k} \) become negligible when the power \( k \) is sufficiently large. Hence, we can truncate \( K \) and employ \( K_\ell \) as an approximation of \( K \) as follows.

Proposition 2.4. Under Assumptions 2.2 and 2.3, let \( n_0 \) be the smallest integer such that for all \( n \geq n_0 \), \( \max_{i \neq j} | x_i^T x_j | \leq 1/\sqrt{2} \) and

\[ \left\| \left( X^T X \right)^{\odot (\ell+1)} - \mathbb{I} \right\|_F \leq \frac{\sigma_{\ell+1}^2}{4 \| \sigma \|_4^2} . \tag{2.5} \]

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

\[ \| K_\ell - K \| \leq \sqrt{2} \| \sigma \|_4^2 \left\| \left( X^T X \right)^{\odot (\ell+1)} - \mathbb{I} \right\|_F \leq \frac{\sigma_{\ell+1}^2}{\sqrt{2}} . \tag{2.6} \]

Remark 2.5 (Comparison to previous work with random dataset). (2.6) is proved by using the inequality \( \| K_\ell - K \| \leq \| K_\ell - K \|_F \) and performing an entry-wise expansion of \( (K_\ell - K) \). 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 \( K \) were expanded in terms of an orthogonal basis with respect to the distribution of random \( X \) and \( W \). In many examples, these two distributions are assumed to be the same, which provides a convenient way to expand and approximate \( K \) with some degree-\( \ell \) 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 \( \sigma(x) = \sum_{k=0}^{\ell} \xi_k(\sigma) h_k(x) \) be an order-\( (\ell+1) \) polynomial with \( \xi_{\ell+1}(\sigma) \neq 0 \). Assume \( |(x_i, x_j)| = \varepsilon \) for all \( i \neq j \) and \( \ell \) is an odd integer. Then

\[ \| K_\ell - K \| = \frac{\xi_{\ell+1}^2(\sigma)}{\ell+1} \varepsilon^{\ell+1} (n-1) \geq \frac{1}{2} \xi_{\ell+1}^2(\sigma) \left\| \left( X^T X \right)^{\odot (\ell+1)} - \mathbb{I} \right\|_F . \]

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 \( K \) 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 \( \ell \)-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 \( K \) has also been studied by Liang et al. (2020) for the ridgeless regression on some random data \( X \) under the polynomial regime \( (n \gg d^p) \). Under a stronger regularity assumption on the kernel function, the authors first applied Taylor expansion to get truncated kernel \( K_\ell \), 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 \( \theta \in \mathbb{R}^N \) for predictors of the form \( f_\theta(X) := \frac{1}{\sqrt{n}} \theta^T \sigma(W X) \), with training data \( X \in \mathbb{R}^{d \times n} \) and training labels \( y \in \mathbb{R}^n \). The loss function of the ridge regression with a ridge parameter \( \lambda \geq 0 \) is defined by

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

The minimizer of (2.7) denoted by \( \hat{\theta} := \arg \min_{\theta} L(\theta) \) has an explicit expression \( \hat{\theta} = \frac{1}{\sqrt{N}} \Phi \left( \frac{1}{\sqrt{n}} \Phi^T \Phi + \lambda \mathbb{I} \right)^{-1} y \).
where $\Phi$ is defined in (1.5). The optimal predictor for this RFR with respect to the loss function in (2.7) is given by

$$
\hat{f}_\lambda^{(RF)}(x) := \frac{1}{\sqrt{N}}\hat{\theta}^T\sigma(Wx),
$$

where we define an empirical kernel $K_N(\cdot, \cdot): \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ as $K_N(x, z) := \frac{1}{n}\sigma(Wx)^\top\sigma(Wz)$, and the $n$-dimension row vector is given by $K_N(x, X) = [K_N(x, x_1), \ldots, K_N(x, x_n)]$.

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

$$
\hat{f}_\lambda^{(K)}(x) := K(x, X)(K + \lambda I_d)^{-1}y.
$$

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 $\hat{f}_\lambda^{(RF)}(x)$ in (2.8) and $\hat{f}_\lambda^{(K)}(x)$ in (2.9) with the kernel $K$ defined in (1.4). As $N$ is sufficiently large, the empirical kernel $K_N$ defined in (1.2) will concentrate around its expectation (1.4). From (2.8) and (2.9), the predictors of RFR and KRR are determined by $K_N$ and $K$, 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 $X$ with the ridge parameter $\lambda \geq 0$ by

$$
\hat{f}_\lambda^{(RF)}(X) := \left(\hat{f}_\lambda^{(RF)}(x_1), \ldots, \hat{f}_\lambda^{(RF)}(x_n)\right)^\top,
$$

$$
\hat{f}_\lambda^{(K)}(X) := \left(\hat{f}_\lambda^{(K)}(x_1), \ldots, \hat{f}_\lambda^{(K)}(x_n)\right)^\top,
$$

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

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

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

With high probability, the training error of a random feature model and the corresponding kernel model with the same ridge parameter $\lambda$ 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 $\lambda \geq 0$, $N/\log^2(N) > C_1n$, and $n \geq n_0$,

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

where $C_1$ and $C_2$ are positive constants depending only on $\sigma^4$ and $\lambda_0$.

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 $X$ and how we generate the labels $y$.

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 $X$ in Louart et al. (2018); Wang and Zhu (2021), a crucial property of our concentration inequality is that it does not depend on $\|X\|$, 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 $\hat{f}_\lambda^{(RF)}$ and $\hat{f}_\lambda^{(K)}$. Following Hastie et al. (2009), LOOCV is defined by

$$
\text{CV}^{(K, \lambda)}_n := \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}_\lambda^{(K)}(x_i))^2,
$$

$$
\text{CV}^{(RF, \lambda)}_n := \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}_\lambda^{(RF)}(x_i))^2,
$$

where $\hat{f}_\lambda^{(K)}$ and $\hat{f}_\lambda^{(RF)}$ are KRR and RFRR estimators, respectively, on training data set $X$ with the data point $x_i$ removed. For simplicity, denote $K_X = K + \lambda I_d$ and $K_{N, \lambda} = K_N + \lambda I_d$. With Schur complement, we can
obtain the “shortcut” formulae for LOOCV as
\[
CV_n^{(K, \lambda)} = \frac{1}{n} y^\top K_\lambda^{-1} D^{-2} K_\lambda^{-1} y, \tag{2.14}
\]
\[
CV_n^{(RF, \lambda)} = \frac{1}{n} y^\top K_{N, \lambda}^{-1} D_N^{-2} K_{N, \lambda}^{-1} y, \tag{2.15}
\]
where \( D \) and \( D_N \) are diagonal matrices with diagonals \([D]_{ii} = [K_\lambda^{-1}]_{ii}\) and \([D_N]_{ii} = [K_{N, \lambda}^{-1}]_{ii}\), 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 \([D]_{ii}\) and \([D_N]_{ii}\) to concentrate around \( \text{tr} K_\lambda^{-1} \) and \( \text{tr} K_{N, \lambda}^{-1} \) respectively. Therefore, as an approximation of LOOCV, we define GCV
\[
GCV_n^{(K, \lambda)} := \left( \lambda \text{tr}(K + \lambda \text{Id})^{-1} \right)^{-2} E_{\text{train}}^{(K, \lambda)},
\]
\[
GCV_n^{(RF, \lambda)} := \left( \lambda \text{tr}(K + \lambda \text{Id})^{-1} \right)^{-2} E_{\text{train}}^{(RF, \lambda)}.
\]

For linear ridge regression models (Hastie et al., 2022), the such approximation is done by applying random matrix theory to replace \( D_{ii} \) with \( \text{tr} K_\lambda^{-1} \) and \( \text{tr} K_{N, \lambda}^{-1} \) 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 estimators of RFRR are close to the corresponding cross-validations of KRR.

**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 \( n \geq n_0 \),
\[
\left| GCV_n^{(K, \lambda)} - GCV_n^{(RF, \lambda)} \right| \leq \frac{c(1 + \lambda^2) \log N \|y\|^2}{\sqrt{nN}}, \tag{2.17}
\]
\[
\left| CV_n^{(K, \lambda)} - CV_n^{(RF, \lambda)} \right| \leq \frac{c(1 + \lambda^2) \log N \|y\|^2}{\sqrt{nN}}, \tag{2.18}
\]
where \( C, c > 0 \) are constants depending only on \( \sigma \).

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 GCV also asymptotically approaches the generalization error of KRR when the deterministic matrix \( K(X, X) \) 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 \( x_i \) 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 \( y = f^*(x) + \varepsilon \), where \( f^*(X) = (f^*(x_1), \ldots, f^*(x_n))^\top \), and \( \varepsilon \sim \mathcal{N}(0, \sigma^2 I_d) \).

We impose the following assumptions.

**Assumption 2.9.** Assume that the target function is a nonlinear function with one neuron defined by \( f^*(x) = \tau(\langle \beta, x \rangle) \), where the random vector \( \beta \sim \mathcal{N}(0, \text{Id}_d) \) and \( \tau \in L^2(\mathbb{R}, \mathcal{G}) \). Suppose that \( \zeta_k(\tau) \neq 0 \) as long as \( k \leq \ell \). Training labels are given by \( y = \tau(x^\top \beta + \varepsilon) \in \mathbb{R}^n \).

In particular, such an assumption includes the case when \( \sigma \) and \( \tau \) are the same activation function.

**Assumption 2.10.** Suppose the test data \( x \in \mathbb{R}^d \) satisfies almost surely, \( \|x\| = 1 \) and
\[
\lim_{n \to \infty} \sqrt{n} \left\| (X^\top x)^{\odot(\ell+1)} \right\|_2 = 0. \tag{2.19}
\]

Assumption 2.10 of the test data \( x \) guarantees similar statistical behavior as the training data points in \( X \), 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 \( x_1, \ldots, x_n \) are i.i.d. samples from some dis-
Our setting, the conditional expectation in (2.20) becomes

Under Assumption 2.10, let \( \beta \) out-of-sample predictors. Viewing \( \beta \) regression (PKRR) induced by the polynomial kernel \( \mathcal{K}_\ell \) works. Notice that Theorem 2.11 covers the ridge-less regression errors of KRR and PKRR satisfy that (2.8) and (2.9), the predictor for kernel regression with respect to \( \mathcal{K}_\ell \) is denoted by

\[
\hat{f}_\ell^{(\ell)}(x) := \mathcal{K}_\ell(x, x') (\mathcal{K}_\ell + \lambda I_d)^{-1} y,
\]

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

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

\[
\Delta_\ell = \left\| \left(X^T X\right)^{\otimes(\ell+1)} - \text{Id} \right\|_F
\]

and

\[
\bar{\Delta}_\ell = \left\| \left(\hat{X}^T \hat{X}\right)^{\otimes(\ell+1)} - \text{Id} \right\|_F.
\]

Under (2.1) and (2.19), we have \( \Delta_\ell, \bar{\Delta}_\ell = o_n(1) \).

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

\[
\left| E_{\text{train}}^{(\ell, \lambda)} - E_{\text{train}}^{(\ell, \lambda)} \right| \leq \frac{C_1 \lambda^2 \|y\|^2}{n} \Delta_\ell, \quad (2.23)
\]

\[
\left| CV_n^{(\ell, \lambda)} - CV_n^{(\ell, \lambda)} \right| \leq \frac{C_1 (1 + \lambda^4) \|y\|^2}{n} \Delta_\ell, \quad (2.24)
\]

\[
\left| GCV_n^{(\ell, \lambda)} - GCV_n^{(\ell, \lambda)} \right| \leq \frac{C_1 (1 + \lambda^4) \|y\|^2}{n} \Delta_\ell, \quad (2.25)
\]

where \( C_1 > 0 \) depend only on \( \sigma \). 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

\[
\left| L_{\ell}^{(\ell)}(x) - L_{\ell}^{(\ell)}(x) \right| \leq C_2 (1 + \lambda) \Delta_\ell, \quad (2.26)
\]

for some constant \( C_2 > 0 \) only depending on \( \sigma, \tau, \sigma_e \).

Based on the definition of \( \ell \) in (2.3), if the training labels satisfy \( \|y\|^2 = O(n) \), the left-hand sides of (2.23)-(2.26) are all vanishing as \( n \to \infty \). Combining the concentration between RFR 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 $K$. For (2.26), we can actually verify the estimators $\hat{f}_K^*(x)$ and $\hat{f}_{\text{RF}}^*(x)$ are polynomials of $x$ with degree at most $\ell$, 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 distribution 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 $P_{>\ell} : L^2(\mathbb{R}, \Gamma) \rightarrow L^2(\mathbb{R}, \Gamma)$ as the projection onto the span of Hermite polynomials defined in (B.1) with degrees at least $\ell + 1$. Specifically, recalling $\beta \sim \mathcal{N}(0, \Sigma)$ and $||x|| = 1$, we can get $(P_{>\ell} f^*)(x) = \sum_{k \geq \ell+1} \zeta_k(\beta) h_k(\beta^T x)$, where $\zeta_k(\beta)$ is defined by (B.2). Denote

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

In the following theorem, we prove that the polynomial approximation barrier for RFRR is related to the $\ell$-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 that RFRR can be further approximated by a simpler polynomial kernel model, whose degree $\ell$ is determined by the $\ell$-orthonormal property in (2.3). Combining these together, RFRR induced by an ultra-wide neural network is asymptotically equivalent to an $\ell$-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 $N/\log^2 N \gg n$,

$$\mathcal{L}(\hat{f}_{\text{RF}}) \geq \|P_{>\ell} f^*\|^2_2 + \sigma_e^2 \mathbb{E}_{x} \left[ K_{m,\ell} K_{\lambda,\ell}^{-2} K_{m,\ell} \right] - o_n(1)$$

$$\geq \|P_{>\ell} f^*\|^2_2 - o_n(1),$$

(2.27)

where $K_{m,\ell} := K_{\ell}(X, X)$, $K_{\lambda,\ell} := \lambda \text{Id} + K_{\ell}(X, X)$.

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

**Remark 2.14 (The variance term).** The second term in the 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,

$$\mathbb{E}_{x} \left[ K_{\ell}(X, x) K_{\ell}(x, X) \right] K_{\ell,\lambda}^{-1} \lesssim \frac{d^n}{n} + \frac{n}{d^{\alpha-1}}$$

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, Section 4). In contrast, under the critical regime $n \approx d^n$, 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 $\|P_{>(2/5)_\alpha} f^*\|^2_2$ on the generalization error for kernel ridge regression with a general rotational invariant kernel (which is $\|P_{>(2/5)_\alpha} f^*\|^2_2$ when data $x$ has unit length), where the dataset $X \in \mathbb{R}^{d \times n}$ is random and satisfies $n \approx d^{\alpha}$. Under more general assumptions on the dataset $X$, we obtain a similar lower bound $\|P_{>(2/5)_\alpha} f^*\|^2_2$ 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.
random data $\mathbf{X}$, where each data point is sampled from $\text{Unif}(S^{d-1})$ with $d = n = 500$. As the width $N$ increases, we observe that the differences for training errors, LOOCV, GCV, and generalization errors between RFRR and KRR are all convergent with a rate of at least $1/\sqrt{N}$. The activation function is a polynomial $p(x) := h_0(x) + \frac{1}{\sqrt{N}} h_1(x) + \frac{1}{2} h_2(x) + \frac{1}{3} h_3(x) + \frac{2}{4} h_4(x) + \frac{3}{5} h_5(x)$. For KRR, we utilize the polynomial KRR $K_{\ell}$ defined by (2.4) with $\ell = 2$ for an approximation of the original $K$. 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 $\ell$-orthonormality as defined in Assumption 2.1 and its 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 $\ell$-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 FRFRR 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 FRFRR in terms of training and generalization errors and cross-validation. In addition, Hermite polynomial expansion of $\sigma$ is a universal way to precisely analyze FRFRR 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 (Miskiakiewicz, 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 FRFRR. 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


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 \( \text{Unif}(S^{d-1}) \) with \( d = 500 \) and training samples \( n = 1000 \). The activation function is a degree-5 polynomial \( p(x) = h_0(x) + \frac{1}{\sqrt{6}} h_1(x) + \frac{1}{6} h_2(x) + \frac{1}{6} h_3(x) + \frac{3}{2} h_4(x) + \frac{3}{2} h_5(x) \), where Hermite polynomials are defined in Definition B.1. As an approximation of the kernel \( \mathbf{K} \) generated by \( \sigma(x) = p(x) \), we can consider \( \mathbf{K}_2 \) defined by

\[
\mathbf{K}_2 = 11^\top + \frac{1}{6} \mathbf{X}^\top \mathbf{X} + \frac{1}{9} (\mathbf{X}^\top \mathbf{X})^\otimes 2 + \frac{26}{36} \mathbf{I}_d.
\]

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

![Figure 3: Differences between KRR and RFRR with various ridge parameters \( \lambda \) in terms of (a) training errors, (b) LOOCV errors, (c) GCV errors, and (d) generalization errors. Here, data \( \mathbf{X} \) is sampled from \( \text{Unif}(S^{d-1}) \) 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 \( \tau \) is Softplus function.](image)

B ADDITIONAL NOTATIONS AND DEFINITIONS

We denote \( \mathbf{I}_d \) as the identity matrix. Let \( \mathbf{K}_\lambda = \mathbf{K} + \lambda \mathbf{I}_d \) where \( \mathbf{K} \) is defined by (1.3) and \( \lambda \geq 0 \) is the ridge parameter. Denote \( \mathbf{K}_{N,\lambda} = \mathbf{K}_N + \lambda \mathbf{I}_d \) where \( \mathbf{K}_N \) is (1.2). Conventionally, let \( |||| \) be the \( \ell_2 \)-norm for vectors and \( \ell_2 \rightarrow \ell_2 \)
The following normalized Hermite polynomials are necessary for expanding \( \sigma \) and approximating \( K \) 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

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

(B.1)

These polynomials \( \{h_k\}_{k=0}^{\infty} \) form an orthogonal basis of \( L^2(\mathbb{R}, \Gamma) \), where \( \Gamma \) 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

\[
\zeta_k(\sigma) = \int_{-\infty}^{\infty} \sigma(x) h_k(x) \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx,
\]

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 \( \sigma \), for \( i, j \in [n] \), we have

\[
K_{ij} = \mathbb{E}_w [\sigma(w_i^\top x_i)\sigma(w_j^\top x_j)] = \sum_{k=0}^{\infty} \xi_k^2(\sigma) \langle x_i, x_j \rangle^k.
\]

Thus, we can expand this kernel as

\[
K = \sum_{k=0}^{\infty} \xi_k^2(\sigma) (X^\top X)^{\odot k}, \quad K - K_\ell = \sum_{k=\ell+1}^{\infty} \xi_k^2(\sigma) \left( (X^\top X)^{\odot k} - \text{Id} \right).
\]

Then by Cauchy’s inequality, for \( n \geq n_0 \),

\[
\|K - K_\ell\|^2 \leq \sum_{i \neq j} \left( \sum_{k=\ell+1}^{\infty} \xi_k^2(\sigma) \langle x_i, x_j \rangle^k \right)^2
\]

\[
\leq \sum_{i \neq j} \left( \sum_{k=\ell+1}^{\infty} \xi_k^2(\sigma) \right) \left( \sum_{k=\ell+1}^{\infty} \langle x_i, x_j \rangle^{2k} \right)
\]

\[
\leq \|\sigma\|_4^4 \sum_{i \neq j} \frac{\langle x_i, x_j \rangle^{2\ell+2}}{1 - \max \|\langle x_i, x_j \rangle\|^2}
\]

\[
\leq 2 \|\sigma\|_4^4 \left\| (X^\top X)^{\odot \ell+1} - \text{Id} \right\|_F^2.
\]

Therefore, from (2.5),

\[
\|K - K_\ell\| \leq \sqrt{2} \|\sigma\|_4^4 \left\| (X^\top X)^{\odot \ell+1} - \text{Id} \right\|_F \leq \frac{1}{2\sqrt{2}} \sigma^2 > \ell.
\]
Since
\[
(X^\top X)_{ij}^{\otimes k} = \langle x_i, x_j \rangle^k = \langle x_i \otimes \cdots \otimes x_i, x_j \otimes \cdots \otimes x_j \rangle,
\]
where \(x_i \otimes \cdots \otimes x_i\) is the \(k\)-th tensor product of \(x_i\), \((X^\top X)_{ij}^{\otimes k}\) is positive semidefinite. Then
\[
\lambda_{\min}(K) \geq \lambda_{\min}(K_\ell) - \|K - K_\ell\| \geq \sigma_{\geq \ell}^2 - 1/2\sigma_{\geq \ell}^2 > 1/2\sigma_{\geq \ell}^2 \equiv \lambda_0,
\]
Notice that \(\lambda_0 > 0\) because \(\sigma_{\geq \ell}^2 > 0\) from Assumption 2.3.

### C.2 Concentration Inequality for Normalized Random Kernel Matrices

Now we introduce the concentration inequality for \(K_N\) 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 \(X\) 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 \(X\) and only assume \(\|\sigma\|_4\) is finite. To consider a normalized version of the kernel matrices, we need to consider \(K_\lambda^{-1}\). Under Assumption 2.3, we use Proposition 2.4 to make sure \(K_\lambda\) 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 exist some positive constants \(C_1, C_2 > 0\) depending on \(\sigma\), such that for any \(N\) satisfying \(N/\log^2(N) > C_1 n\) and any \(\lambda \geq 0, n \geq n_0\), we have
\[
\left\|K_\lambda^{-1/2} (K_N - K) K_\lambda^{-1/2}\right\| \leq C_2 \log(n) \sqrt{\frac{n}{N}},
\]
with probability at least \(1 - N^{-2}\), where \(K_\lambda = K + \lambda \text{Id}\).

**Proof.** Denote \(\tilde{\sigma}(x) := \sigma(x) 1_{|x| \leq B}\), where \(B\) is a parameter to be decided later. Define
\[
K_N = \frac{1}{N} \sum_{i=1}^N \sigma(w_i^\top X)^\top \sigma(w_i^\top X), \quad K = \mathbb{E}_{w}[\sigma(w^\top X)^\top \sigma(w^\top X)],
\]
\[
\tilde{K}_N = \frac{1}{N} \sum_{i=1}^N \tilde{\sigma}(w_i^\top X)^\top \tilde{\sigma}(w_i^\top X), \quad \tilde{K} = \mathbb{E}_{w}[\tilde{\sigma}(w^\top X)^\top \tilde{\sigma}(w^\top X)].
\]
For simplicity, we denote \(\tilde{K}_\lambda := \tilde{K} + \lambda \text{Id}\). Define
\[
H_i := \frac{1}{N} \tilde{K}_\lambda^{-1/2} \tilde{\sigma}(w_i^\top X)^\top \tilde{\sigma}(w_i^\top X) \tilde{K}_\lambda^{-1/2}.
\]
Notice that Proposition 2.4 implies that \(\|K_\lambda^{-1}\| \leq \lambda_0^{-1}\) for \(\lambda \geq 0\). Firstly, based on the truncated function \(\tilde{\sigma}(x)\), we have that for some universal constant \(c > 0\),
\[
\mathbb{P}(K_N \neq \tilde{K}_N) \leq \mathbb{P} \left( \max_{i \in [N], k \in [n]} |w_i^\top x_k| > B \right) \leq N n \mathbb{P}(|\xi| > B) \leq c n \exp(-B^2/2),
\]
where \(\xi \sim N(0, 1)\). Define the event by \(A_i := \{w : |w^\top x_i| \leq B\}\) for \(i \in [n]\). Entry-wisely, we have
\[
\left\|K - \tilde{K}\right\|_{i,j} = \mathbb{E}_w[\sigma(w^\top x_i)\sigma(w^\top x_j)1_{A_i \cap A_j^c}] \leq \mathbb{E}_w[\sigma(w^\top x_i)^{1/2}\mathbb{E}[1_{A_i \cap A_j^c}]^{1/2}] \leq \sqrt{2}\mathbb{E}[\sigma(\xi)^{1/2}P(A_j^c)]^{1/2} \leq C_0 \exp(-B^2/4),
\]
for some constant \(C_0 > 0\) which only depends on \(\|\sigma\|_4\). Therefore, \(\|K - \tilde{K}\| \leq \|K - \tilde{K}\|_F \leq C_0 n \exp(-B^2/4)\) and
\[
\left\|K_\lambda^{-1/2} (K - \tilde{K}) K_\lambda^{-1/2}\right\| \leq \frac{C_0}{\lambda_0} n \exp(-B^2/4).
\]
For
\[ B \geq \sqrt{4 \log \left( \frac{2Cn}{\lambda_0} \right)}, \quad \text{(C.4)} \]
the above equation also implies that \( \| K - \tilde{K} \| \leq \frac{\lambda_0}{2} \), and
\[
\left\| \tilde{K}^{1/2} K^{-1/2}_\lambda \right\|^2 = \left\| K^{-1/2}_\lambda \tilde{K} K^{-1/2}_\lambda \right\| \\
\leq \left\| K^{-1/2}_\lambda (K - \tilde{K}) K^{-1/2}_\lambda \right\| + \left\| K^{-1/2}_\lambda \tilde{K} K^{-1/2}_\lambda \right\| \leq \frac{3}{2}. \quad \text{(C.5)}
\]
Therefore, the smallest eigenvalues of \( K_\lambda \) and \( \tilde{K}_\lambda \) satisfy
\[ \lambda_{\min}(\tilde{K}_\lambda) \geq \lambda_{\min}(K_\lambda) - \| K - \tilde{K} \| \geq \frac{\lambda_0}{2} > 0. \quad \text{(C.6)} \]
It suffices to analyze \( \left\| K^{-1/2}_\lambda (\tilde{K}_N - K) K^{-1/2}_\lambda \right\| \) because of (C.2) and the following equation:
\[ P \left( \left\| K^{-1/2}_\lambda (\tilde{K}_N - K) K^{-1/2}_\lambda \right\| \geq t \right) \leq P \left( \left\| K^{-1/2}_\lambda (\tilde{K}_N - K) K^{-1/2}_\lambda \right\| \geq t \right) + P \left( K_N \neq \tilde{K}_N \right). \quad \text{(C.7)} \]
Meanwhile, by (C.3), (C.4), and (C.5), we know that
\[
\left\| K^{-1/2}_\lambda (\tilde{K}_N - K) K^{-1/2}_\lambda \right\| \leq \left\| K^{-1/2}_\lambda (\tilde{K}_N - K) K^{-1/2}_\lambda \right\| + \left\| K^{-1/2}_\lambda (\tilde{K} - K) K^{-1/2}_\lambda \right\| \\
\leq \left\| K^{-1/2}_\lambda K^{-1/2}_\lambda \right\|^2 \left\| K^{-1/2}_\lambda (\tilde{K}_N - K) K^{-1/2}_\lambda \right\| + \left\| K^{-1/2}_\lambda (\tilde{K} - K) K^{-1/2}_\lambda \right\| \\
\leq \frac{3}{2} \left\| K^{-1/2}_\lambda (\tilde{K}_N - \tilde{K}) K^{-1/2}_\lambda \right\| + \frac{Cn}{\lambda_0} \exp(-B^2/4). \quad \text{(C.8)}
\]
Hence, we only need to prove the concentration inequality for \( \tilde{K}^{-1/2}_\lambda (\tilde{K}_N - \tilde{K}) \tilde{K}^{-1/2}_\lambda = \sum_{i=1}^{N} H_i - \mathbb{E}H_i \). In terms of the definition of \( \tilde{\sigma} \) and (C.6), we know that, almost surely,
\[ \| H_i - \mathbb{E}H_i \| \leq 2 \| H_i \| \leq \frac{4}{\lambda_0 N} \| \tilde{\sigma}(w_i^\top X) \|^2 \leq \frac{4B^2 n}{\lambda_0 N}, \]
where we take expectation with respect to \( w_i \). Analogously, applying (C.6), we have
\[
H_i^2 = \frac{1}{N^2} \tilde{K}^{-1/2}_\lambda \tilde{\sigma}(w_i^\top X)^\top \tilde{\sigma}(w_i^\top X) \tilde{K}^{-1/2}_\lambda \tilde{\sigma}(w_i^\top X)^\top \tilde{\sigma}(w_i^\top X) \tilde{K}^{-1/2}_\lambda \\
\leq \frac{2}{\lambda_0 N^2} \tilde{K}^{-1/2}_\lambda \tilde{\sigma}(w_i^\top X)^\top \tilde{\sigma}(w_i^\top X) \tilde{K}^{-1/2}_\lambda \\
\leq \frac{2B^2 n}{\lambda_0 N^2} \tilde{K}^{-1/2}_\lambda \tilde{\sigma}(w_i^\top X)^\top \tilde{\sigma}(w_i^\top X) \tilde{K}^{-1/2}_\lambda. \]
Notice that \( \mathbb{E}[\tilde{\sigma}(w_i^\top X)^\top \tilde{\sigma}(w_i^\top X)] = \tilde{K} \). Hence, \( \mathbb{E}[\tilde{K}^{-1/2}_\lambda \tilde{\sigma}(w_i^\top X)^\top \tilde{\sigma}(w_i^\top X) \tilde{K}^{-1/2}_\lambda] = \frac{1}{\lambda\lambda_\lambda} \text{Id} \), and
\[ \mathbb{E}[\| H_i - \mathbb{E}H_i \|^2] \leq \mathbb{E}H_i^2 \leq \frac{2B^2 n}{\lambda_0 N^2} \text{Id}. \]
Thus, applying Theorem 5.4.1 of (Vershynin, 2018), we obtain
\[ P \left( \sum_{i=1}^{N} H_i - \mathbb{E}H_i \right. \left. > t \right) \leq 2n \exp \left( -\frac{t^2/2}{\nu + at/3} \right). \quad \text{(C.9)} \]
where \( v \leq \frac{2B^2 n}{\lambda_0 N^2} \) and \( a = \frac{4B^2 n}{\lambda_0 N^2} \). Take \( t = B^2 \sqrt{n/\lambda N} \) and \( B = C \sqrt{\log \lambda N} \). Then for \( \lambda N \geq B^4 n \), 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
\[ P \left( \left\| \tilde{K}^{-1/2}_\lambda (\tilde{K}_N - \tilde{K}) \tilde{K}^{-1/2}_\lambda \right\| > C'' \log(N) \sqrt{\frac{n}{N}} \right) \leq \frac{1}{2} N^{-2}, \quad \text{(C.10)} \]
for sufficiently large $C'$. Here both $C', C'' > 0$ are determined by $\lambda_0$. Notice that for all large $N$, the second term of (C.8) can be also bounded by $C'' \log(N) \sqrt{\frac{p}{N}}$ for some constant $C'' > 0$ only depending on $\|\sigma\|_4$ and $\lambda_0$. Combining (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 $n \geq n_0$,

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

as desired, where $C_1, C_2$ are constants depending only on $\|\sigma\|_4$ and $\lambda_0$. \hfill \Box

### 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 $t = C_1 \log(N) \sqrt{\frac{p}{N}}$. 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:

$$ \left\| (K_N + \lambda \text{Id})^{-1/2} (K - K_N) (K_N + \lambda \text{Id})^{-1/2} \right\| \leq t, $$

$$ \left\| K^{-1/2}_\lambda (K_N + \lambda \text{Id})^{1/2} \right\| \leq \sqrt{1 + t}, $$

$$ \left\| K^{1/2}_\lambda (K_N + \lambda \text{Id})^{-1/2} \right\| \leq (1 - t)^{-1/2}, $$

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

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

$$ (K_N - K) \ll t K_\lambda, $$

$$ (K - K_N) \ll t K_\lambda, $$

$$ (K - K_N) \ll \frac{t}{1 - t} (K_N + \lambda \text{Id}), $$

$$ (K_N + \lambda \text{Id}) \ll (1 + t) K_\lambda, $$

$$ K_\lambda \ll \frac{1}{1 - t} (K_N + \lambda \text{Id}), \quad (C.11) $$

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

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 $K(X, X)$ and $K_N(X, X)$ 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 $n \geq n_0$,

$$ \left| E_{\text{train}}^{(RF, \lambda)} - E_{\text{train}}^{(K, \lambda)} \right| = \frac{\lambda^2}{n} \left| \text{Tr}[(K + \lambda \text{Id})^{-2} y y^T] - \text{Tr}[(K_N + \lambda \text{Id})^{-2} y y^T] \right| $$

$$ = \frac{\lambda^2}{n} \left| [y]^T [(K + \lambda \text{Id})^{-2} - (K_N + \lambda \text{Id})^{-2}] y \right| $$

$$ \leq \frac{\lambda^2}{n} \left\| (K + \lambda \text{Id})^{-2} - (K_N + \lambda \text{Id})^{-2} \right\| \cdot \|y\|^2 $$

$$ \leq \frac{\lambda^2}{n} \left\| y \right\|^2 \| (K + \lambda \text{Id})^{-1} - (K_N + \lambda \text{Id})^{-1} \| \| (K + \lambda \text{Id})^{-1} \| + \| (K_N + \lambda \text{Id})^{-1} \| $$

$$ \leq \frac{5\lambda^2}{\lambda_0 n} \left\| (K + \lambda \text{Id})^{-1} - (K_N + \lambda \text{Id})^{-1} \right\|, \quad (C.12) $$

where in the last line, we employ the fact that $\| (K_N + \lambda \text{Id})^{-1} \| \leq 4\lambda_0^{-1}$ and $\| (K + \lambda \text{Id})^{-1} \| \leq \lambda_0^{-1}$ from Corollary C.2 and Proposition 2.4, respectively.
Take \( C_2 = \sqrt{2} C_1 \) in Proposition C.1. For any \( N \) satisfying \( N / \log^2(N) > 2 C_1^2 n \), we can make \( 0 < t < 3/4 \), where \( t = C_1 \log(N) \sqrt{\frac{2}{N}} \). 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
\[
\| (K + \lambda \text{Id})^{-1} - (K_N + \lambda \text{Id})^{-1} \| \\
= \| (K_\lambda)^{-1/2}(K_N - K)(K_\lambda)^{-1/2}(K_N + \lambda \text{Id})^{-1/2}(K_N + \lambda \text{Id})^{-1/2} \| \\
\leq \frac{1}{2 \lambda_0} \| (K_\lambda)^{-1/2}(K_N - K)(K_\lambda)^{-1/2} \| \| (K_N + \lambda \text{Id})^{-1/2} \| \\
\leq \frac{t}{2 \lambda_0 \sqrt{1 - t}} \leq \frac{t}{\lambda_0}.
\]

Hence, from (C.12), we get
\[
\| E^{(RF, \lambda)}_{\text{train}} - E^{(K, \lambda)}_{\text{train}} \| \leq \frac{5 \lambda^2 \| y \|^2_t}{\lambda_0 n},
\]
which finishes the proof of Theorem 2.7.

\[\square\]

### C.4 Proof of Theorem 2.8

We start with the following estimate on the normalized trace \( \text{tr} K_{\lambda}^{-1} \).

**Lemma C.3.** Under Assumption 2.2, we have \( \text{tr} K_{\lambda} = \lambda + \| \sigma \|^2_2 \) and when \( n \geq n_0 \),
\[
(\lambda + \| \sigma \|^2_2)^{-1} \leq \text{tr} K_{\lambda}^{-1} \leq \lambda_0^{-1}.
\]

*Proof.* By definition of \( K \), we know \( \text{Tr}[K] = n \mathbb{E}_{w}[\sigma(w^T x)^2] = n \mathbb{E}[\sigma(\xi)^2] = n \| \sigma \|^2_2 \) for \( \xi \sim \mathcal{N}(0, 1) \). Hence, \( \text{Tr} K_{\lambda} = n \left( \lambda + \| \sigma \|^2_2 \right) \). Denote \( \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0 \) as the eigenvalues of \( K \). Then, by Cauchy–Schwartz inequality, we have
\[
n = \sum_{i=1}^{n} \frac{1}{\sqrt{\lambda_i + \lambda}} \sqrt{\lambda_i + \lambda} \leq \left( \text{Tr} K_{\lambda}^{-1} \right)^{1/2} \left( \text{Tr} K_{\lambda} \right)^{1/2}.
\]
Therefore, we can get \( (\lambda + \| \sigma \|^2_2)^{-1} \leq \text{tr} K_{\lambda}^{-1} \). Meanwhile, based on Proposition 2.4, \( \text{tr} K_{\lambda}^{-1} \leq \| K_{\lambda}^{-1} \| \leq \lambda_{\text{min}}^{-1}(K) \leq \lambda_0^{-1} \). Notice that \( \lambda_0 = \frac{1}{2} \sigma_{\max}^2 \leq \| \sigma \|^2_2 \).

Recall that \( K_{N,\lambda} = K_N + \lambda \text{Id} \). The following lemma for the approximation of \( K_{\lambda} \) with \( K_{N,\lambda} \) 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 \( n \geq n_0 \),
\[
\| \text{tr} K_{\lambda}^{-1} - \text{tr} K_{N,\lambda}^{-1} \| \leq C_0 \log(N) \sqrt{n},
\]
and
\[
\frac{1}{2(\lambda + \| \sigma \|^2_2)} \leq \text{tr} K_{N,\lambda}^{-1} \leq \frac{3}{2 \lambda_0},
\]
where constants \( C, C_0 > 0 \) only depends on \( \lambda_0 \) and \( \| \sigma \|^4_4 \).

*Proof.* From (C.13) and Lemma C.3, by taking \( t = C_1 \log(N) \sqrt{\frac{2}{N}} \in (0, 1) \), we have
\[
\left| \text{tr} K_{\lambda}^{-1} - \text{tr} K_{N,\lambda}^{-1} \right| \leq \left| K_{\lambda}^{-1} - K_{N,\lambda}^{-1} \right| \leq \frac{t}{\lambda_0(1 - t)},
\]
\[
(\lambda + \| \sigma \|^2_2)^{-1} - \frac{t}{\lambda_0(1 - t)} \leq \text{tr} K_{N,\lambda}^{-1} \leq \lambda_0^{-1} + \frac{t}{\lambda_0(1 - t)}.
\]

(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 $0 \leq t \leq \min\{1/2, \lambda_0/4(\lambda + \|\sigma\|_2^2)\}$. Then,
\[
\frac{t}{\lambda_0(1-t)} \leq \frac{1}{2(\lambda + \|\sigma\|_2^2)} \leq \frac{1}{2\lambda_0}. 
\] (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 $\lambda \geq 0$,
\[
CV_n^{(K,\lambda)} = \frac{1}{n} y^\top K^{-1}_\lambda D^{-2} K^{-1}_\lambda y, \\
CV_n^{(RF,\lambda)} = \frac{1}{n} y^\top K_{N,\lambda}^{-1} D_N^{-2} K_{N,\lambda}^{-1} y,
\]
where $D$ and $D_N$ are diagonal matrices with diagonals $[D]_{ii} = [K^{-1}_\lambda]_{ii}$ and $[D_N]_{ii} = [K_{N,\lambda}^{-1}]_{ii}$, for $i \in [n]$, respectively. When $n \geq n_0$, we have
\[
(\lambda + \|\sigma\|_2^2)^{-1} \leq \|D\| \leq \lambda_0^{-1}. 
\] (C.17)

Additionally, for a constant $C > 0$ depending on $\lambda_0, \|\sigma\|_2$, when $N/\log^2(N) > C(1 + \lambda^2)n$,
\[
\frac{1}{2}(\lambda + \|\sigma\|_2^2)^{-1} \leq \|D_N\| \leq 2\lambda_0^{-1}, \\
\|D^{-2} - D_N^{-2}\| \leq C_0(1 + \lambda^4) \log(N) \sqrt{\frac{n}{N}},
\] (C.18) (C.19)

with probability at least $1 - N^{-2}$, for some constant $C_0 > 0$ which only depends on $\lambda_0$ and $\|\sigma\|_2$.

**Proof.** For $i \in [n]$, denote $y^{-i} \in \mathbb{R}^{n-1}$ by the vector $y$ with the $i$-th entry removed, $X^{-i}$ by the data $X$ with the $i$-th column removed, and $K_{-i,\lambda}$ by the matrix $K_{\lambda}$ 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 $j \neq i$, the $(i, j)$ entry of $K_\lambda^{-1}$ is given by
\[
[K^{-1}_\lambda]_{i,j} = -([K^{-1}_\lambda]_{ii}) \sum_{k \neq i} [K]_{i,k} [K^{-1}_{-i,\lambda}]_{k,j}. 
\] (C.20)

Thus, from definition (2.13), we can exploit (C.20) to obtain
\[
y_i - \hat{f}^{(K)}_{\lambda, \cdot}(x_i) = y_i - K(x_i, X^{-i}) K_{-i,\lambda}^{-1} y^{-i} = y_i + \left[\frac{[K^{-1}_\lambda]_{i,i}}{[K^{-1}_\lambda]_{ii}}\right] y^{-i} = \left[\frac{[K^{-1}_\lambda]_{i,i}}{[K^{-1}_\lambda]_{ii}}\right] y^{-i} + \left[\frac{[K^{-1}_\lambda]_{i,i}}{[K^{-1}_\lambda]_{ii}}\right] (y_i - y),
\]
for any $i \in [n]$, where $[K^{-1}_\lambda]_{i,i}$ is the $i$-th row of $K^{-1}_\lambda$ with the $i$-th entry removed, and $[K^{-1}_\lambda]_{i,i}$ is the $i$-th row of $K^{-1}_\lambda$. Hence, in matrix form, we can get
\[
CV_n^{(K,\lambda)} = \frac{1}{n} \sum_{i=1}^{n} y^\top \left[\frac{[K^{-1}_\lambda]_{i,i}}{[K^{-1}_\lambda]_{ii}}\right] [K^{-1}_\lambda]_{i,i} y = \frac{1}{n} y^\top K^{-1}_\lambda D^{-2} K^{-1}_\lambda y.
\]

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

Secondly, applying Theorem A.4 of (Bai and Silverstein, 2010), we have
\[
[D]_{ii} = [K^{-1}_\lambda]_{ii} = \frac{1}{[K]_{ii} - K(x_i, X^{-i}) K_{-i,\lambda}^{-1} K(X^{-i}, x_i)},
\]
for any $i \in [n]$. Recall that, in the proof of Lemma C.3, we have shown $[K]_{ii} = \lambda + \|\sigma\|_2^2$ for all $i \in [n]$. Therefore, we have
\[
(\lambda + \|\sigma\|_2^2)^{-1} \leq [K^{-1}_\lambda]_{ii} \leq \|K^{-1}_\lambda\| \leq \lambda_0^{-1}, \quad \forall i \in [n].
\]
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 \( t = C_1 \log(N)/\sqrt{n/N} \leq \min\{1/2, \lambda_0/4(\lambda + \| \sigma \|_2^2)\} \) and \( N/\log^2(N) > (1 + \lambda^2)n \), we have

\[
\| D - D_N \| \leq \| K_\lambda^{-1} - K_{N,\lambda}^{-1} \| \leq \frac{t}{\lambda_0(1-t)},
\]

(C.21)

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

\[
\frac{1}{2}(\lambda + \| \sigma \|_2^2)^{-1} \leq (\lambda + \| \sigma \|_2^2)^{-1} - \| D - D_N \| \leq \| D_N \| \leq \lambda_0^{-1} + \| D - D_N \| \leq 2\lambda_0^{-1}.
\]

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

\[
\| D^{-2} - D_N^{-2} \| \leq \| D \|^{-2} \| D_N \|^{-2} (\| D_N \| + \| D \|) \| D - D_N \|
\]

\[
\leq 12\frac{(\lambda + \| \sigma \|_2^2)^4}{\lambda^3} t \leq C_0(1 + \lambda^4) \log(N) \sqrt{\frac{n}{N}}.
\]

This finally completes the proof of this lemma. \( \square \)

**Proof of Theorem 2.8.** We start with (2.17). Recall (2.10), (2.11), and \( K_{N,\lambda} = K_N + \lambda \text{Id} \). Using the expression (2.16), we have

\[
|GCV_n^{(K,\lambda)} - GCV_n^{(RF,\lambda)}| \leq \frac{1}{\lambda^2} \left( (\text{tr} K^{-1})^{-2} - (\text{tr} K_{N,\lambda}^{-1})^{-2} \right) E_{\text{train}}^{(K,\lambda)} + \frac{1}{\lambda^2 (\text{tr} K^{-1})^2} E_{\text{train}}^{(K,\lambda)} - E_{\text{train}}^{(RF,\lambda)}
\]

\[
\leq \frac{1}{n} \| K_\lambda^{-1} y \|^2 \left( (\text{tr} K^{-1})^{-2} - (\text{tr} K_{N,\lambda}^{-1})^{-2} \right)
\]

\[
+ C_2(\lambda + \| \sigma \|_2^2)^2 \log N \| y \|^2 \sqrt{\frac{n}{N}},
\]

(C.22)

where (C.23) is due to (2.12) and Lemma C.3, and \( C_2 \) is a constant depending on \( \| \sigma \|_2 \) and \( \lambda_0 \).

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

\[
\frac{1}{n} \| K_\lambda^{-1} y \|^2 \left( (\text{tr} K^{-1})^{-2} - (\text{tr} K_{N,\lambda}^{-1})^{-2} \right)
\]

\[
\leq \left( \text{tr} K^{-1} \right)^{-2} \left( \text{tr} K_{N,\lambda}^{-1} \right)^{-2} \text{tr} (K_\lambda^{-1} - K_{N,\lambda}^{-1}) \text{tr} (K_\lambda^{-1} + K_{N,\lambda}^{-1}) \frac{1}{n} \| K_\lambda^{-2} \| \| y \|^2
\]

\[
\leq \frac{20(\lambda + \| \sigma \|_2^2)^4}{\lambda^3 n} \| y \|^2 C_0 \log(N) \sqrt{\frac{n}{N}} \leq C_1(1 + \lambda^4) \| y \|^2 \log(N) \sqrt{\frac{n}{N}},
\]

for some constant \( C_1 \) which only relies on \( \| \sigma \|_2 \) and \( \lambda_0 \). Hence, the bounds of (C.22) and (C.23) imply

\[
|GCV_n^{(K,\lambda)} - GCV_n^{(RF,\lambda)}| \leq \frac{c(1 + \lambda^4) \log N \| y \|^2}{\sqrt{nN}}
\]

for a constant \( c \) depending on \( \lambda_0, \| \sigma \|_2 \), 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

\[
|CV_n^{(K,\lambda)} - CV_n^{(RF,\lambda)}| \leq \frac{\| y \|^2}{n} \left( (K_\lambda^{-1} - K_{N,\lambda}^{-1}) D^{-2} K_\lambda^{-1} \right) + \frac{\| y \|^2}{n} \left( K_{N,\lambda}^{-1} (D^{-2} - D_N^{-2}) K_\lambda^{-1} \right)
\]

\[
+ \frac{\| y \|^2}{n} \left( K_{N,\lambda}^{-1} D_N^{-2} (K_\lambda^{-1} - K_{N,\lambda}^{-1}) \right) \leq \frac{c(1 + \lambda^2) \log N \| y \|^2}{\sqrt{nN}}, \quad (C.24)
\]

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. \( \square \)
C.5 Proof of Theorem 2.11

For simplicity, we denote \( K_{N,\lambda} = (K_N + \lambda \text{Id}), K_\lambda = (K + \lambda \text{Id}), K_{m,N} := K_N(x) \in \mathbb{R}^n \) and \( K_m := K(X,x) \in \mathbb{R}^n \). Define
\[
K^{(2)}_N := \mathbb{E}_x[K_N(X,x)K_N(x,X)], \quad K^{(2)} := \mathbb{E}_x[K(X,x)K(x,X)],
\]
where we take expectation with respect to test data \( x \) defined in Assumption 2.10. Recalling Assumption 2.9, we have
\[
f^*(x) = \tau((\beta, x)), \quad y = \tau(X^\top \beta) + \epsilon,
\]
where \( \beta \sim \mathcal{N}(0, \text{Id}) \). Denote the kernel given by \( \tau \) as \( \Psi := \mathbb{E}_{\|\beta\|}[\tau(X^\top \beta)\tau(\beta^\top X)] \) and the vector by \( u := \mathbb{E}_{\|\beta\|}[\tau(X^\top \beta)f^*(x)] \in \mathbb{R}^n \).

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

**Lemma C.6.** There exist some constant \( C > 0 \) depending on \( \lambda_0, \|\sigma\| \) such that, with probability at least \( 1 - N^{-1} \), when \( n \geq \max\{n_0, n_1\} \),
\[
K^{\top}_{m,N}K^{-1}_{N,\lambda}K_{m,N} < \|\sigma\|^2 + \|\sigma\|^4 + \lambda,
\]
when \( N/\log^2(N) > C(1 + \lambda^2)n \). Moreover, we have
\[
K^{\top}_mK^{-1}_m < \|\sigma\|^2 + \lambda.
\]

**Proof.** Consider an enlarged block matrix \( \tilde{K} \in \mathbb{R}^{(n+1)\times(n+1)} \) defined by
\[
\tilde{K} := \begin{pmatrix} K & K_m \\ K_m^\top & K(x,x) \end{pmatrix},
\]
(25)
where \( K(x,x) = \mathbb{E}[\sigma(w^\top x)(\sigma(w^\top x)] = \|\sigma\|^2_2 \). Let \( \tilde{K}_\lambda := \tilde{K} + \lambda \text{Id} \). Analogously to (2.4), let us define
\[
\tilde{K}_{\ell} := \sum_{k=0}^{\ell} c_k^2(\sigma)(\tilde{X}^\top \tilde{X}) \bigoplus^{k} + \sigma^2_{\ell} \text{Id} \in \mathbb{R}^{(n+1)\times(n+1)},
\]
(26)
where \( \tilde{X} = [X, x] \in \mathbb{R}^{d\times(n+1)} \) is the concatenation of training and test data points. By Assumption 2.10, analogously to the proof of Proposition 2.4, we have
\[
\left\| \tilde{K}_{\ell} - \tilde{K} \right\|^2 
\leq 2\|\sigma\|^4_4 \left\| (\tilde{X}^\top \tilde{X}) \bigoplus^{\ell+1} - \text{Id} \right\|^2_F
\leq 2\|\sigma\|^4_4 \left\| (\tilde{X}^\top \tilde{X}) \bigoplus^{\ell+1} - \text{Id} \right\|^2_F + 4\|\sigma\|^4_4 \left\| (\tilde{X}^\top x)^\bigoplus^{(\ell+1)} \right\|^2_2 
\leq \frac{3}{8} \lambda_0^2.
\]
(27)
Since \( \lambda_{\text{min}}(\tilde{K}) \geq \lambda_0 > 0 \), we have \( \lambda_{\text{min}}(\tilde{K}_\lambda) \geq \frac{1}{2} \lambda_0 \) and \( \tilde{K}_\lambda \) is positive definite for any \( \lambda \geq 0 \). By Theorem 7.7.7 of Horn and Johnson (2012), since both \( K_\lambda \) and \( \tilde{K}_\lambda \) are positive definite, the Schur complement of \( \tilde{K}_\lambda \) given by \( K(x,x) + \lambda - K_m^{-1}K_m \) is positive, which concludes our second result in this lemma.

Similarly, consider the block matrix \( \hat{K}_N \in \mathbb{R}^{(n+1)\times(n+1)} \) defined by
\[
\hat{K}_N := K_N(\tilde{X}, \tilde{X}) = \begin{pmatrix} K_N & K_{m,N} \\ K_{m,N}^\top & K_{N}(x,x) \end{pmatrix}.
\]
Let \( \hat{K}_{N,\lambda} := \lambda \text{Id} + \hat{K}_N \). Combining Assumption 2.10 and (C.27), we can easily ensure Proposition C.1 and Corollary C.2 still hold for \( \hat{K}_\lambda \) and \( \hat{K}_{N,\lambda} \). Therefore, with probability at least \( 1 - N^{-2} \), \( \lambda_{\text{min}}(\hat{K}_{N,\lambda}) \geq \frac{1}{2} \lambda_0 \), for sufficiently large constant \( C > 0 \) with \( N/\log^2(N) > C(1 + \lambda^2)n \), which implies that \( \hat{K}_{N,\lambda} \) is positive definite with probability \( 1 - N^{-2} \).
Again, from Theorem 7.7.7 of Horn and Johnson (2012), we can get $K_N(x, x) + \lambda - K_{m,N}^{-1}K_{m,N} > 0$ with probability at least $1 - N^{-2}$. Thus,

$$0 \leq K_{m,N}^{-1}K_{m,N} < K_N(x, x) + \lambda.$$ 

Notice that $K_N(x, x) = \frac{1}{N} \sum_{i=1}^{N} \sigma(w_i^T x)^2$, $\mathbb{E}[K_N(x, x)] = \|\sigma\|^2_2$, and

$$\mathbb{E} \left( K_N(x, x) - \|\sigma\|^2_2 \right)^2 = \frac{1}{N} \text{Var}(\sigma(w^T x)^2) = \frac{\|\sigma\|^4 - \|\sigma\|^2_2}{N},$$

where $w \sim \mathcal{N}(0, \text{Id})$. Therefore, by Markov inequality, we conclude that, with probability at least $1 - 1/N$, $K_N(x, x) \leq \|\sigma\|^2_2 + \|\sigma\|^2_2$. Therefore, with probability at least $1 - N^{-1}$, we have $K_{m,N}^{-1}K_{m,N} < \|\sigma\|^2_2 + \|\sigma\|^2_2 + \lambda$. \hfill $\square$

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

$$\|K_{\lambda}^{-1/2}\Psi K_{\lambda}^{-1/2}\| \leq C,$$

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

$$\Psi_{\ell} := \sum_{k=0}^{\ell} \zeta_k^2(\tau) \left( X^T X \right)^{\otimes k} + \tau^2_{\ell} \text{Id}.$$ \hfill (C.28)

Define $K_{\ell,\lambda} := K_{\ell} + \lambda \text{Id}$. By the assumption and definition of $K_{\ell}$, there exists some constant $C > 0$ such that $\Psi_{\ell} \leq C K_{\ell,\lambda}$ for any $\lambda \geq 0$. Here this constant $C$ only relies on the Hermite coefficients $\zeta_k(\tau)$, $\lambda_0$ and $\zeta_k(\sigma)$ for $0 \leq k \leq \ell$. Next, applying Proposition 2.4 for nonlinear function $\tau$, we have

$$\|\Psi - \Psi_{\ell}\| \leq \sqrt{2} \|\tau\|^2_4 \left\|X^T X\right\|^{\otimes \ell+1} - \text{Id} \| \leq \frac{\sqrt{2} \|\tau\|^2_4}{4 \|\tau\|^2_4}.$$ \hfill (C.29)

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

$$K_{\lambda}^{-1/2}K_{\ell,\lambda}K_{\lambda}^{-1/2} \leq \frac{3}{2} \text{Id},$$

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

$$\|K_{\lambda}^{-1/2}\Psi K_{\lambda}^{-1/2}\| \leq \left\|K_{\lambda}^{-1/2}(\Psi - \Psi_{\ell})K_{\lambda}^{-1/2}\right\| + \left\|K_{\lambda}^{-1/2}\Psi_{\ell}K_{\lambda}^{-1/2}\right\| \leq \lambda_0^{-1} \|\Psi - \Psi_{\ell}\| + \left\|K_{\lambda}^{-1/2}K_{\ell,\lambda}^{-1/2}K_{\lambda}^{-1/2}\Psi_{\ell}K_{\ell,\lambda}K_{\ell,\lambda}^{-1/2}\right\| \leq \frac{\sqrt{2} \|\tau\|^2_4}{4 \lambda_0 \|\tau\|^2_4} + \left\|K_{\ell,\lambda}^{-1/2}\Psi K_{\ell,\lambda}^{-1/2}\right\| \leq \frac{\sqrt{2} \|\tau\|^2_4}{4 \lambda_0 \|\tau\|^2_4} + \frac{3}{2} C.$$

Therefore, there is a constant depending only on $\sigma$, $\tau$ as the upper bound for $\|K_{\lambda}^{-1/2}\Psi K_{\lambda}^{-1/2}\|$. This completes the proof of this lemma. \hfill $\square$

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

$$\|K_{\lambda}^{-\frac{1}{2}}u\| \leq \|\mathbb{E}_{\beta}[\tau(\beta^T X)\tau(\beta^T X)]K_{\lambda}^{-\frac{1}{2}}\| \leq C(1 + \lambda^{1/2}),$$

Proof. Denote $K_{\tau,m} := \mathbb{E}_{\beta}[\tau(\beta^T X)\tau(\beta^T X)]^T$ and $\Psi_{\lambda} := \Psi + \lambda \text{Id}$. Analogously to (C.25) and (C.26), we can consider

$$\Phi_{\lambda} = \mathbb{E}_{\beta}[\tau(\beta^T X)^T \tau(\beta^T X)] + \lambda \text{Id} = \left( \Psi_{\lambda} \otimes \mathbb{E}_{\beta}[\tau(\beta^T x)^2] + \lambda \right).$$
Lemma C.9. Suppose that the assumptions of Theorem 2.11 hold. For any \( n \), which only depends on \( \sigma, \tau \). Additionally, following the same proof of Lemma C.7, we can also obtain

\[
\| E_\beta \left[ \tau (\beta^\top x)^2 \right] \|_2^2 = K_{\tau,m} K_{\lambda}^{-1} K_{\tau,m}^\top \leq (\lambda + \| \tau \|_2^2) \left\| \Psi_{\lambda}^2 K_{\lambda}^{-1} \Psi_{\lambda}^2 \right\|.
\]

Additionally, following the same proof of Lemma C.7, we can also obtain \( \left\| \Psi_{\lambda}^2 K_{\lambda}^{-1} \Psi_{\lambda}^2 \right\| \leq C \), for some constant \( C > 0 \) which only depends on \( \sigma, \tau \). This concludes the proof. \( \square \)

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

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

\[
\begin{align*}
\delta_1 &= E_{\beta, \epsilon} \left[ y^\top K_{\lambda}^{-1} \left( K_N^{(2)} - K_m^{(2)} \right) K_{\lambda}^{-1} y \right], \\
\delta_2 &= E_{\beta, \epsilon} \left[ f^*(x) (K_N(x, X) - K(x, X)) K_{\lambda}^{-1} f^*(X) \right].
\end{align*}
\]

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 \( \sigma, \tau \) such that

\[
\begin{align*}
|\delta_1| &\leq C (1 + \lambda) \log(N) \sqrt{\frac{n}{N}}, \\
|\delta_2| &\leq C (1 + \lambda) \log(N) \sqrt{\frac{n}{N}}.
\end{align*}
\]

**Proof.** Let \( v := K_{\lambda}^{-1} y \) and \( \tilde{g}(x) := K_m^\top v \). Notice that \( \delta_1 = \delta_{1,1} + \delta_{1,2} \), where

\[
\begin{align*}
\delta_{1,1} &= E_{\beta, \epsilon} [v^\top E_{x} [(K_{m,N} - K_m)(K_{m,N} - K_m)^\top] v], \\
\delta_{1,2} &= 2E_{\beta, \epsilon} [v^\top E_{x} [(K_{m,N} - K_m)K_m^\top] v] = 2E_{\beta, \epsilon} [v^\top (K_{m,N} - K_m)\tilde{g}(x)],
\end{align*}
\]

Taking expectation with respect to \( W \), we can obtain

\[
\begin{align*}
0 &\leq E[\delta_{1,1}] = \frac{1}{N^2} \sum_{i,j=1}^{N} E_{\beta, \epsilon} \left[ v^\top E_{W, x} \left[ (\sigma(w_i^\top x)\sigma(w_j^\top X) - K_{m}) \left( \sigma(w_j^\top x)\sigma(w_j^\top X) - K_{m}^\top \right) \right] v \right] \\
&= \frac{1}{N^2} \sum_{i=1}^{N} E_{\beta, \epsilon} \left[ v^\top E_{W, x} \left[ (\sigma(w_i^\top x)^2 \sigma(w_i^\top X)^\top - K_{m}) \sigma(w_i^\top x)\sigma(w_i^\top X) - K_{m}^\top \right] v \right] \\
&= \frac{1}{N^2} \sum_{i=1}^{N} E_{\beta, \epsilon} \left[ v^\top E_{W, x} \left[ \sigma(w_i^\top x)^2 \sigma(w_i^\top X)^\top \sigma(w_i^\top X) - K_{m} K_{m}^\top \right] v \right] \\
&\leq \frac{1}{N} E_{\beta, \epsilon} \left[ v^\top E_{W, x} \left[ \sigma(w_i^\top x)^2 \sigma(w_i^\top X)^\top \sigma(w_i^\top X) \right] v \right],
\end{align*}
\]

where in the last line we apply the fact \( K_m = E_{w_i} [\sigma(w_i^\top x)\sigma(w_i^\top X)^\top] \in \mathbb{R}^n \) for any \( i \)-th row of \( W \).
Furthermore, by applying Lemma C.7 and Cauchy–Schwartz inequality, we have

\[
\mathbb{E}_{\beta,z} \left[ \mathbb{E}_{w,z} \left[ \sigma(w^\top x)^2 \sigma(w^\top X)^\top \sigma(w^\top X) w \right] \right]
\]

\[
= \text{Tr} \left( (K_\lambda^{-1} \mathbb{E}_{w,z} [\sigma(w^\top x)^2 \sigma(w^\top X)^\top \sigma(w^\top X)]) K_\lambda^{-1} (\Psi + \sigma_z^2 \text{Id}) \right)
\]

\[
= \mathbb{E}_{w,z} \left[ \sigma(w^\top x)^2 \sigma(w^\top X) K_\lambda^{-1} (\Psi + \sigma_z^2 \text{Id}) K_\lambda^{-1} \sigma(w^\top X)^\top \right]
\]

\[
\leq \|\sigma\|_4^2 \mathbb{E}_{w,z} \left[ \|\sigma(w^\top X)\|^4 \right]^{1/2} \left\| K_\lambda^{-1} (\Psi + \sigma_z^2 \text{Id}) K_\lambda^{-1} \right\|^2
\]

\[
\leq \|\sigma\|_4^2 \cdot \frac{1}{\lambda_0} \left( C + \frac{\sigma_z^2}{\lambda_0} \right) \mathbb{E}_{w} \left[ \|\sigma(w^\top X)\|^4 \right]^{1/2}
\]

\[
= \|\sigma\|_4^2 \cdot \frac{1}{\lambda_0} \left( C + \frac{\sigma_z^2}{\lambda_0} \right) \mathbb{E}_{w} \left[ \left( \sum_{i=1}^n \sigma(w^\top x_i)^2 \right)^2 \right]^{1/2}
\]

\[
\leq \|\sigma\|_4^2 \cdot \frac{1}{\lambda_0} \left( C + \frac{\sigma_z^2}{\lambda_0} \right) \cdot n
\]

(C.33)

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

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

for some constant \( C_{1,1} > 0 \) which only relies on \( \sigma \) and \( \sigma_z \). Then, Markov inequality deduces that for sufficiently large \( n \),

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

Next, we consider \( \delta_2 \). Let \( z_1, z_2 \) be two i.i.d. copies of \( x \), and \( \beta_1, \beta_2 \) be two i.i.d. copies of \( \beta \). Let \( u_i := K_\lambda^{-1} \tau(\beta_i^\top X) \) and \( g_i(x) := \tau(\beta_i^\top x) \) for \( i = 1, 2 \). Notice that \( \mathbb{E}_{w,z} [\sigma(w^\top z_1) \sigma(X^\top w_i)] = K(X, z_1) \). Then, taking expectation with respect to \( W \), we can obtain

\[
\mathbb{E}[\delta_2^2] = \mathbb{E}_{\beta_1, \beta_2} \left[ u_1^\top \mathbb{E}_{W,z_1,z_2} [(K_N(X, z_1) - K(X, z_1)) g_1(z_1) g_2(z_2) (K_N(z_2, X) - K(z_2, X))] u_2 \right]
\]

\[
= \frac{1}{N^2} \sum_{i,j=1}^N \mathbb{E} \left[ u_1^\top \left( \sigma(w_i^\top z_1) \sigma(X^\top w_i) - K(X, z_1) \right) g_1(z_1) g_2(z_2) (\sigma(w_j^\top z_2) \sigma(w_j^\top X) - K(z_2, X)) u_2 \right]
\]

\[
= \frac{1}{N^2} \sum_{i=1}^N \mathbb{E} \left[ g_1(z_1) g_2(z_2) \sigma(w_i^\top z_1) \sigma(w_i^\top z_2) u_1^\top \left( \sigma(X^\top w_i) \sigma(w_i^\top X) - K(X, z_1) K(z_2, X) \right) u_2 \right]
\]

\[
\leq \frac{1}{N} \mathbb{E}_{\beta_1, \beta_2, w, z_1, z_2} \left[ g_1(z_1) g_2(z_2) \sigma(w_i^\top z_1) \sigma(w_i^\top z_2) \cdot u_1^\top \sigma(X^\top w) \sigma(w^\top X) u_2 \right],
\]

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

\[
\mathbb{E}_{\beta, x} \left[ \sigma(w_i^\top x) \sigma(w_i^\top z_1) \right] \geq 0.
\]

Let \( v_i := K_\lambda^{-1/2} \mathbb{E}_{\beta, x} \left[ \tau(\beta_i^\top z_i) \tau(\beta_i^\top X) \right] \) for \( i = 1, 2 \). Then, Lemma C.8 shows that \( \|v_i\| \leq C(1+\lambda) \) for some universal
constant $C$. Thus, similarly with the derivation of (C.33), we can deduce that
\[
\begin{align*}
\mathbb{E}_{\beta,\delta_2,w,z_1,z_2} \left[ g_1(z_1)g_2(z_2)\sigma(w^\top z_1)\sigma(w^\top z_2) \cdot u_1^\top \sigma(X^\top w)\sigma(w^\top X)u_2 \right] \\
\leq \mathbb{E}_{w,z_1,z_2} \left[ \sigma(w^\top z_1)\sigma(w^\top z_2) \cdot v_1 K_{\lambda}^{-1/2}\sigma(X^\top w)\sigma(w^\top X)K_{\lambda}^{-1/2}v_2 \right] \\
\leq C(1 + \lambda)^2 \|\sigma\|^2 \mathbb{E}_w \left[ \left( \sigma(w^\top X)K_{\lambda}^{-1}\sigma(X^\top w) \right)^2 \right]^{\frac{1}{2}} \\
\leq C \frac{(1 + \lambda)^2 \|\sigma\|^2}{\lambda_0} n,
\end{align*}
\]
where the last line is analogous to (C.33). Therefore, $\mathbb{E}[\delta_2^2] \leq C_2(1 + \lambda)^2 \frac{n}{\lambda_0}$. This indicates, for any $t > 0$,
\[
\mathbb{P}(|\delta_2| > 2t(1 + \lambda)) \leq \frac{C_2 n}{Nt^2}.
\]
Hence, by taking $t = \log(N)\sqrt{C_2 n/N}$, we can conclude the bound of $\delta_2$ in (C.31) with probability at least $1 - \frac{1}{4} \log^{-2}(N)$.

The analysis of $\delta_{1,2}$ is similar to the analysis for $\delta_2$. By definition, we have
\[
\delta_{1,2} = 2 \text{Tr} \left[ \mathbb{E}_w \left( (K_{m,N} - K_m)K_m^{\top}K_m^{-1}(\Psi + \sigma^2 I)K_m^{\top} \right) \right] = 2 \mathbb{E}_w \left[ K_m^{\top}K_m^{-1}(\Psi + \sigma^2 I)K_m^{-1}(K_{m,N} - K_m) \right].
\]
Then, consider $z_1, z_2$ as i.i.d. copies of $x$. Let $A := K_{\lambda}^{-1}(\Psi + \sigma^2 I)K_{\lambda}^{-1}$ and $K_{m,i} := \mathbb{E}_w[\sigma(w^\top z_i)\sigma(w^\top X)]^\top \in \mathbb{R}^n$, for $i = 1, 2$. Then, we have
\[
\begin{align*}
\mathbb{E}[\delta_{1,2}^2] &= \frac{4}{N^2} \sum_{i,j=1}^{N} \mathbb{E} \left[ K_{m,1} A \left( \sigma(w_i^\top z_1)\sigma(X^\top w_i) - K_{m,1} \right) \left( \sigma(w_j^\top z_2)\sigma(w_j^\top X) - K_{m,2} \right) A K_{m,2} \right] \\
&= \frac{4}{N^2} \sum_{i=1}^{N} \mathbb{E} \left[ K_{m,1} A \left( \sigma(w_i^\top z_1)\sigma(X^\top w_i) - K_{m,1} \right) \left( \sigma(w_i^\top z_2)\sigma(w_i^\top X) - K_{m,2} \right) A K_{m,2} \right] \\
&= \frac{4}{N^2} \sum_{i=1}^{N} \mathbb{E} \left[ K_{m,1} A \left( \sigma(w_i^\top z_1)\sigma(w_i^\top z_2)\sigma(X^\top w_i)\sigma(w_i^\top X) - K_{m,1}K_{m,2} \right) A K_{m,2} \right] \\
&\leq \frac{4}{N} \mathbb{E}_{w,z_1,z_2} \left[ \sigma(w^\top z_1)\sigma(w^\top z_2)K_{m,1}A\sigma(w^\top X)\sigma(w^\top X)AK_{m,2} \right] \\
&\leq \frac{4}{N} \mathbb{E} \left[ \sigma(w^\top z_1)^2\sigma(w^\top z_2)^2 \right] \frac{1}{2} \mathbb{E} \left[ K_{m,1}K_{m,2}^{\frac{1}{2}} \right] \left( K_{m,2}K_{m,2}^{\frac{1}{2}} \right) \left( A \right) \left( A \right)^{\frac{1}{2}} \left( K_{m,1}K_{m,1}^{\frac{1}{2}} \right) \left( A \right) \left( A \right)^{\frac{1}{2}} \\
&\leq \frac{4C^2(1 + \lambda)^2 \|\sigma\|^2}{\lambda_0 n} \mathbb{E} \left[ \sum_{i=1}^{n} (\sigma(w^\top x_i)^2)^{1/2} \right] \leq C_{1,2}(1 + \lambda)^2 n \frac{n}{N},
\end{align*}
\]
for some constant $C_{1,2} > 0$, where (i) is because of positiveness of $A$ and (ii) is due to Lemmas C.7 and C.8. Thus, Markov inequality allows us to obtain for a constant $C > 0$,
\[
\mathbb{P} \left( |\delta_{1,2}| > C(1 + \lambda)\log(N) \sqrt{\frac{n}{N}} \right) \leq \frac{1}{4} \log^{-2}(N),
\]
Together with (C.34), we can conclude the bound for $\delta_1$ 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 $K = K(X, X)$ and $K_N = K_N(X, X)$. Hence, we can further decompose the test errors (2.20) for both RFRR and KRR in the following way:

$$
\mathcal{L}(\hat{f}_\lambda^{(K)}) = \mathbb{E}[|f^*(x)|^2] + \text{Tr} \left[ (K + \lambda \text{Id})^{-1} \mathbb{E}[y y^\top] (K + \lambda \text{Id})^{-1} \mathbb{E}[K(X, x)K(x, X)] \right] - 2 \text{Tr} \left[ (K + \lambda \text{Id})^{-1} \mathbb{E}[y f^*(x)K(x, X)] \right],
$$

$$
\mathcal{L}(\hat{f}_\lambda^{(RF)}) = \mathbb{E}[|f^*(x)|^2] + \text{Tr} \left[ (K_N + \lambda \text{Id})^{-1} \mathbb{E}[y y^\top] (K_N + \lambda \text{Id})^{-1} \mathbb{E}[K_N(X, x)K_N(x, X)] \right] - 2 \text{Tr} \left[ (K_N + \lambda \text{Id})^{-1} \mathbb{E}[y f^*(x)K_N(x, X)] \right],
$$

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

$$
E_1 := \text{Tr} \left[ (K_N + \lambda \text{Id})^{-1} \mathbb{E}[y y^\top] (K_N + \lambda \text{Id})^{-1} \mathbb{E}[K_N(X, x)K_N(x, X)] \right],
$$

$$
E_1 := \text{Tr} \left[ (K + \lambda \text{Id})^{-1} \mathbb{E}[y y^\top] (K + \lambda \text{Id})^{-1} \mathbb{E}[K(X, x)K(x, X)] \right],
$$

$$
E_2 := \text{Tr} \left[ (K_N + \lambda \text{Id})^{-1} \mathbb{E}[y f^*(x)K_N(x, X)] \right],
$$

$$
E_2 := \text{Tr} \left[ (K + \lambda \text{Id})^{-1} \mathbb{E}[y f^*(x)K(x, X)] \right].
$$

Therefore, by taking the expectation with respect to $\beta$ and $\varepsilon$, we have

$$
E_1 = \mathbb{E}_x \left[ K_N(x, X)(K_N + \lambda \text{Id})^{-1} (\Psi + \sigma_\varepsilon^2 \text{Id}) (K_N + \lambda \text{Id})^{-1} K_N(X, x) \right],
$$

$$
E_1 = \mathbb{E}_x \left[ K(x, X)(K + \lambda \text{Id})^{-1} (\Psi + \sigma_\varepsilon^2 \text{Id}) (K + \lambda \text{Id})^{-1} K(X, x) \right],
$$

$$
E_2 = \text{Tr} \left[ (K_N + \lambda \text{Id})^{-1} \mathbb{E}[u K_N(x, X)] \right],
$$

$$
E_2 = \text{Tr} \left[ (K + \lambda \text{Id})^{-1} \mathbb{E}[u K(x, X)] \right],
$$

where $\Psi = \mathbb{E}_\beta [\tau(X^\top \beta)\tau(\beta^\top X)]$ and $u = \mathbb{E}_\beta [\tau(X^\top \beta)f^*(x)] \in \mathbb{R}^n$. 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

$$
J_{1,1} := \mathbb{E}_x \left[ K_{m,N}^\top (K_{N,\lambda}^{-1} - K_\lambda^{-1}) (\Psi + \sigma_\varepsilon^2 \text{Id}) K_{N,\lambda}^{-1} K_{m,N} \right],
$$

$$
J_{1,2} := \mathbb{E}_x \left[ K_{m,N}^\top K_\lambda^{-1} (\Psi + \sigma_\varepsilon^2 \text{Id}) (K_{N,\lambda}^{-1} - K_\lambda^{-1}) K_{m,N} \right],
$$

$$
J_{1,3} := \mathbb{E}_{\beta, \varepsilon} \left[ y^\top K_\lambda^{-1} \left( K_N^{(2)} - K^{(2)} \right) K_\lambda^{-1} y \right],
$$

$$
J_{2,1} := \mathbb{E}_x \left[ K_N(x, X) (K_{N,\lambda}^{-1} - K_\lambda^{-1}) u \right],
$$

$$
J_{2,2} := \mathbb{E}_x \left[ (K_N(x, X) - K(x, X)) K_\lambda^{-1} u \right].
$$

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

$$
K_{N,\lambda}^{-1} - K_\lambda^{-1} = K_{N,\lambda}^{-1} (K - K_N) K_\lambda^{-1}
$$

$$
= K_{N,\lambda}^{-\frac{1}{2}} K_{N,\lambda}^{-\frac{1}{2}} K_\lambda^{-\frac{1}{2}} K_\lambda^{-\frac{1}{2}} (K - K_N) K_\lambda^{-\frac{1}{2}} K_\lambda^{-\frac{1}{2}}.
$$
Hence, we can apply Proposition C.1, Corollary C.2, Lemmas C.6, C.7 and C.8 to conclude that

\[ |J_{1,1}| \leq \mathbb{E}_x \left[ \left( K_{m,n} K_{N}^{-1/2} \right)^2 \cdot \left( K_{N}^{-1/2} K_{m}^{-1/2} \right)^2 \cdot K_{N}^{-1/2} (K - K_N) K_{N}^{-1/2} \right] \cdot \left( K_{N}^{-1/2} (\Psi + \sigma^2 I_d) K_{N}^{-1/2} \right) \]

\[ \leq C(1 + \lambda) \log(N) \sqrt{\frac{n}{N}}, \]

\[ |J_{1,2}| \leq \mathbb{E}_x \left[ \left( K_{m,n} K_{N}^{-1/2} \right)^2 \cdot \left( K_{N}^{-1/2} K_{m}^{-1/2} \right)^2 \cdot \left( K_{N}^{-1/2} K_{N}^{-1/2} (K - K_N) K_{N}^{-1/2} \right) \right] \cdot \left( K_{N}^{-1/2} (\Psi + \sigma^2 I_d) K_{N}^{-1/2} \right) \]

\[ \leq C(1 + \lambda) \log(N) \sqrt{\frac{n}{N}}, \]

\[ |J_{2,1}| \leq \mathbb{E}_x \left[ \left( K_{m,n} K_{N}^{-1/2} K_{N}^{-1/2} (K - K_N) K_{N}^{-1/2} \right) \right] \cdot \left( K_{N}^{-1/2} (\Psi + \sigma^2 I_d) K_{N}^{-1/2} \right) \]

\[ \leq C(1 + \lambda) \log(N) \sqrt{\frac{n}{N}}, \]

for some constant \( C > 0 \) depending on the norms of \( \tau \) and \( \sigma, \lambda_0 \) and \( \sigma_e \) 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 \( \lambda_{\min}(K_\ell) \geq 2\lambda_0 \) and \( \lambda_{\min}(K) \geq \lambda_0 \). 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

\[ |E_{\text{train}}^{(\ell,\lambda)} - E_{\text{train}}^{(K,\lambda)}| = \frac{\lambda^2}{n} \left| y^\top \left[ (K_\ell + \lambda I) - (K + \lambda I) \right] y \right| \]

\[ \leq \frac{\lambda^2}{n} \left\| (K_\ell + \lambda I) - (K + \lambda I) \right\| \left\| y \right\|^2 \]

\[ \leq \frac{3\lambda^2 \left\| y \right\|^2}{2\lambda_0 n} \left\| (K_\ell + \lambda I)^{-1} - (K + \lambda I)^{-1} \right\| \]

\[ \leq \frac{3\lambda^2 \left\| y \right\|^2}{2\lambda_0 n} \left\| K_\ell - K \right\| \leq \frac{C\lambda^2 \left\| y \right\|^2}{\lambda_0 n} \frac{1}{\lambda_0 - 1} \left( X^\top X \right)^{\circ (\ell+1)} - \text{Id} \]

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

\[ \left\| (K_\ell + \lambda I)^{-1} - (K + \lambda I)^{-1} \right\| = \left\| K_\ell^{-1} (K - K_\ell) K_\ell^{-1} \right\| \leq \frac{1}{(\lambda + \lambda_0)\lambda_0} \left\| K_\ell - K \right\|. \]

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

\[ \left( \lambda + \left\| \sigma \right\|^2 \right)^{-1} \leq \text{tr} K_{\ell,\lambda}^{-1} \leq \lambda_0^{-1}. \]

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

\[ |GCV_n^{(K,\lambda)} - GCV_n^{(\ell,\lambda)}| \leq \lambda^{-2} \left( \left\| (K_\ell^{-1}) - (K^{-1}) \right\|^2 \right) E_{\text{train}}^{(K,\lambda)} \]

\[ + \lambda^{-2} \left( \left\| (K_\ell^{-1}) - (K^{-1}) \right\|^2 \right) E_{\text{train}}^{(\ell,\lambda)} \]

\[ \leq \lambda^{-2} \left( \left\| (K_\ell^{-1}) - (K^{-1}) \right\|^2 \right) E_{\text{train}}^{(K,\lambda)} + \lambda^{-2} \left( \left\| (K_\ell^{-1}) - (K^{-1}) \right\|^2 \right) E_{\text{train}}^{(\ell,\lambda)} \].
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:

\[
\lambda^{-2} \left| \left( \text{tr} \left[ K^{-1}_\lambda \right] \right)^{-2} \left( \text{tr} \left[ K^{-1}_{\ell,\lambda} \right] \right) \right| E_{\text{train}}^{(K,\lambda)} \leq \left( \text{tr} \left[ K^{-1}_\lambda \right] \right)^{-2} \left( \text{tr} \left( K^{-1}_\lambda - K^{-1}_{\ell,\lambda} \right) \right) \left| \text{tr} \left( K^{-1}_{\ell,\lambda} + K^{-1}_{\ell,\lambda} \right) \right| \frac{1}{n} \| K^{-2}_{\ell,\lambda} \| \| y \|^2 \\
\leq \frac{8(\lambda + \|\sigma\|_2^2)^4}{\lambda_0^n} \| K - K_{\ell} \| \leq \frac{8\sqrt{2}(\lambda + \|\sigma\|_2^2)\|\sigma\|^2_2}{\lambda_0^n} \left\| \left( X^T X \right)^{\otimes \ell + 1} - \text{Id} \right\|_F.
\]

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

\[
\lambda^{-2} \left( \left( \text{tr} \left[ K^{-1}_\lambda \right] \right)^{-2} E_{\text{train}}^{(K,\lambda)} - E_{\text{train}}^{(\ell,\lambda)} \right) \leq \frac{C(\lambda + \|\sigma\|^2_2)\|\sigma\|^2_4}{\lambda_0^n} \left\| \left( X^T X \right)^{\otimes \ell + 1} - \text{Id} \right\|_F,
\]

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

\[
\text{CV}_{n}^{(\ell,\lambda)} - \text{CV}_{n}^{(K,\lambda)} \leq C_2(1 + \lambda^4) \frac{\| y \|^2}{n} \left\| \left( X^T X \right)^{\otimes \ell + 1} - \text{Id} \right\|_F,
\]

for some constant $C_2 > 0$ which only relies on $\|\sigma\|_2$, $\|\sigma\|_4$, and $\lambda_0$. 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 $\beta$ and $\varepsilon$, we have $\left| \mathcal{L}(\hat{f}_\lambda^{(\ell)}(x)) - \mathcal{L}(\hat{f}_\lambda^{(K)}(x)) \right| \leq |E_1' - \hat{E}_1| + |E_2' - \hat{E}_2|$, where

\[
E_1' := \mathbb{E}_{x} \left[ K_\ell(x, x) K^{-1}_{\ell,\lambda} (\Psi + \sigma^2 \varepsilon \text{Id}) K^{-1}_{\ell,\lambda} K_\ell(x, x) \right], \\
\hat{E}_1 := \mathbb{E}_{x} \left[ (\Psi + \sigma^2 \varepsilon \text{Id}) K^{-1}_\lambda K(x, x) \right], \\
E_2' := \text{Tr} \left[ K^{-1}_{\ell,\lambda} \mathbb{E}_{u} [u K_\ell(x, x)] \right], \\
\hat{E}_2 := \text{Tr} \left[ K^{-1}_\lambda \mathbb{E}_{u} [u K(x, x)] \right].
\]

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

\[
\| K_{m,\ell} - K_{m} \| \leq \| K_{m,\ell} - K_{m} \|_2 \leq \sqrt{2}\|\sigma\|^2_4 \left\| \left( X^T x \right)^{\otimes (\ell + 1)} \right\|_2 \leq \frac{1}{\sqrt{2}} \lambda_0.
\]

Moreover, analogously to Lemma C.6, we have

\[
\left\| K_{\ell,\lambda}^{\top} K^{-1/2}_{\ell,\lambda} \right\| \leq \|\sigma\|^2_2 + \lambda.
\]

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

\[
\left\| K^{-1/2}_{\ell,\lambda} \Psi K^{-1/2}_{\ell,\lambda} \right\|, \left\| K^{-1/2}_{\ell,\lambda} u \right\| \leq C,
\]

for some constant $C > 0$ depending only on $\sigma, \tau$. Therefore, because of Proposition 2.4, Lemmas C.6 and C.7, and (C.39),
(C.40) and (C.41), we can deduce that
\[
|E_1' - 
\begin{align*}
E_2' - \bar{E}_2 & \leq \frac{1}{N} \sum_{i=1}^{N} \left( \mathbb{E} \left[ \left( \sum_{k \neq i} \left( f(x_i) - f(x_j) \right) \right)^2 \right] \right) \\
& \leq C_1'(1 + \lambda) \left\| \left( \tilde{X}^\top \tilde{X} \right)^{\odot \ell+1} \right\|_F, \end{align*}
\]
for some constant \( C_1' > 0 \). Similarly, due to Lemma C.8, (C.39), (C.40) and (C.41), we can obtain
\[
|E_1' - \bar{E}_1| \leq \frac{1}{N} \sum_{i=1}^{N} \left( \mathbb{E} \left[ \left( \sum_{k \neq i} \left( f(x_i) - f(x_j) \right) \right)^2 \right] \right) \\
\leq C_2'(1 + \lambda) \left\| \left( \tilde{X}^\top \tilde{X} \right)^{\odot \ell+1} \right\|_F, \end{align*}
\]
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 \( N/ \log^2(N) \geq C_1(1 + \lambda^2)n \) and \( n \geq \max\{n_0, n_1\} \), with probability at least \( 1 - \log^{-1}(N) \),
\[
\mathcal{L}(\tilde{f}_\lambda^{(RF)}) \geq \| P_{> \ell} f^* \|_2^2 - C_1(1 + \lambda) \log(N) \sqrt{\frac{n}{N}} - C_2 \sqrt{n} \left\| \left( X^\top x \right)^{\odot \ell+1} \right\|_2^2,
\]
and
\[
\mathcal{L}(\tilde{f}_\lambda^{(RF)}) \geq \| P_{> \ell} f^* \|_2^2 + \sigma_x^2 \mathbb{E} x \left[ K_{m, \ell}^T K_{\ell, \ell} K_{m, \ell} - C_2 \sqrt{n} \left\| \left( X^\top x \right)^{\odot \ell+1} \right\|_2 \right. \\
- C_2(1 + \lambda) \left( \left\| \left( X^\top x \right)^{\odot \ell+1} \right\|_F + \log(N) \sqrt{\frac{n}{N}} \right),
\]
where \( C_1 \) depends only on \( \sigma \), and \( C_2 > 0 \) depends only on \( \sigma, \tau \) and \( \sigma_x \). In particular, when \( N/ \log^2 N \gg n \), with high probability,
\[
\mathcal{L}(\tilde{f}_\lambda^{(RF)}) \geq \| P_{> \ell} f^* \|_2^2 + \sigma_x^2 \mathbb{E} x \left[ K_{m, \ell}^T K_{\ell, \ell} K_{m, \ell} \right] - o_n(1)
\]
\[
\geq \| P_{> \ell} f^* \|_2^2 - o_n(1).
\]

**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
\[
f^* (x) = \tau(\beta^\top x) = \sum_{k=0}^{\infty} \zeta_k(\tau) h_k(\beta^\top x), \]
\[
(P_{\leq \ell} f^*)(x) = \sum_{k < \ell} \zeta_k(\tau) h_k(\beta^\top x), \quad (P_{> \ell} f^*)(x) = \sum_{k \geq \ell+1} \zeta_k(\tau) h_k(\beta^\top x). \]
Similarly, we define
\[
f^* (X) = \tau(\beta^\top X) = \sum_{k=0}^{\infty} \zeta_k(\tau) h_k(\beta^\top X) \in \mathbb{R}^n, \]
\[
(P_{\leq \ell} f^*)(X) = \sum_{k \leq \ell} \zeta_k(\tau) h_k(\beta^\top X), \quad (P_{> \ell} f^*)(X) = \sum_{k \geq \ell+1} \zeta_k(\tau) h_k(\beta^\top X),
\]
(C.43)
By the property of Hermite polynomials in (2.3), we know
\[ \mathbb{E}_\beta[h_j(\beta^\top x)h_k(\beta^\top x_i)] = \delta_{jk}(x, x_i)^k. \]
This implies
\[ \|f^*\|_2^2 = \mathbb{E}_\beta[f^*(x)^2] = \sum_{k=0}^\infty \zeta_k(\tau)^2 = \|\tau\|_2^2, \]
and from Assumption 2.10,
\[ \|P_{\leq \ell} f^*\|_2^2 = \sum_{k=0}^\ell \zeta_k^2(\tau), \quad \|P_{> \ell} f^*\|_2^2 = \sum_{k=\ell+1}^\infty \zeta_k^2(\tau), \]
\[ \mathbb{E}[P_{\leq \ell} f^*(x)P_{> \ell} f^*(x)] = 0. \quad \text{(C.44)} \]
From (2.9), the predictor of the KRR is given by
\[ \hat{f}_{\lambda}^{(K)}(x) := K(x, X)(K(X, X) + \lambda \text{Id})^{-1} (f^*(X) + \epsilon), \]
where
\[ K(x, X) = \sum_{k=0}^\infty \zeta_k^2(\sigma)(x^\top X)^{\otimes k} \in \mathbb{R}^{1 \times n}, \]
and from Assumption 2.10,
\[ \|K(x, X)\| \leq \sqrt{2n} \|\sigma\|_4^2. \quad \text{(C.45)} \]
Define
\[ P_{\leq \ell} \hat{f}_{\lambda}^{(K)}(x) := K(x, X)(K(X, X) + \lambda \text{Id})^{-1} (P_{\leq \ell} f^*(X)), \]
\[ P_{> \ell} \hat{f}_{\lambda}^{(K)}(x) := K(x, X)(K(X, X) + \lambda \text{Id})^{-1} (P_{> \ell} f^*(X) + \epsilon). \]
From the orthogonal relation in (2.3) and (C.43),
\[ \mathbb{E}_\beta,\epsilon[P_{\leq \ell} \hat{f}_{\lambda}^{(K)}(x)P_{> \ell} \hat{f}_{\lambda}^{(K)}(x)] = 0, \quad \text{(C.46)} \]
\[ \mathbb{E}_\beta[(P_{> \ell} f^*)(X)(P_{> \ell} f^*)(x)] = \sum_{k=\ell+1}^\infty \zeta_k^2(\tau)((x^\top x_1)^k, \ldots, (x^\top x_n)^k). \]
Then by the linearity of expectation, we have
\[ \mathbb{E}_\beta,\epsilon[\hat{f}_{\lambda}^{(K)}(x)P_{> \ell} f^*(x)] = \sum_{k>\ell} \zeta_k^2(\tau)K(x, X)(K(X, X) + \lambda \text{Id})^{-1}((x^\top x_1)^k, \ldots, (x^\top x_n)^k)^\top, \]
which implies
\[ \mathbb{E}_\beta,\epsilon[P_{> \ell} \hat{f}_{\lambda}^{(K)}(x)P_{> \ell} f^*(x)] \leq \|K(x, X)\| \lambda_0^{-1} \sum_{k=\ell+1}^\infty \zeta_k^2(\tau) \left\| (X^\top x)^{\otimes k} \right\|_2 \]
\[ \leq \sqrt{2n} \|\sigma\|_4^2 \lambda_0^{-1} \|\tau\|_4^4 \left( \sum_{k=\ell+1}^\infty \left\| (X^\top x)^{\otimes k} \right\|_2^2 \right)^{1/2} \]
\[ \leq 2\sqrt{2n} \|\sigma\|_4^2 \lambda_0^{-1} \|\tau\|_4^4 \left\| (X^\top x)^{\otimes (\ell+1)} \right\|_2, \quad \text{(C.47)} \]
where the second inequality is due to (C.45), and the third inequality comes from Cauchy’s inequality. Recall the general-
This completes the proof of the lower bound of (C.42).