LEARNING CURVES FOR GAUSSIAN PROCESS REGRESSION WITH POWER-LAW PRIORS AND TARGETS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study the power-law asymptotics of learning curves for Gaussian process regression (GPR). When the eigenspectrum of the prior decays with rate α and the eigenexpansion coefficients of the target function decay with rate β, we show that the generalization error behaves as $\tilde{O}(n^{\max\{\frac{1}{\alpha}-1, \frac{1-2\beta}{\alpha}\}})$ with high probability over the draw of n input samples. Under similar assumptions, we show that the generalization error of kernel ridge regression (KRR) has the same asymptotics. Infinitely wide neural networks can be related to KRR with respect to the neural tangent kernel (NTK), which in several cases is known to have a power-law spectrum. Hence our methods can be applied to study the generalization error of infinitely wide neural networks. We present toy experiments demonstrating the theory.

1 INTRODUCTION

Gaussian processes (GPs) provide a flexible and interpretable framework for learning and adaptive inference, and are widely used for constructing prior distributions in non-parametric Bayesian learning. From an application perspective, one crucial question is how fast do GPs learn, i.e., how much training data is needed to achieve a certain level of generalization performance. Theoretically, this is addressed by analyzing so-called "learning curves", which describe the generalization error as a function of the training set size n. The rate at which the curve approaches zero determines the difficulty of learning tasks and conveys important information about the asymptotic performance of GP learning algorithms. In this paper, we study the learning curves for Gaussian process regression. Our main result characterizes the asymptotics of the generalization error in cases where the eigenvalues of the GP kernel and the coefficients of the eigenexpansion of the target function have a power-law decay. In the remainder of this introductory section, we review related work and outline our main contributions.

Gaussian processes A GP model is a probabilistic model on an infinite-dimensional parameter space (Williams and Rasmussen, 2006; Orbanz and Teh, 2010). In GP regression (GPR), for example, this space can be the set of all continuous functions. Assumptions about the learning problem are encoded by way of a prior distribution over functions, which gets transformed into a posterior distribution given some observed data. The mean of the posterior is then used for prediction. The model uses only a finite subset of the available parameters to explain the data and this subset can grow arbitrarily large as more data are observed. In this sense, GPs are "non-parametric" and contrast with parametric models, where there is a fixed number of parameters. For regression with Gaussian noise, a major appeal of the GP formalism is that the posterior is analytically tractable. GPs are also one important part in learning with kernel machines (Kanagawa et al., 2018) and modeling using GPs has recently gained considerable traction in the neural network community.

Neural networks and kernel learning From a GP viewpoint, there exists a well known correspondence between kernel methods and infinite neural networks (NNs) first studied by Neal (1996). Neal showed that the outputs of a randomly initialized one-hidden layer neural network (with appropriate scaling of the variance of the initialization distribution) converges to a GP over functions in the limit of an infinite number of hidden units. Follow-up work extended this correspondence with analytical expressions for the kernel covariance for shallow NNs by Williams (1997), and more recently for deep fully-connected NNs (Lee et al., 2018; de G. Matthews et al., 2018), convolutional NNs with many channels (Novak et al., 2019; Garriga-Alonso et al., 2019), and more general architectures (Yang, 2019). The correspondence enables *exact* Bayesian inference in the associated GP model for

infinite-width NNs on regression tasks and has led to some recent breakthroughs in our understanding of overparameterized NNs (Jacot et al., 2018; Lee et al., 2019; Arora et al., 2019; Belkin et al., 2018; Daniely et al., 2016; Yang and Salman, 2019; Bietti and Mairal, 2019). The most prominent kernels associated with infinite-width NNs are the Neural Network Gaussian Process (NNGP) kernel when only the last layer is trained (Lee et al., 2018; de G. Matthews et al., 2018), and the Neural Tangent Kernel (NTK) when the entire model is trained (Jacot et al., 2018). Empirical studies have shown that inference with such infinite network kernels is competitive with standard gradient descent-based optimization for fully-connected architectures (Lee et al., 2020).

Learning curves A large-scale empirical characterization of the generalization performance of state-of-the-art deep NNs showed that the associated learning curves often follow a power law of the form $n^{-\beta}$ with the exponent β ranging between 0.07 and 0.35 depending on the data and the algorithm (Hestness et al., 2017; Spigler et al., 2020). Power-law asymptotics of learning curves have been theoretically studied in early works for the Gibbs learning algorithm (Amari et al., 1992; Amari and Murata, 1993; Haussler et al., 1996) that showed a generalization error scaling with exponent $\beta = 0.5$, 1 or 2 under certain assumptions. More recent results from statistical learning theory characterize the shape of learning curves depending on the properties of the hypothesis class (Bousquet et al., 2021). In the context of GPs, approximations and bounds on learning curves have been investigated in several works (Sollich, 1999; Sollich and Halees, 2002; Sollich, 2001; Opper and Vivarelli, 1999; Opper and Malzahn, 2002; Williams and Vivarelli, 2000; Malzahn and Opper, 2001a;b; Seeger et al., 2008; Van Der Vaart and Van Zanten, 2011; Le Gratiet and Garnier, 2015), with recent extensions to kernel regression from a spectral bias perspective (Bordelon et al., 2020; Canatar et al., 2021). For a review on learning curves in relation to its shape, monotonicity and some open problems, see the works of Loog et al. (2019); Viering et al. (2019); Viering and Loog (2021).

Power-law decay of the kernel eigenspectrum The rate of decay of the eigenvalues of the GP kernel conveys important information about its smoothness. Intuitively, if a process is "rough" with more power at high frequencies, then the eigenspectrum decays more slowly. On the other hand, kernels that define smooth processes have a fast-decaying eigenspectrum (Stein, 2012; Williams and Rasmussen, 2006). The precise eigenvalues $(\lambda_p)_{p\geq 1}$ of the operators associated to many kernels and input distributions are not known explicitly, except for a few special cases (Williams and Rasmussen, 2006). Often, however, the asymptotic properties are known. An early work by Widom (1963) gave the asymptotic rate of decay of the eigenvalues of stationary kernels for input distributions with bounded support under certain regularity assumptions on the spectral density of the covariance function. For inputs distributed uniformly on the unit interval, Ritter et al. (1995) showed that r-times mean square differentiable processes feature an asymptotic power-law decay of the form $\lambda_p \propto p^{-(2r+2)}$. More recently, Ronen et al. (2019) showed that for inputs distributed uniformly on a hypersphere, the eigenfunctions of the arc-cosine kernel are spherical harmonics and the eigenvalues follow a power-law decay. The spectral properties of the NTK are integral to the analysis of training convergence and generalization of NNs, and several recent works empirically justify and rely on a power law assumption for the NTK spectrum (Bahri et al., 2021; Canatar et al., 2021; Lee et al., 2020; Nitanda and Suzuki, 2021). Velikanov and Yarotsky (2021) showed that the power-law asymptotics of the infinite network NTK are determined primarily by the singularities of the kernel and has the form $\lambda_p \propto p^{-\nu}$ with $\nu = 1 + \frac{1}{d}$, where d is the input dimension.

Related works The closest approach to the work in this paper is probably the one presented in the works of Sollich (1999); Sollich and Halees (2002); Sollich (2001), where average-case learning curves for GPR are derived under the assumption that the model is correctly specified, with recent extensions to kernel regression focusing on noiseless data sets (Bordelon et al., 2020; Spigler et al., 2020) and Gaussian design analysis (Cui et al., 2021). A related but complementary line of work studies the convergence rates and posterior consistency properties of Bayesian non-parametric models (Barron, 1998; Seeger et al., 2008; Van Der Vaart and Van Zanten, 2011).

Contributions We characterize the asymptotics of the generalization error of GPR under the assumption that the eigenvalues of the GP kernel and the coefficients in the expansion of the target function over corresponding eigenvectors follow a power law. Our main contributions are as follows:

- When the eigenspectrum of the prior decays with rate α and the eigenexpansion coefficients of the target function decay with rate β, we show that with high probability over the draw of n input

samples, the negative log-marginal likelihood behaves as $\tilde{O}(n^{\max\{\frac{1}{\alpha},\frac{1-2\beta}{\alpha}+1\}})$ (Theorem 7) and the generalization error behaves as $\tilde{O}(n^{\max\{\frac{1}{\alpha}-1,\frac{1-2\beta}{\alpha}\}})$ (Theorem 9). In the special case that the model is correctly specified, i.e., the GP prior is the true one from which the target functions are actually generated, we recover a result due to Sollich and Halees (2002) (vide Remark 10).

- Under similar assumptions as in the previous item, we leverage the equivalence between GPR and kernel ridge regression (KRR) (Kanagawa et al., 2018) to show that the generalization error of KRR behaves as $\tilde{O}(n^{\max\{\frac{1}{\alpha}-1,\frac{1-2\beta}{\alpha}\}})$ (Theorem 12). Specializing to the case of KRR with Gaussian design, we recover a result due to Cui et al. (2021) (vide Remark 14).
- We present a few toy experiments demonstrating the theory for GPR with the arc-cosine kernel, which is the conjugate kernel of an infinitely wide shallow ReLU network with two inputs and no biases in the hidden layer (Cho and Saul, 2009).

2 Bayesian learning and generalization error for GPs

In GP regression, our goal is to learn a target function $f\colon \Omega \mapsto \mathbb{R}$ between an input $x \in \Omega$ and output $y \in \mathbb{R}$ based on training samples $D_n = \{(x_i, y_i)\}_{i=1}^n$. We consider an additive noise model $y_i = f(x_i) + \epsilon_i$, where $\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$. If ρ denotes the marginal density of the inputs x_i, then the pairs (x_i, y_i) are generated according to the density $q(x, y) = \rho(x)q(y|x)$, where $q(y|x) = \mathcal{N}(y|f(x), \sigma^2)$. We assume that there is a prior distribution Π_0 on f which is defined as a zero-mean GP with covariance function $k : \Omega \times \Omega \to \mathbb{R}$, i.e., $f \sim \mathcal{GP}(0, k)$. This means that for any finite set $\mathbf{x} = (x_1, \ldots, x_n)^T$, the random vector $f(\mathbf{x}) = (f(x_1), \ldots, f(x_n))^T$ follows the multivariate normal distribution $\mathcal{N}(0, K_n)$ with covariance matrix $K_n = (k(x_i, x_j))_{i,j=1}^n \in \mathbb{R}^{n\times n}$. By Bayes' rule, the posterior distribution of the target f given the training data is given by

$$d\Pi_n(f|D_n) = \frac{1}{Z(D_n)} \prod_{i=1}^n \mathcal{N}(y_i|f(x_i), \sigma^2) d\Pi_0(f),$$

where $d\Pi_0(f)$ is the prior distribution and $Z(D_n) = \int \prod_{i=1}^n \mathcal{N}(y_i|f(x_i), \sigma^2) d\Pi_0(f)$ is the *marginal likelihood* or *model evidence*. The GP prior and the Gaussian noise assumption allows for exact Bayesian inference and the posterior distribution over functions is again a GP with mean and covariance function given by

$$\bar{m}(x) = K_{x\mathbf{x}}(K_n + \sigma^2 I_n)^{-1}\mathbf{y},\ x \in \Omega \tag{1}$$

$$\bar{k}(x, x') = k(x, x') - K_{x\mathbf{x}}(K_n + \sigma^2 I_n)^{-1} K_{\mathbf{x}x'},\ x, x' \in \Omega, \tag{2}$$

where $K_{x\mathbf{x}} = K_{\mathbf{x}x}^T = (k(x_1, x), \ldots, k(x_n, x))^T$ and $\mathbf{y} = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$ (Williams and Rasmussen, 2006).

The performance of GPR depends on how well the posterior approximates f as the number of training samples n tends to infinity. The distance of the posterior to the ground truth can be measured in various ways. We consider two such measures, namely the Bayesian generalization error (Seeger et al., 2008; Haussler and Opper, 1997; Opper and Vivarelli, 1999) and the excess mean square error (Sollich and Halees, 2002; Le Gratiet and Garnier, 2015; Bordelon et al., 2020; Cui et al., 2021).

Definition 1 (Bayesian generalization error)**.** *The Bayesian generalization error is defined as the Kullback-Leibler divergence between the true density $q(y|x)$ and the Bayesian predictive density $p_n(y|x, D_n) = \int p(y|f(x))d\Pi_n(f|D_n)$,*

$$G(D_n) = \int q(x, y) \log \frac{q(y|x)}{p_n(y|x, D_n)} dxdy. \tag{3}$$

A related quantity of interest is the *stochastic complexity* (SC), also known as the *free energy*, which is just the negative log-marginal likelihood. We shall primarily be concerned with a normalized version of the stochastic complexity which is defined as follows:

$$F^0(D_n) = -\log \frac{Z(D_n)}{\prod_{i=1}^n q(y_i|x_i)} = -\log \frac{\int \prod_{i=1}^n \mathcal{N}(y_i|f(x_i), \sigma^2) d\Pi_0(f)}{\prod_{i=1}^n q(y_i|x_i)}. \tag{4}$$

The generalization error (3) can be expressed in terms of the normalized SC as follows (Watanabe, 2009, Theorem 1.2):

$$G(D_n) = \mathbb{E}_{(x_{n+1},y_{n+1})}F^0(D_{n+1}) - F^0(D_n), \tag{5}$$

where $D_{n+1} = D_n \cup \{(x_{n+1}, y_{n+1})\}$ is the dataset obtained by augmenting D_n with a test point (x_{n+1}, y_{n+1}).

If we only wish to measure the performance of the mean of the Bayesian predictive density, then we can use the excess mean square error:

Definition 2 (Excess mean square error)**.** *The excess mean square error is defined as*

$$M(D_n) = \mathbb{E}_{(x_{n+1},y_{n+1})}(\bar{m}(x_{n+1}) - y_{n+1})^2 - \sigma^2 = \mathbb{E}_{x_{n+1}}(\bar{m}(x_{n+1}) - f(x_{n+1}))^2. \tag{6}$$

Proposition 3 (Normalized stochastic complexity for GPR)**.** *The normalized SC $F^0(D_n)$ (4) for GPR with prior $\mathcal{GP}(0, k)$ is given as*

$$F^0(D_n) = \frac{1}{2}\log\det(I + \tfrac{K_n}{\sigma^2}) + \frac{1}{2\sigma^2}\mathbf{y}^T(I + \tfrac{K_n}{\sigma^2})^{-1}\mathbf{y} - \frac{1}{2\sigma^2}(\mathbf{y} - f(\mathbf{x}))^T(\mathbf{y} - f(\mathbf{x})), \tag{7}$$

where $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$. The expectation of the normalized SC w.r.t. the noise $\boldsymbol{\epsilon}$ is given as

$$\mathbb{E}_{\boldsymbol{\epsilon}}F^0(D_n) = \frac{1}{2}\log\det\big(I + \tfrac{K_n}{\sigma^2}\big) - \frac{1}{2}\mathrm{Tr}\Big(I - \big(I + \tfrac{K_n}{\sigma^2}\big)^{-1}\Big) + \frac{1}{2\sigma^2}f(\mathbf{x})^T\big(I + \tfrac{K_n}{\sigma^2}\big)^{-1}f(\mathbf{x}). \tag{8}$$

This is a basic result and has applications in relation to model selection in GPR (Williams and Rasmussen, 2006). For completeness, we give a proof of Proposition 3 in Appendix A. Seeger et al. (2008, Theorem 1) gave an upper bound on the normalized stochastic complexity for the case when f lies in the reproducing kernel Hilbert space (RKHS) of the GP prior. It is well known, however, that sample paths of GP almost surely fall outside the corresponding RKHS (Van Der Vaart and Van Zanten, 2011) limiting the applicability of the result.

In the next section, we derive the asymptotics of $\mathbb{E}_{\boldsymbol{\epsilon}}F^0(D_n)$, the expected generalization error $\mathbb{E}_{\boldsymbol{\epsilon}}G(D_n) = \mathbb{E}_{\boldsymbol{\epsilon}}\mathbb{E}_{(x_{n+1},y_{n+1})}F^0(D_n+1) - \mathbb{E}_{\boldsymbol{\epsilon}}F^0(D_n)$, and the excess mean square error $\mathbb{E}_{\boldsymbol{\epsilon}}M(D_n)$.

3 Asymptotic analysis of GP regression with power-law priors

We begin by introducing some notations and assumptions. According to Mercer's theorem, the covariance function of the GP prior can be decomposed as $k(x_1, x_2) = \sum_{p=1}^{\infty}\lambda_p\phi_p(x_1)\phi_p(x_2)$, where $(\phi_p(x))_{p\geq 1}$ are the eigenfunctions of the operator $L_k\colon L^2(\Omega,\rho) \mapsto L^2(\Omega,\rho)$; $(L_k f)(x) = \int_\Omega k(x,s)f(s)\mathrm{d}\rho(s)$, and $(\lambda_p)_{p\geq 1}$ are the corresponding positive eigenvalues (Williams and Rasmussen, 2006). We index the sequence of eigenvalues in decreasing order, that is $\lambda_1 \geq \lambda_2 \geq \cdots > 0$. The target function $f(x)$ is decomposed into the orthonormal set $(\phi_p(x))_{p\geq 1}$ and its orthogonal complement $\{\phi_p(x) : p \geq 1\}^\perp$ as

$$f(x) = \sum_{p=1}^{\infty}\mu_p\phi_p(x) + \mu_0\phi_0(x) \in L^2(\Omega,\rho), \tag{9}$$

where $\mu = (\mu_0, \mu_1, \ldots, \mu_p, \ldots)^T$ are the coefficients of the decomposition, and $\phi_0(x)$ satisfies $\|\phi_0(x)\|_2 = 1$ and $\phi_0(x) \in \{\phi_p(x) : p \geq 1\}^\perp$. For given sample inputs $\mathbf{x}$, let $\phi_p(\mathbf{x}) = (\phi_p(x_1), \ldots, \phi_p(x_n))^T$, $\Phi = (\phi_0(\mathbf{x}), \phi_1(\mathbf{x}), \ldots, \phi_p(\mathbf{x}), \ldots)$ and $\Lambda = \mathrm{diag}\{0, \lambda_1, \ldots, \lambda_p, \ldots\}$. Then the covariance matrix K_n can be written as $K_n = \Phi\Lambda\Phi^T$, and the function values on the sample inputs can be written as $f(\mathbf{x}) = \Phi\mu$.

We shall make the following assumptions in order to derive the power-law asymptotics of the normalized stochastic complexity and the generalization error of GPR:

Assumption 4 (Power law decay of eigenvalues)**.** *The eigenvalues $(\lambda_p)_{p\geq 1}$ follow the power law*

$$\underline{C_\lambda}p^{-\alpha} \leq \lambda_p \leq \overline{C_\lambda}p^{-\alpha}, \tag{10}$$

where $\underline{C_\lambda}$, $\overline{C_\lambda}$ and α are three positive constants which satisfy $0 < \underline{C_\lambda} \leq \overline{C_\lambda}$ and $\alpha > 1$.

As mentioned in the introduction, this assumption is adopted in many recent works (Bahri et al., 2021; Canatar et al., 2021; Nitanda and Suzuki, 2021). Velikanov and Yarotsky (2021) derived the exact value of the exponent α when the kernel function has a homogeneous singularity on its diagonal, which is the case for instance for the arc-cosine kernel.

Assumption 5 (Power law decay of coefficients of decomposition)**.** *The coefficients $(\mu_p)_{p\geq 1}$ of the decomposition* (9) *of the target function follow the power law*

$$|\mu_p| \leq C_\mu p^{-\beta}, \tag{11}$$

where C_μ and β are two positive constants which satisfy $C_\mu > 0$ and $\beta > 1/2$.

When the orthonormal basis $(\phi_p(x))_p$ is the Fourier basis or the spherical harmonics basis, the coefficients $(\mu_p)_p$ decay at least as fast as a power law so long as the target function $f(x)$ satisfies certain smoothness conditions (Bietti and Mairal, 2019). Velikanov and Yarotsky (2021) gave examples of some natural classes of functions for which Assumption 5 is satisfied, such as functions that have a bounded support with smooth boundary and are smooth on the interior of this support, and derived the corresponding exponents β.

Assumption 6 (Boundedness of the target function and eigenfunctions)**.** *The target function $f(x)$ and the eigenfunctions $(\phi_p(x))_{p\geq 1}$ satisfy*

$$\|f\|_\infty \leq C_f, \quad \|\phi_p\|_\infty \leq C_\phi, \quad \forall p \geq 1,$$

where C_f and C_ϕ are positive constants.

3.1 Asymptotics of the normalized stochastic complexity

In this section we derive the asymptotics of the normalized SC (8). Define

$$T_1(D_n) = \frac{1}{2}\log\det(I + \frac{\Phi\Lambda\Phi^T}{\sigma^2}) - \frac{1}{2}\mathrm{Tr}(I - (I + \frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}), \tag{12}$$

$$T_2(D_n) = \frac{1}{2\sigma^2} f(\mathbf{x})^T (I + \frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1} f(\mathbf{x}). \tag{13}$$

We then have $\mathbb{E}_\epsilon F_n^0 = T_1(D_n) + T_2(D_n)$. We shall analyze the asymptotics of $T_1(D_n)$ and $T_2(D_n)$ for the following two cases: $\mu_0 > 0$ and $\mu_0 = 0$.

In the following, we regard α, β, and σ as constants and omit them in the big-O notation. We also suppress polylogarithmic factors in the big-O-tilde notation.

First, we consider the case $\mu_0 = 0$. In this case, the target function $f(x)$ lies in the span of all eigenfunctions with positive eigenvalues.

Theorem 7 (Asymptotics of the normalized SC, $\mu_0 = 0$)**.** *Assume that $\mu_0 = 0$. Under Assumptions 4, 5 and 6, with probability at least $1-\delta$ over sample inputs $(x_i)_{i=1}^n$, the expectation of the normalized SC* (4) *w.r.t. the noise ϵ has the following asymptotic behavior:*

$$\mathbb{E}_\epsilon F^0(D_n) = \tilde{O}(n^{\max\{\frac{1}{\alpha}, \frac{1-2\beta}{\alpha}+1\}}). \tag{14}$$

The complete proof of Theorem 7 is given in Appendix C.1. We give a sketch of the proof below.

Proof sketch of Theorem 7. We first note that $\mathbb{E}_\epsilon F_n^0 = T_1(D_n) + T_2(D_n)$. In order to analyze the terms $T_1(D_n)$ and $T_2(D_n)$, we will consider truncated versions of these quantities and bound the corresponding errors.

Given a truncation parameter $R \in \mathbb{N}$, let $\Phi_R = (\phi_0(\mathbf{x}), \phi_1(\mathbf{x}), \ldots, \phi_R(\mathbf{x}))$, $\Lambda_R = (0, \lambda_1, \ldots, \lambda_R)$ and $\mu_R = (\mu_0, \mu_1, \ldots, \mu_R)$. We define the truncated version of $T_1(D_n)$ as follows:

$$T_{1,R}(D_n) = \frac{1}{2}\log\det(I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2}) - \frac{1}{2}\mathrm{Tr}(I - (I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}). \tag{15}$$

Similarly, define $\Phi_{>R} = (\phi_{R+1}(\mathbf{x}), \phi_{R+2}(\mathbf{x}), \ldots, \phi_p(\mathbf{x}), \ldots)$, $\Lambda_{>R} = (\lambda_{R+1}, \ldots, \lambda_p, \ldots)$, $f_R(x) = \sum_{p=1}^R \mu_p\phi_p(x)$, $f_R(\mathbf{x}) = (f_R(x_1), \ldots, f_R(x_n))^T$, $f_{>R}(x) = f(x) - f_R(x)$, and $f_{>R}(\mathbf{x}) = (f_{>R}(x_1), \ldots, f_{>R}(x_n))^T$. The truncated version of $T_2(D_n)$ is then defined as

$$T_{2,R}(D_n) = \frac{1}{2\sigma^2} f_R(\mathbf{x})^T (I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1} f_R(\mathbf{x})^T. \tag{16}$$

Let $R = n^{\frac{2}{1+\alpha}-\kappa}$ where $0 \leq \kappa < \frac{\alpha-1}{\alpha(1+\alpha)}$. The proof of consists of three steps:

- In the first step, we bound the residuals $|T_{1,R}(D_n) - T_1(D_n)|$ and $|T_{2,R}(D_n) - T_2(D_n)|$. The key component of this step is Lemma 22, where we show that

$$\|\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T\|_2 = \tilde{O}(\max\{nR^{-\alpha}, n^{\frac{1}{2}}R^{\frac{1-2\alpha}{2}}, R^{1-\alpha}\}).$$

  This allows us to show that with high probability over the draw of the inputs $(x_i)_{i=1}^n$, for sufficiently large n the residuals are bounded as (see Lemma 28):

$$|T_{1,R}(D_n) - T_1(D_n)| = O(nR^{1-\alpha}) = O\Big(n^{\frac{3-\alpha}{1+\alpha}+\kappa(\alpha-1)}\Big),$$
$$|T_{2,R}(D_n) - T_2(D_n)| = \tilde{O}\bigg(\frac{1}{\delta}n^{\max\{(\frac{2}{1+\alpha}-\kappa)\frac{1-2\beta}{2}, 1+(\frac{3\alpha+1}{\alpha(1+\alpha)}-\kappa)\frac{1-2\beta}{2}, -1-C, 1+\frac{1-2\beta}{\alpha}-C\}}\bigg),$$

  where $C = \frac{\alpha-1}{\alpha+1} - \kappa\alpha$ is a positive constant.

- In the second step, we bound the truncated components, $T_{1,R}$ and $T_{2,R}$. In Lemma 30 we show that

$$T_{1,R}(D_n) = \frac{1}{2}\log\det(I + \frac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R) - \frac{1}{2}\mathrm{Tr}\Phi_R(\sigma^2 I + \Lambda_R\Phi_R^T\Phi_R)^{-1}\Lambda_R\Phi_R^T,$$
$$T_{2,R}(D_n) = \frac{1}{2}\mu_R^T\Phi_R^T\Phi_R(\sigma^2 I + \Lambda_R\Phi_R^T\Phi_R)^{-1}\mu_R.$$

  Then the fact that $\Phi_R^T\Phi_R$ concentrates around nI (see Lemma 20) allows us to show that

$$T_{1,R}(D_n) = O(n^{\frac{1}{\alpha}}), \quad \text{and} \quad T_{2,R}(D_n) = \tilde{O}(n^{\max\{0, 1+\frac{1-2\beta}{\alpha}\}}).$$

- In the third step, we compare the orders of the asymptotics of $T_{1,R}$ and $T_{2,R}$, and the residuals $|T_{1,R}(D_n) - T_1(D_n)|$ and $|T_{2,R}(D_n) - T_2(D_n)|$. We show in Lemma 31 that $T_{1,R}$ dominates the residual, and likewise for $T_{2,R}$, when we have

$$\mathbb{E}_\epsilon F^0(D_n) = T_1(D_n) + T_2(D_n) = O(n^{\frac{1}{\alpha}}) + \tilde{O}(n^{\max\{0, 1+\frac{1-2\beta}{\alpha}\}}) = \tilde{O}(n^{\max\{\frac{1}{\alpha}, \frac{1-2\beta}{\alpha}+1\}}).$$

□

For $\mu_0 > 0$, we note the following result:

Theorem 8 (Asymptotics of the normalized SC, $\mu_0 > 0$)**.** *Assume that $\mu_0 > 0$. Under Assumptions 4, 5 and 6, the normalized SC* (8) *has the following asymptotic behavior:*

$$\mathbb{E}_\epsilon F^0(D_n) = \tilde{O}(n). \tag{17}$$

Proof of Theorem 8. $T_2(D_n)$ is estimated as follows:

$$T_2(D_n) = f(\mathbf{x})^T(I + \frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f(\mathbf{x}) \leq \|f(\mathbf{x})\|_2^2\|_2(I + \frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}\| \leq \|f(\mathbf{x})\|_2^2 \leq n\|f\|_\infty^2 = O(n).$$

By Lemma 31, we have $T_1(D_n) = O(n^{\frac{1}{\alpha}})$. Hence $\mathbb{E}_\epsilon F^0(D_n) = T_1(D_n) + T_2(D_n) = O(n)$. □

3.2 Asymptotics of the Bayesian generalization error

In this section we derive the asymptotics of the Bayesian generalization error (3). Define

$$G_1(D_n) = \mathbb{E}_{\boldsymbol{\epsilon}}\mathbb{E}_{(x_{n+1},y_{n+1})}(T_1(D_{n+1}) - T_1(D_n)), \tag{18}$$
$$G_2(D_n) = \mathbb{E}_{\boldsymbol{\epsilon}}\mathbb{E}_{(x_{n+1},y_{n+1})}(T_2(D_{n+1}) - T_2(D_n)). \tag{19}$$

Using (5), we have

$$\mathbb{E}_{\boldsymbol{\epsilon}}G(D_n) = \mathbb{E}_{\boldsymbol{\epsilon}}\mathbb{E}_{(x_{n+1},y_{n+1})}F^0(D_{n+1}) - \mathbb{E}_{\boldsymbol{\epsilon}}F^0(D_n) = G_1(D_n) + G_2(D_n). \tag{20}$$

We derive the asymptotics of the expected generalization error (20) by analyzing the asymptotics of the components $G_1(D_n)$ and $G_2(D_n)$ for the following two cases: $\mu_0 = 0$ and $\mu_0 > 0$. First, we consider the case $\mu_0 = 0$.

Theorem 9 (Asymptotics of the Bayesian generalization error, $\mu_0 = 0$)**.** *Under Assumptions 4, 5 and 6, with probability at least $1-\delta$ over sample inputs $(x_i)_{i=1}^n$, the expectation of the Bayesian generalization error* (3) *w.r.t. the noise $\boldsymbol{\epsilon}$ has the following asymptotic behavior:*

$$\mathbb{E}_{\boldsymbol{\epsilon}} G(D_n) = O(n^{\max\{\frac{1}{\alpha}-1, \frac{1-2\beta}{\alpha}\}}). \tag{21}$$

Intuitively, the exponent $\frac{1-\alpha}{\alpha}$ captures the rate at which the model suppresses the noise, while the exponent $\frac{1-2\beta}{\alpha}$ captures the rate at which the model learns the target function. So, larger the value of β, the easier we learn the target.

The complete proof of Theorem 9 is given in Appendix C.2. We give a brief proof sketch below.

Proof sketch of Theorem 9. Let $R = n^{\zeta}$, where ζ is a constant. The proof consists of three steps:

- In the first step, we consider the truncated versions of $G_1(D_n)$ and $G_2(D_n)$, which are, respectively, $G_{1,R}(D_n) = \mathbb{E}_{(x_{n+1},y_{n+1})}(T_{1,R}(D_{n+1}) - T_{1,R}(D_n))$, and $G_{2,R}(D_n) = \mathbb{E}_{(x_{n+1},y_{n+1})}(T_{2,R}(D_{n+1}) - T_{2,R}(D_n))$.

  In Lemmas 32 and 34, we show that

  $$G_{1,R}(D_n) \leq \left|\frac{1}{2\sigma^2}\operatorname{Tr}(I + \frac{\Lambda_R \Phi_R^T \Phi_R}{\sigma^2})^{-1}\Lambda_R\right| + \left|\frac{1}{2\sigma^2}\operatorname{Tr}(I + \frac{\Lambda_R \Phi_R^T \Phi_R}{\sigma^2})^{-1}\Lambda_R(I + \frac{\Lambda_R \Phi_R^T \Phi_R}{\sigma^2})^{-1}\right|,$$

  $$G_{2,R}(D_n) \leq \frac{1}{2\sigma^2}\|(I + \frac{1}{\sigma^2}\Lambda_R \Phi_R^T \Phi_R)^{-1}\mu_R\|_2^2.$$

  Using the fact that $\Phi_R^T \Phi_R$ concentrates around nI (see Lemma 20), we show $G_{1,R}(D_n) = O(n^{\frac{1-\alpha}{\alpha}} \log R)$ and $G_{2,R}(D_n) = \tilde{O}(n^{\max\{-2, \frac{1-2\beta}{\alpha}\}})$.

- In the second step, we show in Lemma 28 that the residual associated with G_1, $|G_1(D_n) - G_{1,R}(D_n)|$ behaves as $O(nR^{1-\alpha})$. For the residual associated with G_2, we show in Lemma 29 that $|G_{2,R}(D_n) - G_2(D_n)|$ behaves as $\tilde{O}\big(\frac{1}{\delta} n R^{\max\{1/2-\beta, 1-\alpha\}}\big)$.

  Choosing $\zeta > \max\{\frac{2\alpha-1}{\alpha(\alpha-1)}, \frac{1+\min\{2, \frac{2\beta-1}{\alpha}\}}{\min\{\beta-1/2, \alpha-1\}} + 1\}$, we get that $G_1(D_n) = \tilde{O}(n^{\frac{1-\alpha}{\alpha}})$, and $G_2(D_n) = \tilde{O}(n^{\max\{-2, \frac{1-2\beta}{\alpha}\}})$.

- Combining the steps above, we get with high probability over the draw of the inputs $(x_i)_{i=1}^n$, $\mathbb{E}_{\boldsymbol{\epsilon}} G(D_n) = G_1(D_n) + G_2(D_n) = O(n^{\max\{\frac{1}{\alpha}-1, \frac{1-2\beta}{\alpha}\}})$. □

Remark 10. *If $f \sim \mathcal{GP}(0, k)$, then using the Karhunen-Loève expansion we have $f(x) = \sum_{p=1}^{\infty} \sqrt{\lambda_p}\omega_p \phi_p(x)$, where $(\omega_p)_{p=1}^{\infty}$ are i.i.d standard Gaussian variables. Comparing with the expansion of $f(x)$ in* (9)*, we find that $\mu_p = \sqrt{\lambda_p}\omega_p \sim p^{-\alpha/2}$ for $p \geq 1$, and hence $\beta = \alpha/2$, and $\mathbb{E}_{\boldsymbol{\epsilon}} G(D_n) = O(n^{\frac{1}{\alpha}-1})$. This rate matches that of an earlier result due to Sollich and Halees (2002), where it is shown that the asymptotic learning curve (as measured by the expectation of the excess mean square error, $\mathbb{E}_f M(D_n)$) scales as $n^{\frac{1}{\alpha}-1}$ when the model is correctly specified, i.e., f is a sample from the same Gaussian process $\mathcal{GP}(0, k)$, and the eigenvalues decay as a power law for large i, $\lambda_i \sim i^{\alpha}$.*

For $\mu_0 > 0$, we note the following result:

Theorem 11 (Asymptotics of the Bayesian generalization error, $\mu_0 > 0$)**.** *Assume that $\mu_0 > 0$. Under Assumptions 4, 5 and 6, the Bayesian generalization error* (3) *has the following asymptotic behavior:*

$$\mathbb{E}_{\boldsymbol{\epsilon}} G(D_n) = \tilde{O}(1). \tag{22}$$

Proof of Theorem 11. According to (98) in the proof of Lemma 34, $G_2(D_n)$ is estimated as follows:

$$G_2(D_n) = \lim_{S\to\infty} \frac{1}{2\sigma^2}\|(I + \frac{1}{\sigma^2}\Lambda_S \Phi_S^T \Phi_S)^{-1}\mu_S\|_2^2 \leq \lim_{S\to\infty} \frac{1}{2\sigma^2}\|\mu_S\|_2^2 = \frac{1}{2\sigma^2}\|f\|_2^2 = O(1).$$

By Lemma 32, we have $G_1(D_n) = O(n^{\frac{1-\alpha}{\alpha}})$. Then $\mathbb{E}_{\epsilon} G(D_n) = G_1(D_n) + G_2(D_n) = O(1)$. □

In general, if $\mu_0 > 0$, then the generalization error remains constant when $n \to \infty$. This means that if the target function contains a component in the kernel of the operator L_k, GP regression is not able to learn the target function.

3.3 Asymptotics of the excess mean squared error

In this section we derive the asymptotics of the excess mean squared error in Definition 2.

Theorem 12 (Asymptotics of excess mean squared error)**.** *Under Assumptions 4, 5 and 6, if $\mu_0 = 0$, with probability at least $1 - \delta$ over sample inputs $(x_i)_{i=1}^n$, the generalization error measured by excess means squared error* (6) *has the following asymptotic:*

$$\mathbb{E}_{\boldsymbol{\epsilon}} M(D_n) = \tilde{O}\Big(n^{\max\{\frac{1}{\alpha}-1, \frac{1-2\beta}{\alpha}\}}\Big). \tag{23}$$

If $\mu_0 > 0$, the generalization error for kernel ridge regression has the following asymptotic:

$$\mathbb{E}_{\boldsymbol{\epsilon}} M(D_n) = O(1). \tag{24}$$

The proof of Theorem 12 uses similar techniques as Theorem 9 and is given in Appendix C.3.

Remark 13 (Correspondence with kernel ridge regression)**.** *The kernel ridge regression (KRR) estimator arises as a solution to the optimization problem*

$$\hat{f} = \arg\min_{f \in \mathcal{H}_k} \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}_k}^2, \tag{25}$$

where the hypothesis space $\mathcal{H}_k$ is chosen to be an RKHS, and $\lambda > 0$ is a regularization parameter. The solution to (25) *is unique as a function, and is given by $\hat{f}(x) = K_{x\mathbf{x}}(K_n + n\lambda I_n)^{-1}\mathbf{y}$, which coincides with the posterior mean function $\bar{m}(x)$ of the GPR* (1) *if $\sigma^2 = n\lambda$ (Kanagawa et al., 2018, Proposition 3.6). Thus, the additive Gaussian noise in GPR plays the role of regularization in KRR. Leveraging this well known equivalence between GPR and KRR we observe that Theorem 12 also describes the generalization error of KRR as measured by the excess mean square error.*

Remark 14. *Cui et al. (2021) study KRR with Gaussian design, where $\Lambda_R^{1/2}(\phi_1(x), \dots, \phi_R(x)))$ is assumed to follow a Gaussian distribution $\mathcal{N}(0, \Lambda_R)$. Theorem 12 recovers the result in Cui et al. (2021, Eq. 10). In contrast to the approach in Cui et al. (2021), we have no requirement on the distribution of $\phi_p(x)$, and hence our result is more generally applicable. For example, Theorem 12 can be applied to KRR with the arc-cosine kernel when the Gaussian design assumption is not valid.*

4 Experiments

We illustrate our theory on a few toy experiments. We let the input x be uniformly distributed on a unit circle, i.e., $\Omega = S^1$ and $\rho = \mathcal{U}(S^1)$. The points on S^1 can be represented by $x = (\cos\theta, \sin\theta)$ where $\theta \in [-\pi, \pi)$. We use the first order arc-cosine kernel function $k(x_1, x_2) = \frac{1}{\pi}(\psi + (\pi - \psi)\cos\psi)$, where $\psi = \langle x_1, x_2 \rangle$ is the angle between x_1 and x_2. Cho and Saul (2009) showed that this kernel is the conjugate kernel of an infinitely wide shallow ReLU network with two inputs and no biases in the hidden layer. GP regression with prior $\mathcal{GP}(0, k)$ corresponds to Bayesian training of this network (Lee et al., 2018). The eigenvalues and eigenfunctions of the kernel are $\lambda_1 = \frac{4}{\pi^2}$, $\lambda_2 = \lambda_3 = \frac{1}{4}$, $\lambda_{2p} = \lambda_{2p+1} = \frac{4}{\pi^2((2p-2)^2-1)^2}$, $p \geq 2$ and $\phi_1(\theta) = 1$, $\phi_2(\theta) = \frac{\sqrt{2}}{2}\cos\theta$, $\phi_3(\theta) = \frac{\sqrt{2}}{2}\sin\theta$, $\phi_{2p}(\theta) = \frac{\sqrt{2}}{2}\cos(2p-2)\theta$,$\phi_{2p+1}(\theta) = \frac{\sqrt{2}}{2}\sin(2p-2)\theta$, $p \geq 2$. Hence Assumption 4 is satisfied with $\alpha = 4$, and the second part of Assumption 6 is satisfied with $\|\phi_p\| \leq \frac{\sqrt{2}}{2}$, $p \geq 1$. We consider the target functions in Table 1, which satisfy Assumption 5 with the indicated β, and μ_0 indicates whether the function lies in the span of eigenfunctions of the kernel.

The training and test data are generated as follows: We independently sample training inputs $x_1, \dots, x_n$ and test input x_{n+1} from $\mathcal{U}(S^1)$ and training outputs y_i, $i = 1, \dots, n$ from $\mathcal{N}(f(x_i), \sigma^2)$, where we choose $\sigma = 0.1$. The prediction $\bar{y}_{n+1}$ on a test point is obtained using (1). We compute the normalized SC using (7) and the excess mean square error using $(\bar{y}_{n+1} - f(x_{n+1}))^2$. For each target we conduct the experiments 20 times and report the mean and standard deviation in Figure 1, which agree with the asymptotics predicted in Theorems 7 and 9.

	function value	β	μ_0	$\mathbb{E}_\epsilon F^0(D_n)$	$\mathbb{E}_\epsilon M(D_n)$
f_1	$\cos 2\theta$	$+\infty$	0	$\tilde{O}(n^{1/4})$	$\tilde{O}(n^{-3/4})$
f_2	θ^2	2	> 0	$\tilde{O}(n)$	$\tilde{O}(1)$
f_3	$(\lvert x\rvert - \pi/2)^2$	2	0	$\tilde{O}(n^{1/4})$	$\tilde{O}(n^{-3/4})$
f_4	$\begin{cases} \pi/2 - \theta, & \theta \in [0, \pi) \\ -\pi/2 - \theta, & \theta \in [-\pi, 0) \end{cases}$	1	0	$\tilde{O}(n^{3/4})$	$\tilde{O}(n^{-1/4})$

Table 1: Target functions used in the experiments, their values of β and μ_0 for the arc-cosine kernel, and theoretical rates for the normalized SC and excess mean squared error from our theorems.

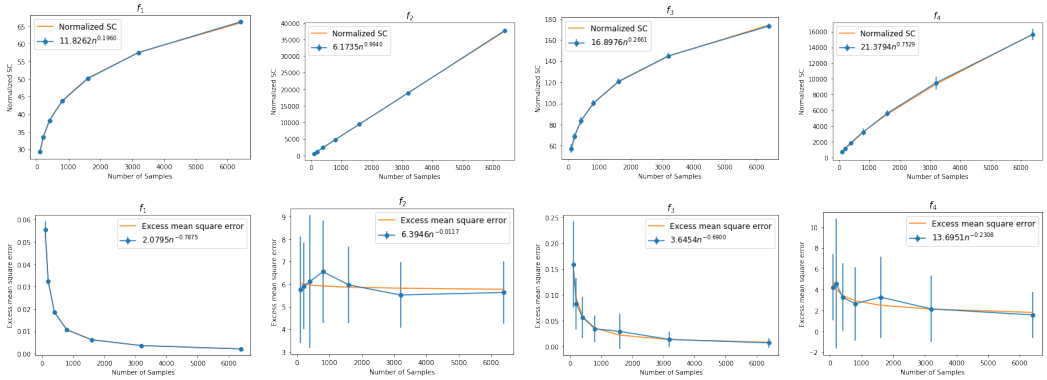

Figure 1: Normalized SC (top) and excess mean squared error (bottom) for GPR with first order arc-cosine kernel and the target functions in Table 1. The rates are in close agreement with our theory.

5 Conclusion

We described the learning curves for GPR for the case that the kernel and target function follow a power law. This setting is frequently encountered in kernel learning and relates to recent advances on neural networks. Our approach is based on a tight analysis of the concentration of the inner product of empirical eigenfunctions $\Phi^T\Phi$ around nI. This allowed us to obtain more general results with more realistic assumptions than previous works. In particular, we recovered some results on learning curves previously obtained under more restricted settings (vide Remarks 14 and 10).

We showed that when $\beta \geq \alpha/2$, meaning that the target function has a compact representation in terms of the eigenfunctions of the kernel, the learning rate is as good as in the correctly specified case. In addition, our result allows us to interpret β from a spectral bias perspective. When $\frac{1}{2} < \beta \leq \frac{\alpha}{2}$, the larger the value of β, the faster the decay of the generalization error. This implies that low-frequency functions are learned faster in terms of the number of training data points.

By leveraging the equivalence between GPR and KRR, we obtained a result on the generalization error of KRR. In the infinite-width limit, training fully-connected deep NNs with gradient descent and infinitesimally small learning rate under least-squared loss is equivalent to solving KRR with respect to the NTK (Jacot et al., 2018; Lee et al., 2019; Domingos, 2020), which in several cases is known to have a power-law spectrum (Velikanov and Yarotsky, 2021). Hence our methods can be applied to study the generalization error of infinitely wide neural networks.

In future work, it would be interesting to estimate the values of α and β for the NNGP kernel and the NTK kernel of deep fully-connected or convolutional NNs and real data distributions and test our theory in these cases. Similarly, it would be interesting to consider extensions to finite width kernels.

REFERENCES

S. Amari and N. Murata. Statistical theory of learning curves under entropic loss criterion. *Neural Computation*, 5(1):140–153, 1993.

S. Amari, N. Fujita, and S. Shinomoto. Four types of learning curves. *Neural Computation*, 4(4): 605–618, 1992.

S. Arora, S. S. Du, W. Hu, Z. Li, R. R. Salakhutdinov, and R. Wang. On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems*, pages 8139–8148, 2019.

Y. Bahri, E. Dyer, J. Kaplan, J. Lee, and U. Sharma. Explaining neural scaling laws. *arXiv preprint arXiv:2102.06701*, 2021.

A. R. Barron. Information-theoretic characterization of Bayes performance and the choice of priors in parametric and nonparametric problems. In D. A. Bernardo J., Berger J. and S. A., editors, *Bayesian statistics*, volume 6, pages 27–52. Oxford University Press, 1998.

M. Belkin, S. Ma, and S. Mandal. To understand deep learning we need to understand kernel learning. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 541–549, 2018.

A. Bietti and J. Mairal. On the inductive bias of neural tangent kernels. In *Advances in Neural Information Processing Systems*, volume 32, pages 12873–12884, 2019.

B. Bordelon, A. Canatar, and C. Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 1024–1034, 2020.

O. Bousquet, S. Hanneke, S. Moran, R. van Handel, and A. Yehudayoff. A theory of universal learning. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, pages 532–541, 2021.

A. Canatar, B. Bordelon, and C. Pehlevan. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nature communications*, 12(1): 1–12, 2021.

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

H. Cui, B. Loureiro, F. Krzakala, and L. Zdeborová. Generalization error rates in kernel regression: The crossover from the noiseless to noisy regime. *arXiv preprint arXiv:2105.15004*, 2021.

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

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

P. Domingos. Every model learned by gradient descent is approximately a kernel machine. *arXiv preprint arXiv:2012.00152*, 2020.

A. Garriga-Alonso, C. E. Rasmussen, and L. Aitchison. Deep convolutional networks as shallow gaussian processes. In *International Conference on Learning Representations*, 2019.

D. Haussler and M. Opper. Mutual information, metric entropy and cumulative relative entropy risk. *The Annals of Statistics*, 25(6):2451–2492, 1997.

D. Haussler, M. Kearns, H. S. Seung, and N. Tishby. Rigorous learning curve bounds from statistical mechanics. *Machine Learning*, 25(2-3):195–236, 1996.

J. Hestness, S. Narang, N. Ardalani, G. Diamos, H. Jun, H. Kianinejad, M. Patwary, M. Ali, Y. Yang, and Y. Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

M. Kanagawa, P. Hennig, D. Sejdinovic, and B. K. Sriperumbudur. Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv preprint arXiv:1807.02582*, 2018.

L. Le Gratiet and J. Garnier. Asymptotic analysis of the learning curve for Gaussian process regression. *Machine Learning*, 98(3):407–433, 2015.

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

J. Lee, L. Xiao, S. Schoenholz, Y. Bahri, R. Novak, J. Sohl-Dickstein, and J. Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

J. Lee, S. Schoenholz, J. Pennington, B. Adlam, L. Xiao, R. Novak, and J. Sohl-Dickstein. Finite versus infinite neural networks: an empirical study. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15156–15172. Curran Associates, Inc., 2020.

M. Loog, T. Viering, and A. Mey. Minimizers of the empirical risk and risk monotonicity. In *Advances in Neural Information Processing Systems*, volume 32, pages 7478–7487, 2019.

D. Malzahn and M. Opper. Learning curves for Gaussian processes regression: A framework for good approximations. In *Advances in Neural Information Processing Systems*, pages 273–279, 2001a.

D. Malzahn and M. Opper. Learning curves for Gaussian processes models: Fluctuations and universality. In *International Conference on Artificial Neural Networks*, pages 271–276, 2001b.

R. M. Neal. *Bayesian Learning for Neural Networks*. Springer-Verlag, Berlin, Heidelberg, 1996. ISBN 0387947248.

A. Nitanda and T. Suzuki. Optimal rates for averaged stochastic gradient descent under neural tangent kernel regime. In *International Conference on Learning Representations*, 2021.

R. Novak, L. Xiao, Y. Bahri, J. Lee, G. Yang, D. A. Abolafia, J. Pennington, and J. Sohl-Dickstein. Bayesian deep convolutional networks with many channels are gaussian processes. In *International Conference on Learning Representations*, 2019.

M. Opper and D. Malzahn. A variational approach to learning curves. In *Advances in Neural Information Processing Systems*, volume 14, pages 463–469, 2002.

M. Opper and F. Vivarelli. General bounds on Bayes errors for regression with Gaussian processes. In *Advances in Neural Information Processing Systems*, volume 11, pages 302–308, 1999.

P. Orbanz and Y. W. Teh. Bayesian nonparametric models. In *Encyclopedia of Machine Learning*, pages 81–89. Springer, 2010.

K. Ritter, G. W. Wasilkowski, and H. Woźniakowski. Multivariate integration and approximation for random fields satisfying Sacks-Ylvisaker conditions. *The Annals of Applied Probability*, pages 518–540, 1995.

B. Ronen, D. Jacobs, Y. Kasten, and S. Kritchman. The convergence rate of neural networks for learned functions of different frequencies. *Advances in Neural Information Processing Systems*, 32:4761–4771, 2019.

M. W. Seeger, S. M. Kakade, and D. P. Foster. Information consistency of nonparametric Gaussian process methods. *IEEE Transactions on Information Theory*, 54(5):2376–2382, 2008.

P. Sollich. Learning curves for Gaussian processes. In *Advances in Neural Information Processing Systems*, pages 344–350, 1999.

P. Sollich. Gaussian process regression with mismatched models. In *Advances in Neural Information Processing Systems*, pages 519–526, 2001.

P. Sollich and A. Halees. Learning curves for Gaussian process regression: Approximations and bounds. *Neural Computation*, 14(6):1393–1428, 2002.

S. Spigler, M. Geiger, and M. Wyart. Asymptotic learning curves of kernel methods: empirical data versus teacher–student paradigm. *Journal of Statistical Mechanics: Theory and Experiment*, 2020 (12):124001, 2020.

M. L. Stein. *Interpolation of spatial data: Some theory for kriging*. Springer Science & Business Media, 2012.

J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

E. A. Valdivia. Relative concentration bounds for the spectrum of kernel matrices. *arXiv preprint arXiv:1812.02108*, 2018.

A. Van Der Vaart and H. Van Zanten. Information rates of nonparametric Gaussian process methods. *Journal of Machine Learning Research*, 12(6), 2011.

M. Velikanov and D. Yarotsky. Universal scaling laws in the gradient descent training of neural networks. *arXiv preprint arXiv:2105.00507*, 2021.

T. Viering and M. Loog. The shape of learning curves: A review. *arXiv preprint arXiv:2103.10948*, 2021.

T. Viering, A. Mey, and M. Loog. Open problem: Monotonicity of learning. In *Conference on Learning Theory*, pages 3198–3201, 2019.

S. Watanabe. *Algebraic geometry and statistical learning theory*. Number 25. Cambridge university press, 2009.

H. Widom. Asymptotic behavior of the eigenvalues of certain integral equations. *Transactions of the American Mathematical Society*, 109(2):278–295, 1963.

C. K. Williams. Computing with infinite networks. In M. C. Mozer, M. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9. MIT Press, 1997.

C. K. Williams and C. E. Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press, 2006.

C. K. Williams and F. Vivarelli. Upper and lower bounds on the learning curve for Gaussian processes. *Machine Learning*, 40(1):77–102, 2000.

G. Yang. Wide feedforward or recurrent neural networks of any architecture are gaussian processes. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

G. Yang and H. Salman. A fine-grained spectral perspective on neural networks. *arXiv preprint arXiv:1907.10599*, 2019.

APPENDIX

A PROOFS RELATED TO THE MARGINAL LIKELIHOOD

Proof of Proposition 3. For brevity, we write F_n^0 for $F^0(D_n)$. Let $\bar{\mathbf{y}} = (\bar{y}_1, \ldots, \bar{y}_n)^T$ be the outputs of the GP regression model on training inputs $\mathbf{x}$. Under the GP prior, the prior distribution of $\bar{\mathbf{y}}$ is $\mathcal{N}(0, K_n)$. Then the evidence of the model is given as follows:

$$\begin{aligned} Z_n &= \int_{\mathbb{R}^n} \left(\prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\bar{y}_i - y_i)^2}{2\sigma^2}} \right) \frac{1}{(2\pi)^{n/2} \det(K_n)^{1/2}} e^{-\frac{1}{2}\bar{\mathbf{y}}^T K_n^{-1} \bar{\mathbf{y}}} \mathrm{d}\bar{\mathbf{y}} \\ &= \frac{1}{(2\pi)^n \sigma^n \det(K_n)^{1/2}} \int_{\mathbb{R}^n} e^{-\frac{1}{2}\bar{\mathbf{y}}^T (K_n^{-1} + \frac{1}{\sigma^2} I)\bar{\mathbf{y}} + \frac{1}{\sigma^2}\bar{\mathbf{y}}^T \mathbf{y} - \frac{1}{2\sigma^2}\mathbf{y}^T \mathbf{y}} \mathrm{d}\bar{\mathbf{y}}. \end{aligned} \tag{26}$$

Letting $\tilde{K}_n^{-1} = K_n^{-1} + \frac{1}{\sigma^2} I$ and $\mu = \frac{1}{\sigma^2}\tilde{K}_n \mathbf{y}$, we have

$$\begin{aligned} Z_n &= \frac{1}{(2\pi)^n \sigma^n \det(K_n)^{1/2}} \int_{\mathbb{R}^n} e^{-\frac{1}{2}(\bar{\mathbf{y}} - \mu)^T \tilde{K}_n^{-1} (\bar{\mathbf{y}} - \mu) - \frac{1}{2\sigma^2}\mathbf{y}^T \mathbf{y} + \frac{1}{2}\mu^T \tilde{K}_n^{-1} \mu} \mathrm{d}\bar{\mathbf{y}} \\ &= \frac{1}{(2\pi)^n \sigma^n \det(K_n)^{1/2}} (2\pi)^{n/2} \det(\tilde{K}_n)^{1/2} e^{-\frac{1}{2\sigma^2}\mathbf{y}^T \mathbf{y} + \frac{1}{2}\mu^T \tilde{K}_n^{-1} \mu} \\ &= \frac{\det(\tilde{K}_n)^{1/2}}{(2\pi)^{n/2} \sigma^n \det(K_n)^{1/2}} e^{-\frac{1}{2\sigma^2}\mathbf{y}^T \mathbf{y} + \frac{1}{2}\mu^T \tilde{K}_n^{-1} \mu}. \end{aligned} \tag{27}$$

The normalized evidence is

$$\begin{aligned} Z_n^0 &= \frac{Z_n}{(2\pi)^{-n/2} \sigma^{-n} e^{-\frac{1}{2\sigma^2}(\mathbf{y} - f(\mathbf{x}))^T (\mathbf{y} - f(\mathbf{x}))}} \\ &= \frac{\det(\tilde{K}_n)^{1/2}}{\det(K_n)^{1/2}} e^{-\frac{1}{2\sigma^2}\mathbf{y}^T \mathbf{y} + \frac{1}{2}\mu^T \tilde{K}_n^{-1} \mu + \frac{1}{2\sigma^2}(\mathbf{y} - f(\mathbf{x}))^T (\mathbf{y} - f(\mathbf{x}))}. \end{aligned} \tag{28}$$

So the normalized stochastic complexity is

$$\begin{aligned} F_n^0 &= -\log Z_n^0 \\ &= -\frac{1}{2}\log\det(\tilde{K}_n)^{1/2} + \frac{1}{2}\log\det(K_n)^{1/2} + \frac{1}{2\sigma^2}\mathbf{y}^T \mathbf{y} - \frac{1}{2}\mu^T \tilde{K}_n^{-1} \mu - \frac{1}{2\sigma^2}(\mathbf{y} - f(\mathbf{x}))^T (\mathbf{y} - f(\mathbf{x})) \\ &= -\frac{1}{2}\log\det(K_n^{-1} + \frac{1}{\sigma^2} I)^{-1} + \frac{1}{2}\log\det(K_n) + \frac{1}{2\sigma^2}\mathbf{y}^T \mathbf{y} - \frac{1}{2\sigma^4}\mathbf{y}^T (K_n^{-1} + \frac{1}{\sigma^2} I)^{-1} \mathbf{y} \\ &\quad - \frac{1}{2\sigma^2}(\mathbf{y} - f(\mathbf{x}))^T (\mathbf{y} - f(\mathbf{x})) \\ &= \frac{1}{2}\log\det(I + \frac{K_n}{\sigma^2}) + \frac{1}{2\sigma^2}\mathbf{y}^T (I + \frac{K_n}{\sigma^2})^{-1} \mathbf{y} - \frac{1}{2\sigma^2}(\mathbf{y} - f(\mathbf{x}))^T (\mathbf{y} - f(\mathbf{x})). \\ &= \frac{1}{2}\log\det(I + \frac{K_n}{\sigma^2}) + \frac{1}{2\sigma^2} f(\mathbf{x})^T (I + \frac{K_n}{\sigma^2})^{-1} f(\mathbf{x}) + \frac{1}{2\sigma^2}\boldsymbol{\epsilon}^T (I + \frac{K_n}{\sigma^2})^{-1} \boldsymbol{\epsilon} - \frac{1}{2\sigma^2}\boldsymbol{\epsilon}^T \boldsymbol{\epsilon} \\ &\quad + \frac{1}{2\sigma^2}\boldsymbol{\epsilon}^T (I + \frac{K_n}{\sigma^2})^{-1} f(\mathbf{x}) \\ &. \end{aligned} \tag{29}$$

After taking the expectation over noises $\boldsymbol{\epsilon}$, we get

$$\mathbb{E}_{\boldsymbol{\epsilon}} F_n^0 = \frac{1}{2}\log\det(I + \frac{K_n}{\sigma^2}) + \frac{1}{2\sigma^2} f(\mathbf{x})^T (I + \frac{K_n}{\sigma^2})^{-1} f(\mathbf{x}) - \frac{1}{2}\mathrm{Tr}(I - (I + \frac{K_n}{\sigma^2})^{-1}). \tag{30}$$

This concludes the proof. □

B HELPER LEMMAS

Lemma 15. *Given constants $a_1, a_2, s_1, s_2 > 0$, if $s_1 > 1$ and $s_2 s_3 > s_1 - 1$, we have that*

$$\sum_{i=1}^{\infty} \frac{a_1 i^{-s_1}}{(1 + a_2 n i^{-s_2})^{s_3}} = O(n^{\frac{1-s_1}{s_2}}). \tag{31}$$

If $s_1 > 1$ and $s_2 s_3 = s_1 - 1$, we have that

$$\sum_{i=1}^{\infty} \frac{a_1 i^{-s_1}}{1 + a_2 n i^{-s_2}} = O(n^{-s_3} \log n). \tag{32}$$

If $s_1 > 1$ and $s_2 s_3 < s_1 - 1$, we have that

$$\sum_{i=1}^{\infty} \frac{a_1 i^{-s_1}}{1 + a_2 n i^{-s_2}} = O(n^{-s_3}). \tag{33}$$

Overall, if $s_1 > 1$,

$$\sum_{i=1}^{\infty} \frac{a_1 i^{-s_1}}{1 + a_2 n i^{-s_2}} = \tilde{O}(n^{\max\{-s_3, \frac{1-s_1}{s_2}\}}). \tag{34}$$

Proof of Lemma 15. First, when $s_1 > 1$ and $s_2 s_3 > s_1 - 1$, we have that

$$\begin{aligned}
\sum_{i=1}^{\infty} \frac{a_1 i^{-s_1}}{(1 + a_2 n i^{-s_2})^{s_3}} &\leq \frac{a_1}{(1 + a_2 n)^{s_3}} + \int_{[1,+\infty]} \frac{a_1 x^{-s_1}}{(1 + a_2 n x^{-s_2})^{s_3}} \mathrm{d}x \\
&= \frac{a_1}{(1 + a_2 n)^{s_3}} + n^{\frac{1-s_1}{s_2}} \int_{[1,+\infty]} \frac{a_1 (\frac{x}{n^{1/s_2}})^{-s_1}}{(1 + a_2 (\frac{x}{n^{1/s_2}})^{-s_2})^{s_3}} \mathrm{d} \frac{x}{n^{1/s_2}} \\
&= \frac{a_1}{(1 + a_2 n)^{s_3}} + n^{\frac{1-s_1}{s_2}} \int_{[1/n^{1/s_2},+\infty]} \frac{a_1 x^{-s_1}}{(1 + a_2 x^{-s_2})^{s_3}} \mathrm{d}x \\
&= O(n^{\frac{1-s_1}{s_2}}).
\end{aligned}$$

Second, when $s_1 > 1$ and $s_2 s_3 = s_1 - 1$, we have that

$$\begin{aligned}
\sum_{i=1}^{\infty} \frac{a_1 i^{-s_1}}{(1 + a_2 n i^{-s_2})^{s_3}} &\leq \frac{a_1}{(1 + a_2 n)^{s_3}} + n^{\frac{1-s_1}{s_2}} \int_{[1/n^{1/s_2},+\infty]} \frac{a_1 x^{-s_1}}{(1 + a_2 x^{-s_2})^{s_3}} \mathrm{d}x \\
&\leq \frac{a_1}{(1 + a_2 n)^{s_3}} + n^{\frac{1-s_1}{s_2}} O(\log n^{(-1/s_2)}) \\
&= O(n^{-s_3} \log n).
\end{aligned}$$

Third, when $s_1 > 1$ and $s_2 s_3 < s_1 - 1$, we have that

$$\begin{aligned}
\sum_{i=1}^{\infty} \frac{a_1 i^{-s_1}}{(1 + a_2 n i^{-s_2})^{s_3}} &\leq \frac{a_1}{(1 + a_2 n)^{s_3}} + n^{\frac{1-s_1}{s_2}} \int_{[1/n^{1/s_2},+\infty]} \frac{a_1 x^{-s_1}}{(1 + a_2 x^{-s_2})^{s_3}} \mathrm{d}x \\
&\leq \frac{a_1}{(1 + a_2 n)^{s_3}} + n^{\frac{1-s_1}{s_2}} O(n^{(-1/s_2)(1-s_1+s_2 s_3)}) \\
&= O(n^{-s_3}).
\end{aligned}$$

Overall, if $s_1 > 1$,

$$\sum_{i=1}^{\infty} \frac{a_1 i^{-s_1}}{1 + a_2 n i^{-s_2}} = \tilde{O}(n^{\max\{-s_3, \frac{1-s_1}{s_2}\}}). \tag{35}$$

□

Lemma 16. *Given constants $a_1, a_2, s_1, s_2 > 0$, if $s_1 \leq 1$, we have that*

$$\sum_{i=1}^{R} \frac{a_1 i^{-s_1}}{(1 + a_2 n i^{-s_2})^{s_3}} = \tilde{O}(\max\{n^{-s_3}, R^{1-s_1}\}). \tag{36}$$

Proof of Lemma 16. First, when $s_1 \leq 1$ and $s_2 s_3 > s_1 - 1$, we have that

$$\begin{aligned}
\sum_{i=1}^{R} \frac{a_1 i^{-s_1}}{(1+a_2 n i^{-s_2})^{s_3}} &\leq \frac{a_1}{(1+a_2 n)^{s_3}} + \int_{[1,R]} \frac{a_1 x^{-s_1}}{(1+a_2 n x^{-s_2})^{s_3}} \mathrm{d}x \\
&= \frac{a_1}{(1+a_2 n)^{s_3}} + n^{\frac{1-s_1}{s_2}} \int_{[1,R]} \frac{a_1 (\frac{x}{n^{1/s_2}})^{-s_1}}{(1+a_2 (\frac{x}{n^{1/s_2}})^{-s_2})^{s_3}} \mathrm{d}\frac{x}{n^{1/s_2}} \\
&= \frac{a_1}{(1+a_2 n)^{s_3}} + n^{\frac{1-s_1}{s_2}} \int_{[1/n^{1/s_2}, R/n^{1/s_2}]} \frac{a_1 x^{-s_1}}{(1+a_2 x^{-s_2})^{s}} \mathrm{d}x \\
&= \frac{a_1}{(1+a_2 n)^{s_3}} + \tilde{O}(n^{\frac{1-s_1}{s_2}} (\frac{R}{n^{1/s_2}})^{1-s_1}) \\
&= \tilde{O}(\max\{n^{-s_3}, R^{1-s_1}\}).
\end{aligned}$$

Second, when $s_1 \leq 1$ and $s_2 s_3 \leq s_1 - 1$, we have that

$$\begin{aligned}
\sum_{i=1}^{R} \frac{a_1 i^{-s_1}}{(1+a_2 n i^{-s_2})^{s_3}} &\leq \frac{a_1}{(1+a_2 n)^{s_3}} + n^{\frac{1-s_1}{s_2}} \int_{[1/n^{1/s_2}, +\infty]} \frac{a_1 x^{-s_1}}{(1+a_2 x^{-s_2})^{s_3}} \mathrm{d}x \\
&\leq \frac{a_1}{(1+a_2 n)^{s_3}} + n^{\frac{1-s_1}{s_2}} \tilde{O}(n^{(-1/s_2)(1-s_1+s_2 s_3)} + (\frac{R}{n^{1/s_2}})^{1-s_1}) \\
&= \tilde{O}(\max\{n^{-s_3}, R^{1-s_1}\}).
\end{aligned}$$

Overall, if $s_1 \leq 1$,

$$\sum_{i=1}^{\infty} \frac{a_1 i^{-s_1}}{(1+a_2 n i^{-s_2})^{s_3}} = \tilde{O}(\max\{n^{-s_3}, R^{1-s_1}\}). \tag{37}$$

□

Lemma 17. *With probability at least $1 - \delta_1$, we have*

$$\|f(\mathbf{x})\|_2^2 = \tilde{O}\left(\frac{1}{\delta_1} n \|f\|_2^2\right)$$

Proof of Lemma 17. Given a positive number $C > 0$, applying Markov's inequality we have

$$\mathbb{P}(f_{>R}^2(X) > C) \leq \frac{1}{C}\|f\|_2^2.$$

Let A be the event that for all sample inputs $(x_i)_{i=1}^n$, $f_{>R}^2(x_i) \leq C$. Then

$$\mathbb{P}(A) \geq 1 - n\mathbb{P}(f_{>R}^2(X) > C) \geq 1 - \frac{1}{C} n \|f\|_2^2. \tag{38}$$

Define $\bar{f}_{>R}^2(x) = \min\{f_{>R}^2(x), C\}$. Since $0 \leq \bar{f}_{>R}^2(x) \leq C$, we have

$$\mathbb{E}(\bar{f}_{>R}^4(X)) \leq C\mathbb{E}(\bar{f}_{>R}^2(X)) \leq C\|f\|_2^2. \tag{39}$$

Applying Bernstein's inequality, we have

$$\begin{aligned}
\mathbb{P}(\sum_{i=1}^{n} \bar{f}_{>R}^2(x_i) > t) &\leq \exp\left(-\frac{t^2}{2(n\mathbb{E}(\bar{f}_{>R}^4(X)) + \frac{Ct}{3})}\right) \\
&\leq \exp\left(-\frac{t^2}{2(nC\|f\|_2^2 + \frac{Ct}{3})}\right) \\
&\leq \exp\left(-\frac{t^2}{4\max\{nC\|f\|_2^2, \frac{Ct}{3}\}}\right).
\end{aligned}$$

Hence, with probability at least $1 - \delta_1/2$ we have

$$\sum_{i=1}^{n} \bar{f}_{>R}^2(x_i) \leq \max\left\{\sqrt{4C \log \frac{2}{\delta_1} n \|f\|_2^2}, \frac{4C}{3} \log \frac{2}{\delta_1}\right\}. \tag{40}$$

When event A happens, $f_{>R}^2(x_i) = \bar{f}_{>R}^2(x_i)$ for all sample inputs. According to (38) and (40), with probability at least $1 - \frac{1}{C}n\|f\|_2^2 - \delta_1/2$, we have

$$\sum_{i=1}^n f_{>R}^2(x_i) = \sum_{i=1}^n \bar{f}_{>R}^2(x_i) \leq \max\left\{\sqrt{4C\log\frac{2}{\delta_1}n\|f\|_2^2}, \frac{4C}{3}\log\frac{2}{\delta_1}\right\}.$$

Choosing $C = \frac{2}{\delta_1}n\|f\|_2^2$, with probability at least $1 - \delta_1$ we have

$$\sum_{i=1}^n f_{>R}^2(x_i) = \sum_{i=1}^n \bar{f}_{>R}^2(x_i) \leq \max\left\{\sqrt{\frac{8}{\delta_1}\log\frac{2}{\delta_1}n^2\|f\|_2^4}, \frac{8}{3\delta_1}n\|f\|_2^2\log\frac{2}{\delta_1}\right\} = \tilde{O}\left(\frac{1}{\delta_1}n\|f\|_2^2\right).$$

□

Corollary 18. *With probability at least $1 - \delta_1$, we have*

$$\|f_{>R}(\mathbf{x})\|_2^2 = \tilde{O}\left(\frac{1}{\delta_1}nR^{1-2\beta}\right).$$

Proof of Corollary 18. The L_2 norm of $f_{>R}(x)$ is given by $\|f_{>R}\|_2^2 = \sum_{p=R+1}^\infty \mu_p^2 \leq \frac{C_\mu}{2\beta-1}R^{1-2\beta}$. Applying Lemma 17 we get the result. □

Corollary 19. *For any $\nu \in \mathbb{R}^R$, with probability at least $1 - \delta_1$ we have*

$$\|\Phi_R\nu\|_2^2 = \tilde{O}\left(\frac{1}{\delta_1}n\|\nu\|_2^2\right).$$

Proof of Corollary 19. Let $g(x) = \sum_{p=1}^R \nu_p\phi_p(x)$. Then $\Phi_R\nu = g(\mathbf{x})$. The L_2 norm of $g(x)$ is given by $\|g\|_2^2 = \sum_{p=1}^R \nu_p^2 = \|\nu\|_2^2$. Applying Lemma 17 we get the result. □

Next we consider the quantity, $\Phi^T\Phi - nI$. Lemma 20 is a generalized version of Valdivia (2018, Proposition 3).

Lemma 20. *Suppose that the eigenvalues $(\lambda_p)_{p\geq 1}$ satisfy Assumption 4, and the eigenfuctions satisfy Assumption 6. Let γ be a positive number such that $\gamma \leq 1$. Then with probability at least $1 - \delta$, we have*

$$\|(I + \frac{n}{\sigma^2}\Lambda_R)^{-\gamma/2}\Lambda_R^{\gamma/2}(\Phi_R^T\Phi_R - nI)\Lambda_R^{\gamma/2}(I + \frac{n}{\sigma^2}\Lambda_R)^{-\gamma/2}\|_2 \leq O(\sqrt{\log\frac{R}{\delta}}n^{\frac{1-2\gamma\alpha+\alpha}{2\alpha}}). \quad (41)$$

Proof of Lemma 20. Let $Y_j = (\phi_1(x_j), \ldots, \phi_R(x_j))^T$ and $Z_j = (I + \frac{n}{\sigma^2}\Lambda_R)^{-\gamma/2}\Lambda_R^{\gamma/2}Y_j$. It is easy to verify that $\mathbb{E}(Z_jZ_j^T) = (I + \frac{n}{\sigma^2}\Lambda_R)^{-\gamma}\Lambda_R^\gamma$. Then the left hand side of (41) is $\sum_{j=1}^n[Z_jZ_j^T - \mathbb{E}(Z_jZ_j^T)]$. We note that

$$\|Z_jZ_j^T - \mathbb{E}(Z_jZ_j^T)\|_2 \leq \max\{\|Z_jZ_j^T\|_2, \|\mathbb{E}(Z_jZ_j^T)\|_2\} \leq \max\{\|Z_j\|_2^2, \frac{\sigma^2}{n^\gamma}\}.$$

Since

$$\|Z_j\|_2^2 = \sum_{p=0}^R \frac{\lambda_p^\gamma}{(1+\frac{n}{\sigma^2}\lambda_p)^\gamma}\phi_p^2(x_j) \leq \sum_{p=0}^R C_\phi^2\frac{\lambda_p^\gamma}{(1+\frac{n}{\sigma^2}\lambda_p)^\gamma} = O(n^{\frac{1-\gamma\alpha}{\alpha}}) \quad \text{(by Lemma 15)}, \quad (42)$$

we have

$$\|Z_jZ_j^T - \mathbb{E}(Z_jZ_j^T)\|_2 \leq O(n^{\frac{1-\gamma\alpha}{\alpha}}).$$

On the other hand,

$$\mathbb{E}[(Z_jZ_j^T - \mathbb{E}(Z_jZ_j^T))^2] = \mathbb{E}[\|Z_j\|_2^2Z_jZ_j^T] - (\mathbb{E}(Z_jZ_j^T))^2.$$

Since

$$\begin{aligned}\mathbb{E}[\|Z_j\|_2^2Z_jZ_j^T] &\preccurlyeq \mathbb{E}[O(n^{\frac{1-\gamma\alpha}{\alpha}})Z_jZ_j^T], \quad \text{(by (42)}\\ &= O(n^{\frac{1-\gamma\alpha}{\alpha}})\mathbb{E}[Z_jZ_j^T],\end{aligned}$$

we have

$$\begin{aligned}\|\mathbb{E}[(Z_jZ_j^T-\mathbb{E}(Z_jZ_j^T))^2]\|_2 &\le \max\{O(n^{\frac{1-\gamma\alpha}{\alpha}})\|\mathbb{E}[Z_jZ_j^T\|_2, \frac{\sigma^4}{n^{2\gamma}}\}\\ &\le \max\{O(n^{\frac{1-\gamma\alpha}{\alpha}})\frac{\sigma^2}{n^{\gamma}}, \frac{\sigma^4}{n^{2\gamma}}\}\\ &= O(n^{\frac{1-2\gamma\alpha}{\alpha}})\end{aligned}$$

Using the matrix Bernstein theorem (Tropp, 2012, Theorem 6.1), we have

$$\begin{aligned}&\mathbb{P}(\|\sum_{j=1}^n[Z_jZ_j^T-\mathbb{E}(Z_jZ_j^T)]\|_2>t)\\ \le& R\exp\left(\frac{-t^2}{2(n\|\mathbb{E}[(Z_jZ_j^T-\mathbb{E}(Z_jZ_j^T))^2]\|_2+\frac{t\max_j\|Z_jZ_j^T-\mathbb{E}(Z_jZ_j^T)\|_2}{3})}\right)\\ \le& R\exp\left(\frac{-t^2}{2(nO(n^{\frac{1-2\gamma\alpha}{\alpha}})+\frac{tO(n^{\frac{1-\gamma\alpha}{\alpha}})}{3})}\right)\\ =& R\exp\left(\frac{-t^2}{\max\{O(n^{\frac{1-2\gamma\alpha+\alpha}{\alpha}}), tO(n^{\frac{1-\gamma\alpha}{\alpha}})\}}\right).\end{aligned}$$

Then with probability at least $1-\delta$, we have

$$\begin{aligned}\|\sum_{j=1}^n[Z_jZ_j^T-\mathbb{E}(Z_jZ_j^T)]\|_2 \le& \max\{\sqrt{\log\frac{R}{\delta}O(n^{\frac{1-2\gamma\alpha+\alpha}{\alpha}})}, \log\frac{R}{\delta}O(n^{\frac{1-\gamma\alpha}{\alpha}})\}\\ =& O(\sqrt{\log\frac{R}{\delta}}n^{\frac{1-2\gamma\alpha+\alpha}{2\alpha}}).\end{aligned}$$

□

Lemma 21. *Suppose that the eigenvalues $(\lambda_p)_{p\ge1}$ satisfy Assumption 4, and the eigenfunctions satisfy Assumption 6. Let $\Phi_{R<p\le S}=(\phi_{R+1}(\mathbf{x}),\dots,\phi_S(\mathbf{x}))$, and $\Lambda_{R<p\le S}=(\lambda_{R+1},\dots,\lambda_S)$. Then with probability at least $1-\delta$, we have*

$$\|\Lambda_{R<p\le S}^{1/2}(\Phi_{R<p\le S}^T\Phi_{R<p\le S}-nI)\Lambda_{R<p\le S}^{1/2}\|_2 \le O(\log\frac{S-R}{\delta}\max\{n^{\frac{1}{2}}R^{\frac{1-2\alpha}{2}}, R^{1-\alpha}\}). \quad (43)$$

Proof of Lemma 21. Let $Y_j=(\phi_{R+1}(x_j),\dots,\phi_S(x_j))^T$ and $Z_j=\Lambda_{R<p\le S}^{1/2}Y_j$. It is easy to verify that $\mathbb{E}(Z_jZ_j^T)=\Lambda_{R<p\le S}$. Then the left hand side of (43) is $\sum_{j=1}^n[Z_jZ_j^T-\mathbb{E}(Z_jZ_j^T)]$. We note that

$$\begin{aligned}\|Z_jZ_j^T-\mathbb{E}(Z_jZ_j^T)\|_2 &\le \max\{\|Z_jZ_j^T\|_2, \|\mathbb{E}(Z_jZ_j^T)\|_2\}\\ &\le \max\{\|Z_j\|_2^2, O(R^{-\alpha})\}.\end{aligned}$$

Since we have

$$\begin{aligned}\|Z_j\|_2^2 &= \sum_{p=R+1}^S\lambda_p\phi_p^2(x_j)\\ &\le \sum_{p=R+1}^S\lambda_pC_\phi^2\\ &= O(R^{1-\alpha}),\end{aligned} \quad (44)$$

it follows that

$$\|Z_jZ_j^T-\mathbb{E}(Z_jZ_j^T)\|_2 \le O(R^{1-\alpha}).$$

On the other hand,

$$\mathbb{E}[(Z_jZ_j^T-\mathbb{E}(Z_jZ_j^T))^2]=\mathbb{E}[\|Z_j\|_2^2Z_jZ_j^T]-(\mathbb{E}(Z_jZ_j^T))^2.$$

By (44), we have

$$\begin{aligned}\mathbb{E}[\|Z_j\|_2^2 Z_j Z_j^T] &\preccurlyeq \mathbb{E}[O(R^{1-\alpha}) Z_j Z_j^T] \\ &= O(R^{1-\alpha})\mathbb{E}[Z_j Z_j^T].\end{aligned}$$

It follows that

$$\begin{aligned}\|\mathbb{E}[(Z_j Z_j^T - \mathbb{E}(Z_j Z_j^T))^2]\|_2 &\leq \max\{O(R^{1-\alpha})\|\mathbb{E}[Z_j Z_j^T]\|_2, O(R^{-2\alpha})\} \\ &\leq \max\{O(R^{1-\alpha})O(R^{-\alpha}), O(R^{-2\alpha})\} \\ &= O(R^{1-2\alpha}).\end{aligned}$$

Using the matrix Bernstein theorem (Tropp, 2012, Theorem 6.1), we get

$$\begin{aligned}&\mathbb{P}(\|\sum_{j=1}^n [Z_j Z_j^T - \mathbb{E}(Z_j Z_j^T)]\|_2 > t) \\ \leq&(S-R)\exp\left(\frac{-t^2}{2(n\|\mathbb{E}[(Z_j Z_j^T - \mathbb{E}(Z_j Z_j^T))^2]\|_2 + \frac{t\max_j \|Z_j Z_j^T - \mathbb{E}(Z_j Z_j^T)\|_2}{3})}\right) \\ \leq&(S-R)\exp\left(\frac{-t^2}{2(nO(R^{1-2\alpha}) + \frac{tO(R^{1-\alpha})}{3})}\right) \\ =&(S-R)\exp\left(\frac{-t^2}{\max\{O(nR^{1-2\alpha}), tO(R^{1-\alpha})\})}\right).\end{aligned}$$

Then with probability at least $1-\delta$, we have

$$\begin{aligned}\|\sum_{j=1}^n [Z_j Z_j^T - \mathbb{E}(Z_j Z_j^T)]\|_2 \leq& \max\{\sqrt{\log\frac{S-R}{\delta} O(nR^{1-2\alpha})}, \log\frac{S-R}{\delta} O(R^{1-\alpha})\} \\ =&O(\log\frac{S-R}{\delta}\max\{n^{\frac{1}{2}} R^{\frac{1-2\alpha}{2}}, R^{1-\alpha}\}).\end{aligned}$$

This concludes the proof. □

Lemma 22. *Under the assumptions of Lemma 21, with probability at least $1-\delta$, we have*

$$\|\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T\|_2 = \tilde{O}(\max\{nR^{-\alpha}, n^{\frac{1}{2}} R^{\frac{1-2\alpha}{2}}, R^{1-\alpha}\}).$$

Proof of Lemma 22. For $S \in \mathbb{N}$, we have

$$\begin{aligned}\|\Phi_{>S}\Lambda_{>S}\Phi_{>S}^T\|_2 &\leq \sum_{p=S+1}^{\infty} \|\Lambda_p \phi_p(\mathbf{x})\phi_p(\mathbf{x})^T\|_2 \\ &= \sum_{p=S+1}^{\infty} \lambda_p \|\phi_p(\mathbf{x})\|_2^2 \\ &\leq \sum_{p=S+1}^{\infty} \lambda_p n C_\phi^2 \\ &= O(nS^{1-\alpha}).\end{aligned}$$

Let $S = R^{\frac{\alpha}{\alpha-1}}$. Then we get $\|\Phi_{>S}\Lambda_{>S}\Phi_{>S}^T\|_2 = O(nR^{-\alpha})$.

Let $\Phi_{R<p\leq S} = (\phi_{R+1}(\mathbf{x}), \ldots, \phi_S(\mathbf{x}))$, $\Lambda_{R<p\leq S} = (\lambda_{R+1}, \ldots, \lambda_S)$. We then have

$$\begin{aligned}\|\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T\|_2 \leq& \|\Phi_{>S}\Lambda_{>S}\Phi_{>S}^T\|_2 + \|\Phi_{R<p\leq S}\Lambda_{R<p\leq S}\Phi_{R<p\leq S}^T\|_2 \\ \leq& O(nR^{-\alpha}) + \|\Lambda_{R<p\leq S}^{1/2}\Phi_{R<p\leq S}^T\Phi_{R<p\leq S}\Lambda_{R<p\leq S}^{1/2}\|_2 \\ \leq& O(nR^{-\alpha}) + n\|\Lambda_{R<p\leq S}\|_2 + \|\Lambda_{R<p\leq S}^{1/2}(\Phi_{R<p\leq S}^T\Phi_{R<p\leq S} - nI)\Lambda_{R<p\leq S}^{1/2}\|_2 \\ \leq& O(nR^{-\alpha}) + O(nR^{-\alpha}) + O(\log\frac{R^{\frac{\alpha}{\alpha-1}} - R}{\delta}\max\{n^{\frac{1}{2}} R^{\frac{1-2\alpha}{2}}, R^{1-\alpha}\}) \\ =& \tilde{O}(\max\{nR^{-\alpha}, n^{\frac{1}{2}} R^{\frac{1-2\alpha}{2}}, R^{1-\alpha}\}),\end{aligned}$$

where in the fourth inequality we use Lemma 21. □

Corollary 23. *If $R = n^{\frac{2}{1+\alpha}-\kappa}$ where $\kappa < \frac{\alpha-1}{\alpha(1+\alpha)}$, we have that with probability at least $1-\delta$,*

$$\|(I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}\frac{\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T}{\sigma^2}\|_2 \leq \|\frac{\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T}{\sigma^2}\|_2 = \tilde{O}(n^{\frac{1-\alpha}{1+\alpha}+\kappa\alpha}) = o(1).$$

Proof of Corollary 23. By Lemma 22 and the assumption $R = n^{\frac{2}{1+\alpha}-\kappa}$, we have

$$\begin{aligned}\|(I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}\frac{\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T}{\sigma^2}\|_2 \leq& \|\frac{\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T}{\sigma^2}\|_2\\ \leq& \tilde{O}(\max\{nR^{-\alpha}, n^{\frac{1}{2}}R^{\frac{1-2\alpha}{2}}, R^{1-\alpha}\})\\ =& \tilde{O}(n^{\frac{1-\alpha}{1+\alpha}+\kappa\alpha}).\end{aligned}$$

□

Lemma 24. *Assume that $\|(I + \frac{n}{\sigma^2}\Lambda_R)^{-\gamma/2}\Lambda_R^{\gamma/2}(\Phi_R^T\Phi_R - nI)\Lambda_R^{\gamma/2}(I + \frac{n}{\sigma^2}\Lambda_R)^{-\gamma/2}\|_2 < 1$ where $\gamma \leq 1$. Then we have*

$$\begin{aligned}&(I + \frac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}\\ =&(I + \frac{n}{\sigma^2}\Lambda_R)^{-1} + \sum_{j=1}^{\infty}(-1)^j\left(\frac{1}{\sigma^2}(I + \frac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R(\Phi_R^T\Phi_R - nI)\right)^j(I + \frac{n}{\sigma^2}\Lambda_R)^{-1}.\end{aligned}$$

Proof of Lemma 24. First note that

$$\begin{aligned}&\|(I + \frac{n}{\sigma^2}\Lambda_R)^{-1/2}\Lambda_R^{1/2}(\Phi_R^T\Phi_R - nI)\Lambda_R^{1/2}(I + \frac{n}{\sigma^2}\Lambda_R)^{-1/2}\|_2\\ <& \|(I + \frac{n}{\sigma^2}\Lambda_R)^{-\gamma/2}\Lambda_R^{\gamma/2}(\Phi_R^T\Phi_R - nI)\Lambda_R^{\gamma/2}(I + \frac{n}{\sigma^2}\Lambda_R)^{-\gamma/2}\|_2\\ <& 1.\end{aligned}$$

Let $\tilde{\Lambda}_{\epsilon,R} = \mathrm{diag}\{\epsilon, \lambda_1, \ldots, \lambda_R\}$. Since $\Lambda_R = \mathrm{diag}\{0, \lambda_1, \ldots, \lambda_R\}$, we have that when ϵ is sufficiently small, $\|\frac{1}{\sigma^2}(I + \frac{n}{\sigma^2}\tilde{\Lambda}_{\epsilon,R})^{-1/2}\tilde{\Lambda}_{\epsilon,R}^{1/2}(\Phi_R^T\Phi_R - nI)\tilde{\Lambda}_{\epsilon,R}^{1/2}(I + \frac{n}{\sigma^2}\tilde{\Lambda}_{\epsilon,R})^{-1/2}\|_2 < 1$. Since all diagonal entries of $\tilde{\Lambda}_{\epsilon,R}$ are positive, we have

$$\begin{aligned}&(I + \frac{1}{\sigma^2}\tilde{\Lambda}_{\epsilon,R}\Phi_R^T\Phi_R)^{-1}\\ =&(I + \frac{n}{\sigma^2}\tilde{\Lambda}_{\epsilon,R} + \frac{1}{\sigma^2}\tilde{\Lambda}_{\epsilon,R}(\Phi_R^T\Phi_R - nI))^{-1}\\ =&\tilde{\Lambda}_{\epsilon,R}^{1/2}(I + \frac{n}{\sigma^2}\tilde{\Lambda}_{\epsilon,R})^{-1/2}\left[I + \frac{1}{\sigma^2}(I + \frac{n}{\sigma^2}\tilde{\Lambda}_{\epsilon,R})^{-1/2}\tilde{\Lambda}_{\epsilon,R}^{1/2}(\Phi_R^T\Phi_R - nI)\tilde{\Lambda}_{\epsilon,R}^{1/2}(I + \frac{n}{\sigma^2}\tilde{\Lambda}_{\epsilon,R})^{-1/2}\right]^{-1}\\ &\quad(I + \frac{n}{\sigma^2}\tilde{\Lambda}_{\epsilon,R})^{-1/2}\tilde{\Lambda}_{\epsilon,R}^{-1/2}\\ =&(I + \frac{n}{\sigma^2}\tilde{\Lambda}_{\epsilon,R})^{-1}\\ &+\sum_{j=1}^{\infty}\left((-1)^j\tilde{\Lambda}_{\epsilon,R}^{1/2}(I + \frac{n}{\sigma^2}\tilde{\Lambda}_{\epsilon,R})^{-1/2}\left(\frac{1}{\sigma^2}(I + \frac{n}{\sigma^2}\tilde{\Lambda}_{\epsilon,R})^{-1/2}\tilde{\Lambda}_{\epsilon,R}^{1/2}(\Phi_R^T\Phi_R - nI)\tilde{\Lambda}_{\epsilon,R}^{1/2}(I + \frac{n}{\sigma^2}\tilde{\Lambda}_{\epsilon,R})^{-1/2}\right)^j\right.\\ &\left.\quad(I + \frac{n}{\sigma^2}\tilde{\Lambda}_{\epsilon,R})^{-1/2}\tilde{\Lambda}_{\epsilon,R}^{-1/2}\right)\\ =&(I + \frac{n}{\sigma^2}\tilde{\Lambda}_{\epsilon,R})^{-1} + \sum_{j=1}^{\infty}(-1)^j\left(\frac{1}{\sigma^2}(I + \frac{n}{\sigma^2}\tilde{\Lambda}_{\epsilon,R})^{-1}\tilde{\Lambda}_{\epsilon,R}(\Phi_R^T\Phi_R - nI)\right)^j(I + \frac{n}{\sigma^2}\tilde{\Lambda}_{\epsilon,R})^{-1}.\end{aligned}$$

Letting $\epsilon \to 0$, we get

$$\begin{aligned}&(I + \frac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}\\ =&(I + \frac{n}{\sigma^2}\Lambda_R)^{-1} + \sum_{j=1}^{\infty}(-1)^j\left(\frac{1}{\sigma^2}(I + \frac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R(\Phi_R^T\Phi_R - nI)\right)^j(I + \frac{n}{\sigma^2}\Lambda_R)^{-1}.\end{aligned}$$

This concludes the proof. □

Lemma 25. *Let $R = n^{\frac{2}{1+\alpha}-\kappa}$ where $\kappa < \frac{\alpha-1}{\alpha(1+\alpha)}$. Then with probability at least $1-\delta$, for sufficiently large n, we have*

$$
\begin{aligned}
&(I + \frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1} - (I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1} \\
=&\sum_{j=1}^{\infty}(-1)^j\left((I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}\frac{\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T}{\sigma^2}\right)^j(I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}.
\end{aligned}
$$

Proof of Lemma 25. Define $\Phi_{>R} = (\phi_{R+1}(\mathbf{x}), \phi_{R+2}(\mathbf{x}), \ldots)$, $\Lambda_{>R} = (\lambda_{R+1}, \lambda_{R+2}, \ldots)$. Then we have

$$
\begin{aligned}
&(I + \frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1} - (I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1} \\
=&(I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2} + \frac{\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T}{\sigma^2})^{-1} - (I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1} \\
=&\left(\left(I + (I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}\frac{\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T}{\sigma^2}\right)^{-1} - I\right)(I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}.
\end{aligned}
$$

By Corollary 23, for sufficiently large n, $\|(I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}\frac{\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T}{\sigma^2}\|_2 < 1$ with probability at least $1-\delta$. Hence

$$
\begin{aligned}
&(I + \frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1} - (I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1} \\
=&\left(\left(I + (I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}\frac{\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T}{\sigma^2}\right)^{-1} - I\right)(I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1} \\
=&\sum_{j=1}^{\infty}(-1)^j\left((I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}\frac{\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T}{\sigma^2}\right)^j(I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}.
\end{aligned}
$$

□

Lemma 26. *Let $R = n^{\frac{2}{1+\alpha}-\kappa}$, where $0 \le \kappa < \frac{\alpha-1}{\alpha(1+\alpha)}$. Then when n is sufficiently large, with probability at least $1-2\delta$, we have*

$$
\|(I + \frac{1}{\sigma^2}\Phi_R\Lambda_R\Phi_R^T)^{-1}f_R(\mathbf{x})\|_2 = \tilde{O}\left(\sqrt{\frac{1}{\delta}n}\cdot n^{\max\{-1, \frac{1-2\beta}{2\alpha}\}}\right). \tag{45}
$$

Proof of Lemma 26. Using the Woodbury matrix identity, we have that

$$
\begin{aligned}
(I + \frac{1}{\sigma^2}\Phi_R\Lambda_R\Phi_R^T)^{-1}f_R(\mathbf{x}) =&\left[I - \Phi_R(\sigma^2 I + \Lambda_R\Phi_R^T\Phi_R)^{-1}\Lambda_R\Phi_R^T\right]\Phi_R\mu_R \\
=&\Phi_R\mu_R - \Phi_R(\sigma^2 I + \Lambda_R\Phi_R^T\Phi_R)^{-1}\Lambda_R\Phi_R^T\Phi_R\mu_R \\
=&\sigma^2\Phi_R(\sigma^2 I + \Lambda_R\Phi_R^T\Phi_R)^{-1}\mu_R.
\end{aligned} \tag{46}
$$

Let $A = (I + \frac{n}{\sigma^2}\Lambda_R)^{-\gamma/2}\Lambda_R^{\gamma/2}(\Phi_R^T\Phi_R - nI)\Lambda_R^{\gamma/2}(I + \frac{n}{\sigma^2}\Lambda_R)^{-\gamma/2}$, where $\gamma > \frac{1+\alpha}{2\alpha}$. By Lemma 20, with probability at least $1-\delta$, we have $\|A\|_2 = \tilde{O}(n^{\frac{1-2\gamma\alpha+\alpha}{2\alpha}})$. When n is sufficiently large, $\|A\|_2$ is less than 1. By Lemma 24, we have

$$
\begin{aligned}
&(I + \frac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1} \\
=&(I + \frac{n}{\sigma^2}\Lambda_R)^{-1} + \sum_{j=1}^{\infty}(-1)^j\left(\frac{1}{\sigma^2}(I + \frac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R(\Phi_R^T\Phi_R - nI)\right)^j(I + \frac{n}{\sigma^2}\Lambda_R)^{-1}.
\end{aligned}
$$

We then have

$$\begin{aligned}
&\|(\sigma^2 I+\Lambda_R\Phi_R^T\Phi_R)^{-1}\mu_R\|_2\\
=&\frac{1}{\sigma^2}\left\|\left((I+\frac{n}{\sigma^2}\Lambda_R)^{-1}+\sum_{j=1}^{\infty}(-1)^j\left(\frac{1}{\sigma^2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R(\Phi_R^T\Phi_R-nI)\right)^j(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\right)\mu_R\right\|_2\\
\leq&\frac{1}{\sigma^2}\left(\|(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\mu_R\|_2+\sum_{j=1}^{\infty}\left\|\left(\frac{1}{\sigma^2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R(\Phi_R^T\Phi_R-nI)\right)^j(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\mu_R\right\|_2\right).
\end{aligned}\tag{47}$$

By Lemma 15,

$$\|(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\mu_R\|_2\leq\sqrt{\sum_{p=1}^{R}\frac{C_\mu^2p^{-2\beta}}{(1+n\underline{C_\lambda}p^{-\alpha}/\sigma^2)^2}}=\tilde{O}(n^{\max\{-1,\frac{1-2\beta}{2\alpha}\}}).\tag{48}$$

Using the fact that $\|A\|_2=\tilde{O}(n^{\frac{1-2\gamma\alpha+\alpha}{2\alpha}})$ and $\|(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R\|_2\leq n^{-1}$, we have

$$\begin{aligned}
&\left\|\left(\frac{1}{\sigma^2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R(\Phi_R^T\Phi_R-nI)\right)^j(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\mu_R\right\|_2\\
=&\left\|(I+\frac{n}{\sigma^2}\Lambda_R)^{-1+\frac{\gamma}{2}}\Lambda_R^{1-\frac{\gamma}{2}}\left(A(I+\frac{n}{\sigma^2}\Lambda_R)^{-1+\gamma}\Lambda_R^{1-\gamma}\right)^{j-1}A(I+\frac{n}{\sigma^2}\Lambda_R)^{-1+\frac{\gamma}{2}}\Lambda_R^{-\frac{\gamma}{2}}\mu_R\right\|_2\\
\leq&n^{-1+\frac{\gamma}{2}+(-1+\gamma)(j-1)}\tilde{O}(n^{\frac{(1-2\gamma\alpha+\alpha)j}{2\alpha}})\|(I+\frac{n}{\sigma^2}\Lambda_R)^{-1+\frac{\gamma}{2}}\Lambda_R^{-\frac{\gamma}{2}}\mu_R\|_2\\
=&\tilde{O}(n^{-\frac{\gamma}{2}+\frac{(1-\alpha)j}{2\alpha}})\|(I+\frac{n}{\sigma^2}\Lambda_R)^{-1+\frac{\gamma}{2}}\Lambda_R^{-\frac{\gamma}{2}}\mu_R\|_2.
\end{aligned}\tag{49}$$

By Lemma 16 and the assumption $R=n^{\frac{2}{1+\alpha}-\kappa}$,

$$\begin{aligned}
\|(I+\frac{n}{\sigma^2}\Lambda_R)^{-1+\frac{\gamma}{2}}\Lambda_R^{-\frac{\gamma}{2}}\mu_R\|_2\leq&\sqrt{\sum_{p=1}^{R}\frac{(\underline{C_\lambda}p^{-\alpha})^{-\gamma}C_\mu^2p^{-2\beta}}{(1+n\underline{C_\lambda}p^{-\alpha}/\sigma^2)^{2-\gamma}}}\\
=&\tilde{O}(\max\{n^{-1+\gamma/2},R^{1/2-\beta+\alpha\gamma/2}\})
\end{aligned}\tag{50}$$

We then have

$$\begin{aligned}
&\left\|\left(\frac{1}{\sigma^2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R(\Phi_R^T\Phi_R-nI)\right)^j(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\mu_R\right\|_2\\
=&\tilde{O}(n^{\frac{(1-\alpha)(j-1)}{2\alpha}})\tilde{O}(\max\{n^{-1+\gamma/2-\frac{\gamma}{2}+\frac{(1-\alpha)}{2\alpha}},R^{1/2-\beta+\alpha\gamma/2}n^{-\frac{\gamma}{2}+\frac{(1-\alpha)}{2\alpha}}\})
\end{aligned}\tag{51}$$

According to (47), (48) and (51), we have

$$\begin{aligned}
&\|(\sigma^2 I+\Lambda_R\Phi_R^T\Phi_R)^{-1}\mu_R\|_2\\
\leq&\tilde{O}(n^{\max\{-1,\frac{1-2\beta}{2\alpha}\}})+\sum_{j=1}^{\infty}\sigma^{-2j}\tilde{O}(n^{\frac{(1-\alpha)(j-1)}{2\alpha}})\tilde{O}(\max\{n^{-1+\gamma/2-\frac{\gamma}{2}+\frac{(1-\alpha)}{2\alpha}},R^{1/2-\beta+\alpha\gamma/2}n^{-\frac{\gamma}{2}+\frac{(1-\alpha)}{2\alpha}}\}))\\
\leq&\tilde{O}(n^{\max\{-1,\frac{1-2\beta}{2\alpha}\}})+\frac{\tilde{O}(\max\{n^{-1+\frac{(1-\alpha)}{2\alpha}},R^{1/2-\beta+\alpha\gamma/2}n^{-\frac{\gamma}{2}+\frac{(1-\alpha)}{2\alpha}}\})}{1-\tilde{O}(n^{\frac{1-\alpha}{2\alpha}})\sigma^{-2}})\\
\leq&\tilde{O}(\max\{n^{\max\{-1,\frac{1-2\beta}{2\alpha},-1+\frac{(1-\alpha)}{2\alpha}\}},R^{1/2-\beta+\alpha\gamma/2}n^{-\frac{\gamma}{2}+\frac{(1-\alpha)}{2\alpha}}\})\\
\leq&\tilde{O}(n^{\max\{-1,\frac{1-2\beta}{2\alpha},-1+\frac{(1-\alpha)}{2\alpha},\frac{1-2\beta}{1+\alpha}+\frac{(\alpha-1)\gamma}{2(1+\alpha)}+\frac{(1-\alpha)}{2\alpha}-\kappa(1/2-\beta+\alpha\gamma/2)\}}).
\end{aligned}\tag{52}$$

Note that

$$\begin{aligned}
&\frac{1-2\beta}{2\alpha}-\left(\frac{1-2\beta}{1+\alpha}+\frac{(\alpha-1)\gamma}{2(1+\alpha)}+\frac{(1-\alpha)}{2\alpha}-\kappa(1/2-\beta+\alpha\gamma/2)\right)\\
=&\frac{1}{2}(\frac{\alpha-1}{\alpha(1+\alpha)}-\kappa)(2\beta-1-\alpha\gamma)-\frac{(1-\alpha)}{2\alpha}.
\end{aligned}\tag{53}$$

Hence as long as $\gamma < \alpha^{-1}(\frac{\alpha-1}{\alpha}(\frac{\alpha-1}{\alpha(1+\alpha)}-\kappa)^{-1}+2\beta-1)$, (53) is greater than zero. Since $\kappa \geq 0$, we can verify that

$$\alpha^{-1}(\frac{\alpha-1}{\alpha}(\frac{\alpha-1}{\alpha(1+\alpha)}-\kappa)^{-1}+2\beta-1) \geq \alpha^{-1}\frac{\alpha-1}{\alpha}(\frac{\alpha-1}{\alpha(1+\alpha)})^{-1} = \frac{1+\alpha}{\alpha}.$$

Then we can choose $\frac{1+\alpha}{2\alpha} < \gamma < \frac{1+\alpha}{\alpha}$ so that (53) is greater than zero. It is easy to verify that $-1 > -1 + \frac{(1-\alpha)}{2\alpha}$. Using (52), we then get

$$\|(\sigma^2 I + \Lambda_R \Phi_R^T \Phi_R)^{-1}\mu_R\|_2 = \tilde{O}(n^{\max\{-1, \frac{1-2\beta}{2\alpha}\}}). \tag{54}$$

By Corollary 19, with probability at least $1-\delta$, we have

$$\begin{aligned}\|\Phi_R(\sigma^2 I + \Lambda_R \Phi_R^T \Phi_R)^{-1}\mu_R\|_2 =& \tilde{O}(\sqrt{\frac{1}{\delta}n}\|(\sigma^2 I + \Lambda_R \Phi_R^T \Phi_R)^{-1}\mu_R\|_2) \\ =& \tilde{O}(\sqrt{\frac{1}{\delta}n}\cdot n^{\max\{-1, \frac{1-2\beta}{2\alpha}\}}).\end{aligned}$$

From (46), we get $\|(I + \frac{1}{\sigma^2}\Phi_R\Lambda_R\Phi_R^T)^{-1}f_R(\mathbf{x})\|_2 = \tilde{O}(\sqrt{\frac{1}{\delta}n}\cdot n^{\max\{-1, \frac{1-2\beta}{2\alpha}\}})$. This concludes the proof. □

Lemma 27. *Let $R = n^{\frac{2}{1+\alpha}-\kappa}$ where $0 \leq \kappa < \frac{\alpha-1}{\alpha(1+\alpha)}$. Then when n is sufficiently large, with probability at least $1-3\delta$ we have*

$$\|(I + \frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f_R(\mathbf{x})\|_2 = \tilde{O}(\sqrt{\frac{1}{\delta}n}\cdot n^{\max\{-1, \frac{1-2\beta}{2\alpha}\}}). \tag{55}$$

Proof of Lemma 27. We have

$$\begin{aligned}&(I + \frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f_R(\mathbf{x}) \\ =&(I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}f_R(\mathbf{x}) + \left((I + \frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1} - (I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}\right)f_R(\mathbf{x}).\end{aligned} \tag{56}$$

By Lemma 26, with probability at least $1-2\delta$ we have,

$$\|(I + \frac{1}{\sigma^2}\Phi_R\Lambda_R\Phi_R^T)^{-1}f_R(\mathbf{x})\|_2 = \tilde{O}(\sqrt{\frac{1}{\delta}n}\cdot n^{\max\{-1, \frac{1-2\beta}{2\alpha}\}}).$$

Using Lemma 25 and Corollary 23, with probability at least $1-\delta$, the second term in the right hand side of (56) is estimated as follows:

$$\begin{aligned}&\|\left((I + \frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1} - (I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}\right)f_R(\mathbf{x})\|_2 \\ =&\|\sum_{j=1}^{\infty}(-1)^j\left((I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}\frac{\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T}{\sigma^2}\right)^j(I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}f_R(\mathbf{x})\|_2 \\ =&\sum_{j=1}^{\infty}\left\|\left((I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}\frac{\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T}{\sigma^2}\right)\right\|_2^j\|(I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}f_R(\mathbf{x})\|_2 \\ =&\sum_{j=1}^{\infty}\tilde{O}(n^{j(\frac{1-\alpha}{1+\alpha}+\kappa\alpha)})\tilde{O}(\sqrt{\frac{1}{\delta}n}\cdot n^{\max\{-1, \frac{1-2\beta}{2\alpha}\}}) \\ =&o(\sqrt{\frac{1}{\delta}n}\cdot n^{\max\{-1, \frac{1-2\beta}{2\alpha}\}}).\end{aligned}$$

Overall, from (56), we have that with probability $1-3\delta$,

$$\|(I + \frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f_R(\mathbf{x})\|_2 = \tilde{O}(\sqrt{\frac{1}{\delta}n}\cdot n^{\max\{-1, \frac{1-2\beta}{2\alpha}\}}).$$

This concludes the proof. □

C PROOF OF THE MAIN RESULTS

C.1 PROOFS RELATED TO THE ASYMPTOTICS OF THE NORMALIZED STOCHASTIC COMPLEXITY

Lemma 28. *Let $R = n^{\frac{2}{1+\alpha}-\kappa}$. Under Assumptions 4, 5 and 6, we have*

$$|T_{1,R}(D_n) - T_1(D_n)| = O(nR^{1-\alpha}) = O\Big(n^{\frac{3-\alpha}{1+\alpha}+\kappa(\alpha-1)}\Big). \tag{57}$$

If we further assume that $0 \leq \kappa < \frac{\alpha-1}{\alpha(1+\alpha)}$, then for sufficiently large n with probability at least $1-4\delta$ we have

$$|T_{2,R}(D_n) - T_2(D_n)| = \tilde{O}\bigg(\frac{1}{\delta} n^{\max\{(\frac{2}{1+\alpha}-\kappa)\frac{1-2\beta}{2}, 1+(\frac{3\alpha+1}{\alpha(1+\alpha)}-\kappa)\frac{1-2\beta}{2}, -1+(\frac{1-\alpha}{1+\alpha}+\kappa\alpha), 1+\frac{1-2\beta}{\alpha}+\frac{1-\alpha}{1+\alpha}+\kappa\alpha\}}\bigg). \tag{58}$$

Proof of Lemma 28. Define $\Phi_{>R} = (\phi_{R+1}(\mathbf{x}), \phi_{R+2}(\mathbf{x}), \ldots, \phi_p(\mathbf{x}), \ldots)$, and $\Lambda_{>R} = (\lambda_{R+1}, \ldots, \lambda_p, \ldots)$. We then have

$$\begin{aligned} |T_1(D_n) - T_{1,R}(D_n)| = &\left|\frac{1}{2}\log\det(I + \frac{1}{\sigma^2}\Phi\Lambda\Phi^T) - \frac{1}{2}\log\det(I + \frac{1}{\sigma^2}\Phi_R\Lambda_R\Phi_R^T)\right| \\ &+ \frac{1}{2}\left|\mathrm{Tr}(I + \frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1} - \mathrm{Tr}(I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}\right|. \end{aligned} \tag{59}$$

As for the first term in the right hand side of (59), we have

$$\begin{aligned} &\left|\frac{1}{2}\log\det(I + \frac{1}{\sigma^2}\Phi\Lambda\Phi^T) - \frac{1}{2}\log\det(I + \frac{1}{\sigma^2}\Phi_R\Lambda_R\Phi_R^T)\right| \\ =&\left|\frac{1}{2}\log\det\bigg((I + \frac{1}{\sigma^2}\Phi_R\Lambda_R\Phi_R^T)^{-1}(I + \frac{1}{\sigma^2}\Phi_R\Lambda_R\Phi_R^T + \frac{1}{\sigma^2}\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T)\bigg)\right| \\ =&\left|\frac{1}{2}\log\det\bigg(I + \frac{1}{\sigma^2}(I + \frac{1}{\sigma^2}\Phi_R\Lambda_R\Phi_R^T)^{-1}\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T\bigg)\right| \\ =&\frac{1}{2}\left|\mathrm{Tr}\log\bigg(I + \frac{1}{\sigma^2}(I + \frac{1}{\sigma^2}\Phi_R\Lambda_R\Phi_R^T)^{-1}\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T\bigg)\right|. \end{aligned} \tag{60}$$

Given a concave function h and a matrix $B \in \mathbb{R}^{n\times n}$ whose eigenvalues $\zeta_1, \ldots, \zeta_n$ are all positive, we have that

$$\mathrm{Tr}\, h(B) = \sum_{p=1}^{n} h(\zeta_i) \leq nh(\frac{1}{n}\sum_{p=1}^{n}\zeta_i) \leq nh(\frac{1}{n}\,\mathrm{Tr}\, B), \tag{61}$$

where we used Jensen's inequality. Using $h(x) = \log(1+x)$ in (61), we have

$$\begin{aligned} &\left|\frac{1}{2}\log\det(I + \frac{1}{\sigma^2}\Phi\Lambda\Phi^T) - \frac{1}{2}\log\det(I + \frac{1}{\sigma^2}\Phi_R\Lambda_R\Phi_R^T)\right| \\ \leq&\frac{n}{2}\log(1 + \frac{1}{n}\,\mathrm{Tr}(\frac{1}{\sigma^2}(I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T)) \\ \leq&\frac{n}{2}\log(1 + \frac{1}{n\sigma^2}\|(I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}\|_2\,\mathrm{Tr}(\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T)) \\ \leq&\frac{n}{2}\log(1 + \frac{1}{n\sigma^2}\sum_{p=R+1}^{\infty}\lambda_p\|\phi_p(\mathbf{x})\|_2^2) \\ =&\frac{n}{2}\log(1 + \frac{C_\phi^2}{\sigma^2}\sum_{p=R+1}^{\infty}\lambda_p) \\ =&O(nR^{1-\alpha}) \\ =&\tilde{O}(n^{\frac{3-\alpha}{1+\alpha}+\kappa(\alpha-1)}), \end{aligned}$$

where in the second inequality we use the fact that $\operatorname{Tr} AB \leq \|A\|_2 \operatorname{Tr} B$ when A and B are symmetric positive definite matrices.

As for the second term in the right hand side of (59), let $A = (I + \frac{\Phi_R \Lambda_R \Phi_R^T}{\sigma^2})^{-1/2}$. Then we have

$$
\begin{aligned}
&\frac{1}{2}\left|\operatorname{Tr}(I + \frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1} - \operatorname{Tr}(I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}\right| \\
=&\frac{1}{2}\left|\operatorname{Tr} A\left[I - (I + A(\frac{\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T}{\sigma^2})A)^{-1}\right]A\right| \\
\leq&\frac{1}{2}\operatorname{Tr}\left[I - (I + A(\frac{\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T}{\sigma^2})A)^{-1}\right] \\
\leq&\frac{n}{2}(1 - (1 + \frac{1}{n}\operatorname{Tr} A(\frac{\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T}{\sigma^2})A)^{-1}) \\
\leq&\frac{n}{2}(1 - (1 + \frac{1}{n}\operatorname{Tr}(\frac{\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T}{\sigma^2}))^{-1}) \\
\leq&\frac{n}{2}(1 - (1 + \frac{1}{n\sigma^2}\sum_{p=R+1}^{\infty}\lambda_p\|\phi_p(\mathbf{x})\|_2^2))^{-1}) \\
=&O(nR^{1-\alpha}) \\
=&\tilde{O}(n^{\frac{3-\alpha}{1+\alpha}+\kappa(\alpha-1)}),
\end{aligned}
$$

where in the first inequality we use the fact that $\|A\|_2 < 1$ and $\operatorname{Tr} ABA \leq \|A\|_2^2 \operatorname{Tr} B$ when A and B are symmetric positive definite matrices, and in the second inequality we use $h(x) = 1 - 1/(1+x)$ in (61). This concludes the proof of the first statement.

As for $|T_2(D_n) - T_{2,R}(D_n)|$, we have

$$
\begin{aligned}
|T_2(D_n) - T_{2,R}(D_n)| = &\left|f(\mathbf{x})^T(I + \frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f(\mathbf{x}) - f_R(\mathbf{x})^T(I + \frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f_R(\mathbf{x})\right| \\
&+ \left|f_R(\mathbf{x})^T(I + \frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f_R(\mathbf{x}) - f_R(\mathbf{x})^T(I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}f_R(\mathbf{x})\right|.
\end{aligned} \tag{62}
$$

For the first term on the right-hand side of (62), we have

$$
\begin{aligned}
&\left|f(\mathbf{x})^T(I + \frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f(\mathbf{x}) - f_R(\mathbf{x})^T(I + \frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f_R(\mathbf{x})\right| \\
\leq&2\left|f_{>R}(\mathbf{x})^T(I + \frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f_R(\mathbf{x})\right| + \left|f_{>R}(\mathbf{x})^T(I + \frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f_{>R}(\mathbf{x})\right| \\
\leq&2\|f_{>R}(\mathbf{x})\|_2\|(I + \frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f_R(\mathbf{x})\|_2 + \|f_{>R}(\mathbf{x})\|_2\|(I + \frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}\|_2\|f_{>R}(\mathbf{x})\|_2 \\
\leq&2\|f_{>R}(\mathbf{x})\|_2\|(I + \frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f_R(\mathbf{x})\|_2 + \|f_{>R}(\mathbf{x})\|_2^2.
\end{aligned}
$$

Applying Corollary 18 and Lemma 27, with probability at least $1 - 4\delta$, we have

$$
\begin{aligned}
&\left|f(\mathbf{x})^T(I + \frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f(\mathbf{x}) - f_R(\mathbf{x})^T(I + \frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f_R(\mathbf{x})\right| \\
\leq&2\tilde{O}\left(\sqrt{\frac{1}{\delta}nR^{1-2\beta}}\right)\tilde{O}(\sqrt{\frac{1}{\delta}n}\cdot n^{\max\{-1,\frac{1-2\beta}{2\alpha}\}}) + \tilde{O}(\frac{1}{\delta}nR^{1-2\beta}) \\
=&2\tilde{O}\left(\frac{1}{\delta}n^{1+(\frac{2}{1+\alpha}-\kappa)\frac{1-2\beta}{2}+\max\{-1,\frac{1-2\beta}{2\alpha}\}}\right) + \tilde{O}(\frac{1}{\delta}n^{1+(\frac{2}{1+\alpha}-\kappa)(1-2\beta)}) \\
=&2\tilde{O}\left(\frac{1}{\delta}n^{1+(\frac{2}{1+\alpha}-\kappa)\frac{1-2\beta}{2}+\max\{-1,\frac{1-2\beta}{2\alpha}\}}\right),
\end{aligned}
$$

where the last equality holds because $(\frac{2}{1+\alpha}-\kappa)\frac{1-2\beta}{2} < \frac{1-2\beta}{2\alpha}$ when $0 \leq \kappa < \frac{\alpha-1}{\alpha(1+\alpha)}$.

As for the second term on the right-hand side of (62), according to Lemma 25, Corollary 23 and Lemma 26, we have

$$\begin{aligned}
&\left|f_R(\mathbf{x})^T(I+\frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f_R(\mathbf{x}) - f_R(\mathbf{x})^T(I+\frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}f_R(\mathbf{x})\right| \\
=&\left|\sum_{j=1}^{\infty}(-1)^j f_R(\mathbf{x})^T\left((I+\frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}\frac{\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T}{\sigma^2}\right)^j(I+\frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}f_R(\mathbf{x})\right| \\
\leq&\sum_{j=1}^{\infty}\|(I+\frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}\|_2^{j-1}\cdot\|\frac{\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T}{\sigma^2}\|_2^j\cdot\|(I+\frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}f_R(\mathbf{x})\|_2^2 \qquad (63)\\
=&\sum_{j=1}^{\infty}\tilde{O}(n^{j(\frac{1-\alpha}{1+\alpha}+\kappa\alpha)})\tilde{O}(\frac{1}{\delta}n^{1+\max\{-2,\frac{1-2\beta}{\alpha}\}}) \\
=&\tilde{O}(\frac{1}{\delta}n^{1+\max\{-2,\frac{1-2\beta}{\alpha}\}+(\frac{1-\alpha}{1+\alpha}+\kappa\alpha)}).
\end{aligned}$$

By (62), we have

$$\begin{aligned}
|T_2(D_n)-T_{2,R}(D_n)| &= \tilde{O}\left(\frac{1}{\delta}n^{1+(\frac{2}{1+\alpha}-\kappa)\frac{1-2\beta}{2}+\max\{-1,\frac{1-2\beta}{2\alpha}\}}\right)+\tilde{O}(\frac{1}{\delta}n^{1+\max\{-2,\frac{1-2\beta}{\alpha}\}+(\frac{1-\alpha}{1+\alpha}+\kappa\alpha)}) \\
&= \tilde{O}\left(\frac{1}{\delta}n^{\max\{(\frac{2}{1+\alpha}-\kappa)\frac{1-2\beta}{2},1+(\frac{3\alpha+1}{\alpha(1+\alpha)}-\kappa)\frac{1-2\beta}{2},-1+(\frac{1-\alpha}{1+\alpha}+\kappa\alpha),1+\frac{1-2\beta}{\alpha}+\frac{1-\alpha}{1+\alpha}+\kappa\alpha\}}\right).
\end{aligned}$$

This concludes the proof of the second statement. □

In Lemma 28, we gave a bound for $|T_{2,R}(D_n)-T_2(D_n)|$ when $n^{\frac{2}{1+\alpha}-\frac{\alpha-1}{\alpha(1+\alpha)}} < R \leq n^{\frac{2}{1+\alpha}}$. For $R > n$, we note the following lemma:

Lemma 29. *Let $R > n$. Under Assumptions 4, 5 and 6, for sufficiently large n and with probability at least $1-3\delta$ we have*

$$|T_{2,R}(D_n)-T_2(D_n)| = \tilde{O}\left(\frac{1}{\delta}nR^{\max\{1/2-\beta,1-\alpha\}}\right). \qquad (64)$$

Proof of Lemma 29. Define $\Phi_{>R} = (\phi_{R+1}(\mathbf{x}),\phi_{R+2}(\mathbf{x}),\ldots,\phi_p(\mathbf{x}),\ldots)$, and $\Lambda_{>R} = (\lambda_{R+1},\ldots,\lambda_p,\ldots)$. Then we have

$$\begin{aligned}
|T_2(D_n)-T_{2,R}(D_n)| =& \left|f(\mathbf{x})^T(I+\frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f(\mathbf{x}) - f_R(\mathbf{x})^T(I+\frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f_R(\mathbf{x})\right| \\
&+\left|f_R(\mathbf{x})^T(I+\frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f_R(\mathbf{x}) - f_R(\mathbf{x})^T(I+\frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}f_R(\mathbf{x})\right|. \qquad (65)
\end{aligned}$$

For the first term on the right-hand side of (65), with probability $1-3\delta$ we have

$$\begin{aligned}
&\left|f(\mathbf{x})^T(I+\frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f(\mathbf{x}) - f_R(\mathbf{x})^T(I+\frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f_R(\mathbf{x})\right| \\
\leq&2\left|f_{>R}(\mathbf{x})^T(I+\frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f_R(\mathbf{x})\right| + \left|f_{>R}(\mathbf{x})^T(I+\frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f_{>R}(\mathbf{x})\right| \\
\leq&2\|f_{>R}(\mathbf{x})\|_2\|(I+\frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}\|_2\|f_R(\mathbf{x})\|_2 + \|f_{>R}(\mathbf{x})\|_2\|(I+\frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}\|_2\|f_{>R}(\mathbf{x})\|_2 \\
\leq&2\|f_{>R}(\mathbf{x})\|_2\|f_R(\mathbf{x})\|_2 + \|f_{>R}(\mathbf{x})\|_2^2 \\
\leq&2\tilde{O}\left(\sqrt{\frac{1}{\delta}nR^{1-2\beta}}\right)\tilde{O}(\sqrt{\frac{1}{\delta}n}\cdot\|f\|_2) + \tilde{O}(\frac{1}{\delta}nR^{1-2\beta}) \\
=&\tilde{O}\left(\frac{1}{\delta}nR^{1/2-\beta}\right),
\end{aligned}$$

where we used Corollary 18 and Lemma 17 for the last inequality.

As for the second term on the right-hand side of (65), by Lemmas 25 and 22, we have

$$\begin{aligned}
&\left| f_R(\mathbf{x})^T (I + \frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1} f_R(\mathbf{x}) - f_R(\mathbf{x})^T (I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1} f_R(\mathbf{x}) \right| \\
=&\left| \sum_{j=1}^{\infty} (-1)^j f_R(\mathbf{x})^T \left((I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1} \frac{\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T}{\sigma^2} \right)^j (I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1} f_R(\mathbf{x}) \right| \\
\leq& \sum_{j=1}^{\infty} \|(I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}\|_2^{j+1} \cdot \| \frac{\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T}{\sigma^2} \|_2^j \cdot \|f_R(\mathbf{x})\|_2^2 \\
=& \sum_{j=1}^{\infty} \tilde{O}(R^{j(1-\alpha)}) \tilde{O}(\frac{1}{\delta} n \|f\|_2^2) \\
=& \tilde{O}(\frac{1}{\delta} n R^{1-\alpha}).
\end{aligned} \tag{66}$$

Using (65), we have

$$|T_2(D_n) - T_{2,R}(D_n)| = \tilde{O}\left(\frac{1}{\delta} n R^{1/2-\beta}\right) + \tilde{O}(\frac{1}{\delta} n R^{1-\alpha}) = \tilde{O}\left(\frac{1}{\delta} n R^{\max\{1/2-\beta, 1-\alpha\}}\right).$$

□

In the next lemma we give the asympototics of $T_{1,R}(D_n)$ and $T_{2,R}(D_n)$.

Lemma 30. *Let $R = n^{\frac{2}{1+\alpha}-\kappa}$ where $0 \leq \kappa < \frac{\alpha-1}{\alpha(1+\alpha)}$. Under Assumptions 4, 5 and 6, with probability at least $1-\delta$, we have*

$$T_{1,R}(D_n) = O(n^{\frac{1}{\alpha}}), \quad T_{2,R}(D_n) = \tilde{O}(n^{\max\{0, 1+\frac{1-2\beta}{\alpha}\}}). \tag{67}$$

Proof of Lemma 30. Let

$$A = (I + \frac{n}{\sigma^2}\Lambda_R)^{-\gamma/2} \Lambda_R^{\gamma/2} (\Phi_R^T\Phi_R - nI) \Lambda_R^{\gamma/2} (I + \frac{n}{\sigma^2}\Lambda_R)^{-\gamma/2}, \tag{68}$$

where $\gamma = \frac{1+\alpha}{2\alpha} + \kappa_1$ and $\kappa_1 > 0$. By Lemma 20, with probability at least $1-\delta$, we have

$$\|A\|_2 = \tilde{O}(n^{\frac{1-2\gamma\alpha+\alpha}{2\alpha}}). \tag{69}$$

When n is sufficiently large, $\|A\|_2$ is less than 1. Let $B = (I + \frac{n}{\sigma^2}\Lambda_R)^{-1/2} \Lambda_R^{1/2} (\Phi_R^T\Phi_R - nI) \Lambda_R^{1/2} (I + \frac{n}{\sigma^2}\Lambda_R)^{-1/2}$. Then $\|B\|_2 = \frac{\sigma^{2(1-\gamma)}}{n^{1-\gamma}} \|A\|_2 = \tilde{O}(n^{\frac{1-\alpha}{2\alpha}})$. Using the Woodbury ma-

trix identity, we compute $T_{1,R}(D_n)$ as follows:

$$
\begin{aligned}
T_{1,R}(D_n) &= \frac{1}{2}\log\det(I+\frac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R) - \frac{1}{2}\mathrm{Tr}\Phi_R(\sigma^2 I+\Lambda_R\Phi_R^T\Phi_R)^{-1}\Lambda_R\Phi_R^T \\
&= \frac{1}{2}\log\det(I+\frac{n}{\sigma^2}\Lambda_R) + \frac{1}{2}\log\det[I+\frac{1}{\sigma^2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}\Lambda_R^{1/2}(\Phi_R^T\Phi_R-nI)\Lambda_R^{1/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}] \\
&- \frac{1}{2}\mathrm{Tr}(\sigma^2 I+\Lambda\Phi_R^T\Phi_R)^{-1}\Lambda\Phi_R^T\Phi_R \\
&= \frac{1}{2}\log\det(I+\frac{n}{\sigma^2}\Lambda_R) + \frac{1}{2}\,\mathrm{Tr}\log[I+\frac{1}{\sigma^2}B] - \frac{1}{2}\mathrm{Tr}(I-\sigma^2(\sigma^2 I+\Lambda\Phi_R^T\Phi_R)^{-1})) \\
&= \frac{1}{2}\log\det(I+\frac{n}{\sigma^2}\Lambda_R) + \frac{1}{2}\,\mathrm{Tr}\sum_{j=1}^{\infty}\frac{(-1)^{j-1}}{j}(\frac{1}{\sigma^2}B)^j \\
&- \frac{1}{2}\mathrm{Tr}\left(I-(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}+\sum_{j=1}^{\infty}(-1)^j\left(\frac{1}{\sigma^2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R(\Phi_R^T\Phi_R-nI)\right)^j(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\right) \\
&\leq \frac{1}{2}\log\det(I+\frac{n}{\sigma^2}\Lambda_R) - \frac{1}{2}\mathrm{Tr}\Big(I-(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\Big) + \frac{1}{2}R\sum_{j=1}^{\infty}\|\frac{1}{\sigma^2}B\|_2^j \\
&- \frac{1}{2}\mathrm{Tr}\left(\sum_{j=1}^{\infty}(-1)^j\frac{1}{\sigma^{2j}}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}B^j(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}\right),
\end{aligned}
\tag{70}
$$

where in the last equality we apply Lemma 24.

Let $h(x)=\log(1+x)-(1-\frac{1}{1+x})$. It is easy to verify that $h(x)$ is increasing on $[0,+\infty)$. As for the first term on the right hand side of (70), we have

$$
\begin{aligned}
&\frac{1}{2}\log\det(I+\frac{n}{\sigma^2}\Lambda_R) - \frac{1}{2}\,\mathrm{Tr}\Big(I-(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\Big) \\
=&\frac{1}{2}\sum_{p=1}^{R}\left(\log(1+\frac{n}{\sigma^2}\lambda_p)-(1-\frac{1}{1+\frac{n}{\sigma^2}\lambda_p})\right) \\
=&\frac{1}{2}\sum_{p=1}^{R}h(\frac{n}{\sigma^2}\lambda_p) \\
\leq&\frac{1}{2}\sum_{p=1}^{R}h(\frac{n}{\sigma^2}\overline{C_\lambda}p^{-\alpha}) \\
\leq&\frac{1}{2}h(\frac{n}{\sigma^2}\overline{C_\lambda}) + \frac{1}{2}\int_{[1,R]}h(\frac{n}{\sigma^2}\overline{C_\lambda}x^{-\alpha})\mathrm{d}x \\
=&\frac{1}{2}h(\frac{n}{\sigma^2}\overline{C_\lambda}) + \frac{1}{2}n^{1/\alpha}\int_{[1/n^{1/\alpha},R/n^{1/\alpha}]}h(\frac{\overline{C_\lambda}}{\sigma^2}x^{-\alpha})\mathrm{d}x \\
=&O(n^{1/\alpha}),
\end{aligned}
$$

where in the last equality we use the fact that $\int_{[0,+\infty]}h(x^{-\alpha})\mathrm{d}x<\infty$.

As for the third term on the right hand side of (70), we have

$$
\begin{aligned}
&\left|\mathrm{Tr}\left(\sum_{j=1}^{\infty}(-1)^j\frac{1}{\sigma^{2j}}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}B^j(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}\right)\right| \\
\leq&\sum_{j=1}^{\infty}\left|\mathrm{Tr}\left(\frac{1}{\sigma^{2j}}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}B^j(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}\right)\right| \\
\leq&R\sum_{j=1}^{\infty}\left\|\frac{1}{\sigma^{2j}}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}B^j(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}\right\|_2 \\
\leq&R\sum_{j=1}^{\infty}\left\|\frac{1}{\sigma^{2j}}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}B^j(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}\right\|_2 \\
\leq&R\sum_{j=1}^{\infty}\left\|\frac{1}{\sigma^{2j}}B^j\right\|_2.
\end{aligned}
$$

Then the asymptotics of $T_{1,R}(D_n)$ is given by

$$
\begin{aligned}
T_{1,R}(D_n) \leq& O(n^{1/\alpha})+R\sum_{j=1}^{\infty}\left\|\frac{1}{\sigma^{2j}}B^j\right\|_2 \\
=&O(n^{1/\alpha})+R\sum_{j=1}^{\infty}\frac{1}{\sigma^{2j}}\tilde{O}(n^{\frac{j(1-\alpha)}{2\alpha}}) \\
=&O(n^{1/\alpha})+R\tilde{O}(n^{\frac{1-\alpha}{2\alpha}}) \\
=&O(n^{1/\alpha})+\tilde{O}(n^{\frac{2}{1+\alpha}-\kappa+\frac{1-\alpha}{2\alpha}}) \\
=&O(n^{\frac{1}{\alpha}})+\tilde{O}(n^{\frac{1}{\alpha}-\frac{(\alpha-1)^2}{2\alpha(1+\alpha)}-\kappa}) \\
=&O(n^{\frac{1}{\alpha}}).
\end{aligned}
$$

This concludes the proof of the first statement.

Using the Woodbury matrix identity, we compute $T_{2,R}(D_n)$ as follows:

$$
\begin{aligned}
T_{2,R}(D_n) &= \frac{1}{2\sigma^2}\mu_R^T\Phi_R^T(I+\frac{1}{\sigma^2}\Phi_R\Lambda_R\Phi_R^T)^{-1}\Phi_R\mu_R \\
&= \frac{1}{2\sigma^2}\mu_R^T\Phi_R^T\big[I-\Phi_R(\sigma^2I+\Lambda_R\Phi_R^T\Phi_R)^{-1}\Lambda_R\Phi_R^T\big]\Phi_R\mu_R \\
&= \frac{1}{2\sigma^2}\mu_R^T\Phi_R^T\Phi_R\mu_R-\frac{1}{2\sigma^2}\mu_R^T\Phi_R^T\Phi_R(\sigma^2I+\Lambda_R\Phi_R^T\Phi_R)^{-1}\Lambda_R\Phi_R^T\Phi_R\mu_R \\
&= \frac{1}{2}\mu_R^T\Phi_R^T\Phi_R(\sigma^2I+\Lambda_R\Phi_R^T\Phi_R)^{-1}\mu_R \\
&= \frac{1}{2}\mu_R^T\Lambda_R^{-1}\Lambda_R\Phi_R^T\Phi_R(\sigma^2I+\Lambda_R\Phi_R^T\Phi_R)^{-1}\mu_R \\
&= \frac{1}{2}\mu_R^T\Lambda_R^{-1}\mu_R-\frac{1}{2}\mu_R^T\Lambda_R^{-1}(I+\frac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}\mu_R.
\end{aligned}
\tag{71}
$$

Using Lemma 24, we have

$$
\begin{aligned}
T_{2,R}(D_n) =& \frac{1}{2}\mu_R^T\Lambda_R^{-1}\mu_R - \frac{1}{2}\mu_R^T\Lambda_R^{-1}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\mu_R \\
&+ \frac{1}{2}\sum_{j=1}^{\infty}(-1)^{j+1}\mu_R^T\Lambda_R^{-1}\Big(\frac{1}{\sigma^2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R(\Phi_R^T\Phi_R - nI)\Big)^j(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\mu_R \\
&= \frac{n}{2}\mu_R^T(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\mu_R \\
&+ \frac{1}{2}\sum_{j=1}^{\infty}(-1)^{j+1}\frac{1}{\sigma^{2j}}\mu_R^T(I+\frac{n}{\sigma^2}\Lambda_R)^{-1+\gamma/2}\Lambda_R^{-\gamma/2}A\Big((I+\frac{n}{\sigma^2}\Lambda_R)^{-1+\gamma}\Lambda_R^{1-\gamma}A\Big)^{j-1} \\
&\qquad (I+\frac{n}{\sigma^2}\Lambda_R)^{-1+\gamma/2}\Lambda_R^{-\gamma/2}\mu \\
&\le \frac{n}{2}\mu_R^T(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\mu_R + \frac{1}{2}\sum_{j=1}^{\infty}\frac{1}{\sigma^{2j}}\|A\|^j\Big(\frac{n}{\sigma^2}\Big)^{(-1+\gamma)(j-1)}\|(I+\frac{n}{\sigma^2}\Lambda_R)^{-1+\gamma/2}\Lambda_R^{-\gamma/2}\mu\|_2^2,
\end{aligned} \tag{72}
$$

where in the second to last equality we used the definition of A (68). As for the first term on the right hand side of (72), by Lemma 15, we have

$$
\frac{n}{2}\mu^T(I+\frac{n}{\sigma^2}\Lambda)^{-1}\mu \le \frac{n}{2}\sum_{p=1}^{R}\frac{C_\beta^2 p^{-2\beta}}{1+\frac{n}{\sigma^2}\underline{C_\lambda}p^{-\alpha}} = \tilde{O}(n^{\max\{0,1+\frac{1-2\beta}{\alpha}\}}).
$$

By Lemma 16, we have

$$
\begin{aligned}
\|(I+\frac{n}{\sigma^2}\Lambda)^{-1+\gamma/2}\Lambda^{-\gamma/2}\mu\|_2^2 \le& \sum_{p=1}^{R}\frac{C_\beta^2 p^{-2\beta}(\underline{C_\lambda}p^{-\alpha})^{-\gamma}}{(1+\frac{n}{\sigma^2}\underline{C_\lambda}p^{-\alpha})^{2-\gamma}} \\
=&\tilde{O}(\max\{n^{-2+\gamma}, R^{1-2\beta+\alpha\gamma}\}) \\
=&\tilde{O}(n^{\max\{-2+\gamma,\frac{2(1-2\beta+\alpha\gamma)}{1+\alpha}-\kappa(1-2\beta+\alpha\gamma)\}}).
\end{aligned}
$$

Using (69), the second term on the right hand side of (72) is computed as follows:

$$
\begin{aligned}
&\frac{1}{2}\sum_{j=1}^{\infty}\frac{1}{\sigma^{2j}}\|A\|^j\Big(\frac{n}{\sigma^2}\Big)^{(-1+\gamma)(j-1)}\|(I+\frac{n}{\sigma^2}\Lambda)^{-1+\gamma/2}\Lambda^{-\gamma/2}\mu\|_2^2 \\
\le&\frac{1}{2}\sum_{j=1}^{\infty}\frac{1}{\sigma^{2j}}\tilde{O}(n^{\frac{j(1-2\gamma\alpha+\alpha)}{2\alpha}})\Big(\frac{n}{\sigma^2}\Big)^{(-1+\gamma)(j-1)}\tilde{O}(n^{\max\{-2+\gamma,\frac{2(1-2\beta+\alpha\gamma)}{1+\alpha}-\kappa(1-2\beta+\alpha\gamma)\}}) \\
=&\tilde{O}(n^{\max\{-2+\gamma+\frac{1-2\gamma\alpha+\alpha}{2\alpha},\frac{2(1-2\beta+\alpha\gamma)}{1+\alpha}+\frac{1-2\gamma\alpha+\alpha}{2\alpha}-\kappa(1-2\beta+\alpha\gamma)\}}) \\
=&\tilde{O}(n^{\max\{-2+\frac{1+\alpha}{2\alpha},\frac{2(1-2\beta+\alpha\gamma)}{1+\alpha}+\frac{1-2\gamma\alpha+\alpha}{2\alpha}-\kappa(1-2\beta+\alpha\gamma)\}}).
\end{aligned}
$$

Since $-2+\frac{1+\alpha}{2\alpha}<0$ and $\gamma=\frac{1+\alpha}{2\alpha}+\kappa_1$, we have

$$
\begin{aligned}
&\frac{2(1-2\beta+\alpha\gamma)}{1+\alpha}+\frac{1-2\gamma\alpha+\alpha}{2\alpha}-\kappa(1-2\beta+\alpha\gamma) \\
=&\frac{2(1-2\beta)}{1+\alpha}+\frac{(\alpha-1)\gamma}{1+\alpha}+\frac{1+\alpha}{2\alpha}-\kappa(1-2\beta+\alpha\gamma) \\
=&\frac{1-2\beta}{\alpha}+\frac{(\alpha-1)(1-2\beta)}{\alpha(\alpha+1)}+\frac{(\alpha-1)(\frac{1+\alpha}{2\alpha}+\kappa_1)}{1+\alpha}+\frac{1+\alpha}{2\alpha}-\kappa(1-2\beta+\alpha(\frac{1+\alpha}{2\alpha}+\kappa_1)) \\
=&\frac{1-2\beta}{\alpha}+1+\frac{(\alpha-1)\kappa_1}{1+\alpha}-\kappa(1-2\beta+\alpha(\frac{1+\alpha}{2\alpha}+\kappa_1))+\frac{(\alpha-1)(1-2\beta)}{\alpha(\alpha+1)} \\
\le&\frac{1-2\beta}{\alpha}+1+\frac{(\alpha-1)\kappa_1}{1+\alpha}-\kappa\alpha(\frac{1+\alpha}{2\alpha}+\kappa_1)=\frac{1-2\beta}{\alpha}+1+(\frac{\alpha-1}{1+\alpha}-\kappa\alpha)\kappa_1-\frac{\kappa(1+\alpha)}{2},
\end{aligned}
$$

where the last inequality holds because $0 \leq \kappa < \frac{\alpha-1}{\alpha(1+\alpha)}$. Choosing κ_1 such that $0 < \kappa_1 < \frac{\kappa(1+\alpha)}{2(\frac{\alpha-1}{1+\alpha}-\kappa\alpha)}$, we have $\frac{2(1-2\beta+\alpha\gamma)}{1+\alpha} + \frac{1-2\gamma\alpha+\alpha}{2\alpha} - \kappa(1-2\beta+\alpha\gamma) < \frac{1-2\beta}{\alpha} + 1$. Hence we have

$$\begin{aligned} T_{2,R}(D_n) =& \tilde{O}(n^{\max\{0,1+\frac{1-2\beta}{\alpha}\}}) + \tilde{O}(n^{\max\{-2+\frac{1+\alpha}{2\alpha}, \frac{2(1-2\beta+\alpha\gamma)}{1+\alpha}+\frac{1-2\gamma\alpha+\alpha}{2\alpha}-\kappa(1-2\beta+\alpha\gamma)\}}) \\ =& \tilde{O}(n^{\max\{0,1+\frac{1-2\beta}{\alpha}\}}). \end{aligned}$$

This concludes the proof of the second statement. □

Lemma 31. *Under Assumptions 4, 5 and 6, with probability at least* $1-5\delta$*, we have*

$$T_1(D_n) = O(n^{\frac{1}{\alpha}}), \tag{73}$$

and

$$T_2(D_n) = \tilde{O}(n^{\max\{0,1+\frac{1-2\beta}{\alpha}\}}). \tag{74}$$

Proof of Lemma 31. Let $R = n^{\frac{2}{1+\alpha}-\kappa}$ where $0 \leq \kappa < \frac{\alpha-1}{\alpha(1+\alpha)}$. By Lemmas 28 and 30, with probability at least $1-5\delta$ we have

$$|T_{1,R}(D_n) - T_1(D_n)| = \tilde{O}(n^{\frac{3-\alpha}{1+\alpha}+\kappa(\alpha-1)}), \tag{75}$$

and

$$|T_{2,R}(D_n) - T_2(D_n)| = \tilde{O}\left(\frac{1}{\delta} n^{\max\{(\frac{2}{1+\alpha}-\kappa)\frac{1-2\beta}{2}, 1+(\frac{3\alpha+1}{\alpha(1+\alpha)}-\kappa)\frac{1-2\beta}{2}, -1+(\frac{1-\alpha}{1+\alpha}+\kappa\alpha), 1+\frac{1-2\beta}{\alpha}+\frac{1-\alpha}{1+\alpha}+\kappa\alpha\}}\right), \tag{76}$$

as well as

$$T_{1,R}(D_n) = O(n^{\frac{1}{\alpha}}), \tag{77}$$

and

$$T_{2,R}(D_n) = \tilde{O}(n^{\max\{0,1+\frac{1-2\beta}{\alpha}\}}). \tag{78}$$

We then have

$$T_1(D_n) \leq T_{1,R}(D_n) + |T_{1,R}(D_n) - T_1(D_n)| = O(n^{\frac{1}{\alpha}}) + \tilde{O}(n^{\frac{3-\alpha}{1+\alpha}+\kappa(\alpha-1)}) = O(n^{\frac{1}{\alpha}}),$$

where in the last equality we use the fact that $\frac{3-\alpha}{1+\alpha} + \kappa(\alpha-1) - \frac{1}{\alpha} = (\kappa - \frac{\alpha-1}{\alpha(1+\alpha)})(\alpha-1) < 0$. This concludes the proof of the first statement.

As for $T_2(D_n)$, we have

$$\begin{aligned} T_2(D_n) &\leq T_{2,R}(D_n) + |T_{2,R}(D_n) - T_2(D_n)| \\ &= \tilde{O}(n^{\max\{0,1+\frac{1-2\beta}{\alpha}\}}) \\ &+ \tilde{O}\left(\frac{1}{\delta} n^{\max\{(\frac{2}{1+\alpha}-\kappa)\frac{1-2\beta}{2}, 1+(\frac{3\alpha+1}{\alpha(1+\alpha)}-\kappa)\frac{1-2\beta}{2}, -1+(\frac{1-\alpha}{1+\alpha}+\kappa\alpha), 1+\frac{1-2\beta}{\alpha}+\frac{1-\alpha}{1+\alpha}+\kappa\alpha\}}\right) \\ &= \tilde{O}(n^{\max\{0,1+\frac{1-2\beta}{\alpha}\}}) + \tilde{O}\left(\frac{1}{\delta} n^{1+(\frac{3\alpha+1}{\alpha(1+\alpha)}-\kappa)\frac{1-2\beta}{2}}\right), \end{aligned}$$

where in the last equality we use the fact that $(\frac{2}{1+\alpha}-\kappa)\frac{1-2\beta}{2} < 0$ and $\frac{1-\alpha}{1+\alpha} + \kappa\alpha < 0$. Since $(\frac{3\alpha+1}{\alpha(1+\alpha)}-\kappa)\frac{1-2\beta}{2} - \frac{1-2\beta}{\alpha} = (\frac{\alpha-1}{\alpha(1+\alpha)}-\kappa)\frac{1-2\beta}{2} < 0$, we have

$$T_{2,R}(D_n) = \tilde{O}(n^{\max\{0,1+\frac{1-2\beta}{\alpha}\}}).$$

This concludes the proof of the second statement. □

Proof of Theorem 7. Using Lemma 31, with probability at least $1-5\tilde{\delta}$, we have

$$\mathbb{E}_\epsilon F^0(D_n) = T_1(D_n) + T_2(D_n) = O(n^{\frac{1}{\alpha}}) + \tilde{O}(n^{\max\{0,1+\frac{1-2\beta}{\alpha}\}}) = \tilde{O}(n^{\max\{\frac{1}{\alpha},0,\frac{1-2\beta}{\alpha}+1\}}).$$

By noting that $\frac{1}{\alpha} > 0$, and letting $\delta = 5\tilde{\delta}$, we get the result. □

C.2 PROOFS RELATED TO THE ASYMPTOTICS OF THE GENERALIZATION ERROR

Lemma 32. *Under Assumptions 4, 5 and 6, with probability at least* $1-\delta$ *over sample inputs* $(x_i)_{i=1}^n$*, we have*

$$G_1(D_n) = \tilde{O}(n^{\frac{1-\alpha}{\alpha}}). \tag{79}$$

Proof of Lemma 32. Let $G_{1,R}(D_n) = \mathbb{E}_{(x_{n+1},y_{n+1})}(T_{1,R}(D_{n+1}) - T_{1,R}(D_n))$. By Lemma 28, we have that

$$\begin{aligned}|G_1(D_n) - G_{1,R}(D_n)| &= \left|\mathbb{E}_{(x_{n+1},y_{n+1})}[T_1(D_{n+1}) - T_{1,R}(D_{n+1})] - [T_1(D_n) - T_{1,R}(D_n)]\right| \\ &= \left|\mathbb{E}_{(x_{n+1},y_{n+1})}O((n+1)R^{1-\alpha})\right| + \left|O(nR^{1-\alpha})]\right| \\ &= O(nR^{1-\alpha})\end{aligned} \tag{80}$$

Define $\eta_R = (\phi_0(x_{n+1}), \phi_1(x_{n+1}), \ldots, \phi_R(x_{n+1}))^T$ and $\widetilde{\Phi}_R = (\Phi_R^T, \eta_R)^T$. As for $G_{1,R}(D_n)$, we have

$$\begin{aligned}G_{1,R}(D_n) &= \mathbb{E}_{(x_{n+1},y_{n+1})}(T_{1,R}(D_{n+1}) - T_{1,R}(D_n)) \\ &= \mathbb{E}_{(x_{n+1},y_{n+1})}\left(\frac{1}{2}\log\det(I + \frac{\widetilde{\Phi}_R\Lambda_R\widetilde{\Phi}_R^T}{\sigma^2}) - \frac{1}{2}\mathrm{Tr}(I - (I + \frac{\widetilde{\Phi}_R\Lambda_R\widetilde{\Phi}_R^T}{\sigma^2})^{-1})\right) \\ &- \left(\frac{1}{2}\log\det(I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2}) - \frac{1}{2}\mathrm{Tr}(I - (I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1})\right) \\ &= \frac{1}{2}\left(\mathbb{E}_{(x_{n+1},y_{n+1})}\log\det(I + \frac{\widetilde{\Phi_R}\Lambda_R\widetilde{\Phi_R}^T}{\sigma^2}) - \log\det(I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})\right) \\ &- \frac{1}{2}\left(\mathbb{E}_{(x_{n+1},y_{n+1})}\mathrm{Tr}(I - (I + \frac{\widetilde{\Phi}_R\Lambda_R\widetilde{\Phi}_R^T}{\sigma^2})^{-1}) - \mathrm{Tr}(I - (I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1})\right).\end{aligned} \tag{81}$$

As for the first term in the right hand side (81), we have

$$\begin{aligned}&\frac{1}{2}\left(\mathbb{E}_{(x_{n+1},y_{n+1})}\log\det(I + \frac{\widetilde{\Phi}_R\Lambda_R\widetilde{\Phi}_R^T}{\sigma^2}) - \log\det(I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})\right) \\ &= \frac{1}{2}\left(\mathbb{E}_{(x_{n+1},y_{n+1})}\log\det(I + \frac{\Lambda_R\widetilde{\Phi}_R^T\widetilde{\Phi}_R}{\sigma^2}) - \log\det(I + \frac{\Lambda_R\Phi_R^T\Phi_R}{\sigma^2})\right) \\ &= \frac{1}{2}\left(\mathbb{E}_{(x_{n+1},y_{n+1})}\log\det(I + \frac{\Lambda_R\Phi_R^T\Phi_R + \eta_R\eta_R^T}{\sigma^2}) - \log\det(I + \frac{\Lambda_R\Phi_R^T\Phi_R}{\sigma^2})\right) \\ &= \frac{1}{2}\left(\mathbb{E}_{(x_{n+1},y_{n+1})}\log\det\left((I + \frac{\Lambda_R\Phi_R^T\Phi_R}{\sigma^2})^{-1}(I + \frac{\Lambda_R\Phi_R^T\Phi_R}{\sigma^2} + \frac{\Lambda_R\eta_R\eta_R^T}{\sigma^2})\right)\right) \\ &= \frac{1}{2}\left(\mathbb{E}_{(x_{n+1},y_{n+1})}\log\det\left(I + (I + \frac{\Lambda_R\Phi_R^T\Phi_R}{\sigma^2})^{-1}\frac{\Lambda_R\eta_R\eta_R^T}{\sigma^2}\right)\right) \\ &= \frac{1}{2}\left(\mathbb{E}_{(x_{n+1},y_{n+1})}\log\left(1 + \frac{1}{\sigma^2}\eta_R^T(I + \frac{\Lambda_R\Phi_R^T\Phi_R}{\sigma^2})^{-1}\Lambda_R\eta_R\right)\right) \\ &\leq \frac{1}{2}\left(\mathbb{E}_{(x_{n+1},y_{n+1})}\frac{1}{\sigma^2}\eta_R^T(I + \frac{\Lambda_R\Phi_R^T\Phi_R}{\sigma^2})^{-1}\Lambda_R\eta_R\right) \\ &= \frac{1}{2\sigma^2}\,\mathrm{Tr}(I + \frac{\Lambda_R\Phi_R^T\Phi_R}{\sigma^2})^{-1}\Lambda_R,\end{aligned}$$

where in the last equality we use the fact that $\mathbb{E}_{(x_{n+1},y_{n+1})}\eta_R\eta_R^T = I$. Let

$$A = (I + \frac{n}{\sigma^2}\Lambda_R)^{-1/2}\Lambda_R^{1/2}(\Phi_R^T\Phi_R - nI)\Lambda_R^{1/2}(I + \frac{n}{\sigma^2}\Lambda_R)^{-1/2}. \tag{82}$$

According to Lemma 20, with probability at least $1-\delta$, we have $\|A\|_2 = \tilde{O}(n^{\frac{1-\alpha}{2\alpha}}(\log R)^{1/2})$. When n is sufficiently large, $\|A\|_2$ is less than 1. By Lemma 24, we have

$$
\begin{aligned}
&\operatorname{Tr}(I + \frac{\Lambda_R \Phi_R^T \Phi_R}{\sigma^2})^{-1}\Lambda_R \\
=& \operatorname{Tr}(I + \frac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R + \sum_{j=1}^{\infty}(-1)^j \operatorname{Tr}\left(\frac{1}{\sigma^2}(I + \frac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R(\Phi_R^T\Phi_R - nI)\right)^j (I + \frac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R \\
=& \operatorname{Tr}(I + \frac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R + \sum_{j=1}^{\infty}(-1)^j \operatorname{Tr}\frac{1}{\sigma^2}(I + \frac{n}{\sigma^2}\Lambda_R)^{-1/2}\Lambda_R^{1/2}A^j(I + \frac{n}{\sigma^2}\Lambda_R)^{-1/2}\Lambda_R^{1/2}.
\end{aligned}
$$

By Lemma 15, we have

$$
\operatorname{Tr}(I + \frac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R \le \sum_{p=1}^{R}\frac{\overline{C_\lambda}p^{-\alpha}}{1 + n\underline{C_\lambda}p^{-\alpha}/\sigma^2} = \tilde{O}(n^{\frac{1-\alpha}{\alpha}}). \tag{83}
$$

Since $\|A\|_2^j = \tilde{O}(n^{\frac{j(1-\alpha)}{2\alpha}}(\log R)^{j/2})$, we have that the absolute values of diagonal entries of A^j are at most $\tilde{O}(n^{\frac{j(1-\alpha)}{2\alpha}})$. Let $(A^j)_{p,p}$ denote the (p,p)-th entry of the matrix A^j. Then we have

$$
\begin{aligned}
&\left|\operatorname{Tr}\frac{1}{\sigma^2}(I + \frac{n}{\sigma^2}\Lambda_R)^{-1/2}\Lambda_R^{1/2}A^j(I + \frac{n}{\sigma^2}\Lambda_R)^{-1/2}\Lambda_R^{1/2}\right| \\
&= \left|\sum_{p=1}^{R}\frac{\lambda_p(A^j)_{p,p}}{1 + n\lambda_p/\sigma^2}\right| \le \sum_{p=1}^{R}\frac{\lambda_p\|A\|_2^j}{1 + n\lambda_p/\sigma^2} \\
&= \tilde{O}(n^{\frac{1-\alpha}{\alpha}})\tilde{O}(n^{\frac{j(1-\alpha)}{2\alpha}}(\log R)^{j/2}),
\end{aligned} \tag{84}
$$

where in the last step we used (83). According to (83) and (84), we have

$$
\begin{aligned}
&\frac{1}{2}\left(\mathbb{E}_{(x_{n+1},y_{n+1})}\log\det(I + \frac{\widetilde{\Phi}_R\Lambda_R\widetilde{\Phi}_R^T}{\sigma^2}) - \log\det(I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})\right) \\
&= \frac{1}{2\sigma^2}\operatorname{Tr}(I + \frac{\Lambda_R\Phi_R^T\Phi_R}{\sigma^2})^{-1}\Lambda_R = \tilde{O}(n^{\frac{1-\alpha}{\alpha}}) + \sum_{j=1}^{\infty}\tilde{O}(n^{\frac{1-\alpha}{\alpha}})\tilde{O}(n^{\frac{j(1-\alpha)}{2\alpha}}) \\
&= \tilde{O}(n^{\frac{1-\alpha}{\alpha}}) + \tilde{O}(n^{\frac{1-\alpha}{\alpha}})\tilde{O}(n^{\frac{1-\alpha}{2\alpha}}(\log R)^{1/2}) = \tilde{O}(n^{\frac{1-\alpha}{\alpha}}(\log R)^{1/2}).
\end{aligned} \tag{85}
$$

Using the Woodbury matrix identity, the second term in the right hand side (81) is given by

$$
\begin{aligned}
&\frac{1}{2}\left(\mathbb{E}_{(x_{n+1},y_{n+1})}\operatorname{Tr}(I - (I + \frac{\widetilde{\Phi}_R\Lambda_R\widetilde{\Phi}_R^T}{\sigma^2})^{-1} - \operatorname{Tr}(I - (I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}\right) \\
&= \frac{1}{2}\left(\mathbb{E}_{(x_{n+1},y_{n+1})}\operatorname{Tr}(\frac{1}{\sigma^2}\widetilde{\Phi}_R(I + \frac{1}{\sigma^2}\Lambda_R\widetilde{\Phi}_R^T\widetilde{\Phi}_R)^{-1}\Lambda_R\widetilde{\Phi}_R^T - \operatorname{Tr}(\frac{1}{\sigma^2}\Phi_R(I + \frac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}\Lambda_R\Phi_R^T\right) \\
&= \frac{1}{2}\left(\mathbb{E}_{(x_{n+1},y_{n+1})}\operatorname{Tr}(\frac{1}{\sigma^2}(I + \frac{1}{\sigma^2}\Lambda_R\widetilde{\Phi}_R^T\widetilde{\Phi}_R)^{-1}\Lambda_R\widetilde{\Phi}_R^T\widetilde{\Phi}_R - \operatorname{Tr}(\frac{1}{\sigma^2}(I + \frac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}\Lambda_R\Phi_R^T\Phi_R\right) \\
&= -\frac{1}{2}\left(\mathbb{E}_{(x_{n+1},y_{n+1})}\operatorname{Tr}(I + \frac{1}{\sigma^2}\Lambda_R\widetilde{\Phi}_R^T\widetilde{\Phi}_R)^{-1} - \operatorname{Tr}(I + \frac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}\right) \\
&= -\frac{1}{2}\left(\mathbb{E}_{(x_{n+1},y_{n+1})}\operatorname{Tr}(I + \frac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R + \frac{1}{\sigma^2}\Lambda_R\eta_R\eta_R^T)^{-1} - \operatorname{Tr}(I + \frac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}\right) \\
&= \frac{1}{2\sigma^2}\left(\mathbb{E}_{(x_{n+1},y_{n+1})}\operatorname{Tr}\frac{(I + \frac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}\Lambda_R\eta_R\eta_R^T(I + \frac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}}{1 + \frac{1}{\sigma^2}\eta_R^T(I + \frac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}\Lambda_R\eta_R}\right),
\end{aligned}
$$

where the last equality uses the Sherman–Morrison formula. Since $\eta_R^T(I+\frac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}\Lambda_R\eta_R = \eta_R^T(\Lambda^{-1}+\frac{1}{\sigma^2}\Phi_R^T\Phi_R)^{-1}\eta_R \geq 0$, we get

$$
\begin{aligned}
&\frac{1}{2\sigma^2}\left(\mathbb{E}_{(x_{n+1},y_{n+1})}\operatorname{Tr}\frac{(I+\frac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}\Lambda_R\eta_R\eta_R^T(I+\frac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}}{1+\frac{1}{\sigma^2}\eta_R^T(I+\frac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}\Lambda_R\eta_R}\right)\\
&\leq \frac{1}{2\sigma^2}\left(\mathbb{E}_{(x_{n+1},y_{n+1})}\operatorname{Tr}(I+\frac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}\Lambda_R\eta_R\eta_R^T(I+\frac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}\right)\\
&= \frac{1}{2\sigma^2}\operatorname{Tr}(I+\frac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}\Lambda_R(I+\frac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}\\
&= \frac{1}{2\sigma^2}\operatorname{Tr}\Lambda_R^{1/2}(I+\frac{1}{\sigma^2}\Lambda_R^{1/2}\Phi_R^T\Phi_R\Lambda_R^{1/2})^{-1}\Lambda_R^{1/2}(I+\frac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}\\
&= \frac{1}{2\sigma^2}\operatorname{Tr}(I+\frac{1}{\sigma^2}\Lambda_R^{1/2}\Phi_R^T\Phi_R\Lambda_R^{1/2})^{-1}\Lambda_R^{1/2}(I+\frac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}\Lambda_R^{1/2}\\
&= \frac{1}{2\sigma^2}\operatorname{Tr}(I+\frac{1}{\sigma^2}\Lambda_R^{1/2}\Phi_R^T\Phi_R\Lambda_R^{1/2})^{-1}\Lambda_R(I+\frac{1}{\sigma^2}\Lambda_R^{1/2}\Phi_R^T\Phi_R\Lambda_R^{1/2})^{-1}\\
&= \frac{1}{2\sigma^2}\|\Lambda_R^{1/2}(I+\frac{1}{\sigma^2}\Lambda_R^{1/2}\Phi_R^T\Phi_R\Lambda_R^{1/2})^{-1}\|_F^2\\
&= \frac{1}{2\sigma^2}\|\Lambda_R^{1/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}(I+\frac{1}{\sigma^2}A)^{-1}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}\|_F^2,
\end{aligned}
$$

where in the penultimate equality we use $\operatorname{Tr}(BB^T)=\|B\|_F^2$, $\|B\|_F$ is the Frobenius norm of A, and in the last equality we use the definition of A (82). Then we have

$$
\begin{aligned}
&\frac{1}{2\sigma^2}\|\Lambda_R^{1/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}(I+\frac{1}{\sigma^2}A)^{-1}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}\|_F^2\\
&= \frac{1}{2\sigma^2}\|\Lambda_R^{1/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}(I+\sum_{j=1}^{\infty}(-1)^j\frac{1}{\sigma^{2j}}A^j)(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}\|_F^2\\
&= \frac{1}{2\sigma^2}\|\Lambda_R^{1/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}+\sum_{j=1}^{\infty}(-1)^j\frac{1}{\sigma^{2j}}\Lambda_R^{1/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}A^j(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}\|_F^2.
\end{aligned}
\tag{86}
$$

By Lemma 15, we have

$$
\|\Lambda_R^{1/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\|_F \leq \sqrt{\sum_{p=1}^{R}\frac{\overline{C_\lambda}p^{-\alpha}}{(1+n\underline{C_\lambda}p^{-\alpha}/\sigma^2)^2})} = \tilde{O}(n^{\frac{1-\alpha}{2\alpha}}). \tag{87}
$$

Since $\|A\|_2^j = \tilde{O}(n^{\frac{j(1-\alpha)}{2\alpha}}(\log R)^{j/2})$, we have

$$
\begin{aligned}
&\|\frac{1}{\sigma^{2j}}\Lambda_R^{1/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}A^j(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2})\|_F\\
&= \frac{1}{\sigma^{2j}}\|\Lambda_R^{1/2}\|_F\|(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}A^j(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2})\|_2\\
&= \frac{1}{\sigma^{2j}}O(1)\tilde{O}(n^{\frac{j(1-\alpha)}{2\alpha}}(\log R)^{j/2}),
\end{aligned}
\tag{88}
$$

where in the first inequality we use the fact that $\|AB\|_F \leq \|A\|_F\|B\|_2$ when B is symmetric. According to (87) and (88), we have

$$\begin{aligned}
&\frac{1}{2}\left(\mathbb{E}_{(x_{n+1},y_{n+1})}\mathrm{Tr}(I-(I+\frac{\widetilde{\Phi}_R\Lambda_R\widetilde{\Phi}_R^T}{\sigma^2})^{-1}-\mathrm{Tr}(I-(I+\frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}\right)\\
&\leq \frac{1}{2\sigma^2}\,\mathrm{Tr}(I+\frac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}\Lambda_R(I+\frac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}\\
\leq&\frac{1}{2\sigma^2}\|\Lambda_R^{1/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}+\sum_{j=1}^{\infty}(-1)^j\frac{1}{\sigma^{2j}}\Lambda_R^{1/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}A^j(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}\|_F^2\\
\leq&\frac{1}{2\sigma^2}\left(\|\Lambda_R^{1/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\|_F+\sum_{j=1}^{\infty}\left\|\frac{1}{\sigma^{2j}}\Lambda_R^{1/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}A^j(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}\right\|_F\right)^2\\
\leq&\frac{1}{2\sigma^2}\left(\tilde{O}(n^{\frac{1-\alpha}{2\alpha}})+\sum_{j=1}^{\infty}\frac{1}{\sigma^{2j}}O(1)\tilde{O}(n^{\frac{j(1-\alpha)}{2\alpha}}(\log R)^{j/2})\right)^2\\
\leq&\frac{1}{2\sigma^2}\left(\tilde{O}(n^{\frac{1-\alpha}{2\alpha}})+\frac{1}{\sigma^2}\tilde{O}(n^{\frac{1-\alpha}{2\alpha}}(\log R)^{1/2})\right)^2=\tilde{O}(n^{\frac{1-\alpha}{\alpha}}(\log R)^{1/2}).
\end{aligned}\tag{89}$$

Combining (85) and (89) we get that $G_{1,R}(D_n)=\tilde{O}(n^{\frac{1-\alpha}{\alpha}}(\log R)^{1/2})$. From (80) we have that $G_1(D_n)\leq G_{1,R}(D_n)+|G_1(D_n)-G_{1,R}(D_n)|=\tilde{O}(n^{\frac{1-\alpha}{\alpha}}(\log R)^{1/2})+O(nR^{1-\alpha})$. Choosing $R=n^{\frac{2\alpha-1}{\alpha(\alpha-1)}}$ we get that $G_1(D_n)=\tilde{O}(n^{\frac{1-\alpha}{\alpha}})$, which concludes the proof. □

Lemma 33. *Assume that $\|\xi\|_2=1$. When n is sufficiently large, with probability at least $1-2\delta$ we have*

$$\|(I+\frac{1}{\sigma^2}\Phi_S\Lambda_S\Phi_S^T)^{-1}\Phi_S\Lambda_S\xi\|_2=O(\sqrt{\frac{1}{\delta}n}\cdot n^{-1}).\tag{90}$$

Proof of Lemma 33. Using the Woodbury matrix identity, we have that

$$\begin{aligned}
((I+\frac{1}{\sigma^2}\Phi_S\Lambda_S\Phi_S^T)^{-1}\Phi_S\Lambda_S\xi=&\left[I-\Phi_S(\sigma^2I+\Lambda_S\Phi_S^T\Phi_S)^{-1}\Lambda_S\Phi_S^T\right]\Phi_S\Lambda_S\xi\\
=&\Phi_S\Lambda_S\xi-\Phi_S(\sigma^2I+\Lambda_S\Phi_S^T\Phi_S)^{-1}\Lambda_S\Phi_S^T\Phi_S\Lambda_S\xi\\
=&\sigma^2\Phi_S(\sigma^2I+\Lambda_S\Phi_S^T\Phi_S)^{-1}\Lambda_S\xi.
\end{aligned}\tag{91}$$

Let $A=(I+\frac{n}{\sigma^2}\Lambda_S)^{-\gamma/2}\Lambda_S^{\gamma/2}(\Phi_S^T\Phi_S-nI)\Lambda_S^{\gamma/2}(I+\frac{n}{\sigma^2}\Lambda_S)^{-\gamma/2}$, where $\gamma>\frac{1+\alpha}{2\alpha}$. By Lemma 20, with probability at least $1-\delta$, we have $\|A\|_2=\tilde{O}(n^{\frac{1-2\gamma\alpha+\alpha}{2\alpha}})$. When n is sufficiently large, $\|A\|_2$ is less than 1. By Lemma 24, we have

$$\begin{aligned}
&(I+\frac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}\\
=&(I+\frac{n}{\sigma^2}\Lambda_S)^{-1}+\sum_{j=1}^{\infty}(-1)^j\left(\frac{1}{\sigma^2}(I+\frac{n}{\sigma^2}\Lambda_S)^{-1}\Lambda_S(\Phi_S^T\Phi_S-nI)\right)^j(I+\frac{n}{\sigma^2}\Lambda_S)^{-1}.
\end{aligned}$$

Then we have

$$\begin{aligned}
&\|(\sigma^2I+\Lambda_S\Phi_S^T\Phi_S)^{-1}\Lambda_S\xi\|_2\\
=&\frac{1}{\sigma^2}\left\|\left((I+\frac{n}{\sigma^2}\Lambda_S)^{-1}+\sum_{j=1}^{\infty}(-1)^j\left(\frac{1}{\sigma^2}(I+\frac{n}{\sigma^2}\Lambda_S)^{-1}\Lambda_S(\Phi_S^T\Phi_S-nI)\right)^j(I+\frac{n}{\sigma^2}\Lambda_S)^{-1}\right)\Lambda_S\xi\right\|_2\\
\leq&\frac{1}{\sigma^2}\left(\|(I+\frac{n}{\sigma^2}\Lambda_S)^{-1}\Lambda_S\xi\|_2+\sum_{j=1}^{\infty}\left\|\left(\frac{1}{\sigma^2}(I+\frac{n}{\sigma^2}\Lambda_S)^{-1}\Lambda_S(\Phi_S^T\Phi_S-nI)\right)^j(I+\frac{n}{\sigma^2}\Lambda_S)^{-1}\Lambda_S\xi\right\|_2\right).
\end{aligned}\tag{92}$$

For the first term in the right hand side of the last equation, we have

$$\|(I+\frac{n}{\sigma^2}\Lambda_S)^{-1}\Lambda_S\xi\|_2 \leq \|(I+\frac{n}{\sigma^2}\Lambda_S)^{-1}\Lambda_S\|_2\|\xi\|_2 \leq \frac{\sigma^2}{n} = O(n^{-1}). \tag{93}$$

Using the fact that $\|A\|_2 = \tilde{O}(n^{\frac{1-2\gamma\alpha+\alpha}{2\alpha}})$ and $\|(I+\frac{n}{\sigma^2}\Lambda_S)^{-1}\Lambda_S\|_2 \leq n^{-1}$, we have

$$\begin{aligned}
&\left\|\left(\frac{1}{\sigma^2}(I+\frac{n}{\sigma^2}\Lambda_S)^{-1}\Lambda_S(\Phi_S^T\Phi_S - nI)\right)^j (I+\frac{n}{\sigma^2}\Lambda_S)^{-1}\Lambda_S\xi\right\|_2 \\
=&\left\|(I+\frac{n}{\sigma^2}\Lambda_S)^{-1+\frac{\gamma}{2}}\Lambda_S^{1-\frac{\gamma}{2}}\left(A(I+\frac{n}{\sigma^2}\Lambda_S)^{-1+\gamma}\Lambda_S^{1-\gamma}\right)^{j-1}A(I+\frac{n}{\sigma^2}\Lambda_S)^{-1+\frac{\gamma}{2}}\Lambda_S^{-\frac{\gamma}{2}}\Lambda_S\xi\right\|_2 \\
\leq& n^{-1+\frac{\gamma}{2}+(-1+\gamma)(j-1)}\tilde{O}(n^{\frac{(1-2\gamma\alpha+\alpha)j}{2\alpha}})\|(I+\frac{n}{\sigma^2}\Lambda_S)^{-1+\frac{\gamma}{2}}\Lambda_S^{1-\frac{\gamma}{2}}\xi\|_2 \\
=&\tilde{O}(n^{-\frac{\gamma}{2}+\frac{(1-\alpha)j}{2\alpha}})\|(I+\frac{n}{\sigma^2}\Lambda_S)^{-1+\frac{\gamma}{2}}\Lambda_S^{1-\frac{\gamma}{2}}\|_2\|\xi\|_2 \\
=&\tilde{O}(n^{-\frac{\gamma}{2}+\frac{(1-\alpha)j}{2\alpha}})O(n^{-1+\gamma/2}) \\
=&\tilde{O}(n^{-1+\frac{(1-\alpha)j}{2\alpha}}).
\end{aligned} \tag{94}$$

Using (92), (93) and (94), we have

$$\begin{aligned}
&\|(\sigma^2 I+\Lambda_S\Phi_S^T\Phi_S)^{-1}\Lambda_S\xi\|_2 \\
=&\tilde{O}(n^{-1}) + \sum_{j=1}^{\infty}\sigma^{-2j}\tilde{O}(n^{-1+\frac{(1-\alpha)j}{2\alpha}}) \\
=&\tilde{O}(n^{-1}) + \sigma^{-2}\tilde{O}(n^{-1+\frac{1-\alpha}{2\alpha}}) \\
=&\tilde{O}(n^{-1}).
\end{aligned} \tag{95}$$

By Corollary 19, with probability at least $1-\delta$, we have

$$\begin{aligned}
\|\Phi_S(\sigma^2 I+\Lambda_S\Phi_S^T\Phi_S)^{-1}\Lambda_S\xi\|_2 =&\tilde{O}(\sqrt{\frac{1}{\delta}n}\|(\sigma^2 I+\Lambda_S\Phi_S^T\Phi_S)^{-1}\Lambda_S\xi\|_2) \\
=&\tilde{O}(\sqrt{\frac{1}{\delta}n}\cdot n^{-1}).
\end{aligned}$$

From (91) we get $\|(I+\frac{1}{\sigma^2}\Phi_S\Lambda_S\Phi_S^T)^{-1}f_S(\mathbf{x})\|_2 = \tilde{O}(\sqrt{\frac{1}{\delta}n}\cdot n^{-1})$. This concludes the proof. □

Lemma 34. *Under Assumptions 4, 5 and 6, with probability at least $1-6\delta$ over sample inputs $(x_i)_{i=1}^n$, we have*

$$G_2(D_n) = \tilde{O}(n^{\max\{-2,\frac{1-2\beta}{\alpha}\}}). \tag{96}$$

Proof of Lemma 34. Let $G_{2,S}(D_n) = \mathbb{E}_{(x_{n+1},y_{n+1})}(T_{2,S}(D_{n+1}) - T_{2,S}(D_n))$. By Lemma 29, when $S > n$, with probability at least $1-3\delta$ we have that

$$\begin{aligned}
|G_2(D_n) - G_{2,S}(D_n)| &= \left|\mathbb{E}_{(x_{n+1},y_{n+1})}[T_2(D_{n+1}) - T_{2,S}(D_{n+1})] - [T_2(D_n) - T_{2,S}(D_n)]\right| \\
&= \left|\mathbb{E}_{(x_{n+1},y_{n+1})}\tilde{O}\left(\frac{1}{\delta}(n+1)S^{\max\{1/2-\beta,1-\alpha\}}\right)\right| + \left|\tilde{O}\left(\frac{1}{\delta}nS^{\max\{1/2-\beta,1-\alpha\}}\right)\right| \\
&= \tilde{O}\left(\frac{1}{\delta}nS^{\max\{1/2-\beta,1-\alpha\}}\right).
\end{aligned} \tag{97}$$

Define $\eta_S = (\phi_0(x_{n+1}), \phi_1(x_{n+1}), \ldots, \phi_S(x_{n+1}))^T$ and $\widetilde{\Phi}_S = (\Phi_S^T, \eta_S)^T$. In Eq. (71) in the proof of Lemma 30, we showed that

$$\begin{aligned}
T_{2,S}(D_n) &= \frac{1}{2\sigma^2}\mu_S^T\Phi_S^T(I+\frac{1}{\sigma^2}\Phi_S\Lambda_S\Phi_S^T)^{-1}\Phi_S\mu_S \\
&= \frac{1}{2}\mu_S^T\Lambda_S^{-1}\mu_S - \frac{1}{2}\mu_S^T\Lambda_S^{-1}(I+\frac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}\mu_S.
\end{aligned}$$

We have

$$
\begin{aligned}
G_{2,S}(D_n) &= \mathbb{E}_{(x_{n+1},y_{n+1})}(T_{2,S}(D_{n+1}) - T_{2,S}(D_n)) \\
&= \mathbb{E}_{(x_{n+1},y_{n+1})}\left(\frac{1}{2}\mu_S^T\Lambda_S^{-1}\mu_S - \frac{1}{2}\mu_S^T\Lambda_S^{-1}(I+\frac{1}{\sigma^2}\Lambda_S\widetilde{\Phi}_S^T\widetilde{\Phi}_S)^{-1}\mu_S\right) \\
&- \left(\frac{1}{2}\mu_S^T\Lambda_S^{-1}\mu_S - \frac{1}{2}\mu_S^T\Lambda_S^{-1}(I+\frac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}\mu_S)\right) \\
&= \mathbb{E}_{(x_{n+1},y_{n+1})}\left(\frac{1}{2}\mu_S^T\Lambda_S^{-1}(I+\frac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}\mu_S - \frac{1}{2}\mu_S^T\Lambda_S^{-1}(I+\frac{1}{\sigma^2}\Lambda_S\widetilde{\Phi}_S^T\widetilde{\Phi}_S)^{-1}\mu_S\right) \\
&= \mathbb{E}_{(x_{n+1},y_{n+1})}\left(\frac{1}{2\sigma^2}\mu_S^T\Lambda_S^{-1}\frac{(I+\frac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}\Lambda_S\eta_S\eta_S^T(I+\frac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}}{1+\frac{1}{\sigma^2}\eta_S^T(I+\frac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}\Lambda_S\eta_S}\mu_S)\right) \\
&= \mathbb{E}_{(x_{n+1},y_{n+1})}\left(\frac{1}{2\sigma^2}\frac{\mu_S^T(I+\frac{1}{\sigma^2}\Phi_S^T\Phi_S\Lambda_S)^{-1}\eta_S\eta_S^T(I+\frac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}\mu_S}{1+\frac{1}{\sigma^2}\eta_S^T(I+\frac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}\Lambda_S\eta_S})\right) \\
&\leq \mathbb{E}_{(x_{n+1},y_{n+1})}\left(\frac{1}{2\sigma^2}\mu_S^T(I+\frac{1}{\sigma^2}\Phi_S^T\Phi_S\Lambda_S)^{-1}\eta_S\eta_S^T(I+\frac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}\mu_S\right) \\
&= \frac{1}{2\sigma^2}\mu_S^T(I+\frac{1}{\sigma^2}\Phi_S^T\Phi_S\Lambda_S)^{-1}(I+\frac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}\mu_S \\
&= \frac{1}{2\sigma^2}\|(I+\frac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}\mu_S\|_2^2,
\end{aligned}
\tag{98}
$$

where in the third to last equality we used the Sherman–Morrison formula, in the last inequality we used the fact that $\eta_S^T(I+\frac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}\Lambda_S\eta_S = \eta_S^T(\Lambda^{-1}+\frac{1}{\sigma^2}\Phi_S^T\Phi_S)^{-1}\eta_S \geq 0$, and in the last equality we used the fact that $\mathbb{E}_{(x_{n+1},y_{n+1})}\eta_S\eta_S^T = I$.

Let $\hat{\mu}_R = (\mu_1,\ldots,\mu_R,0,\ldots,0) \in \mathbb{R}^S$. Then we have

$$
\|(I+\frac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}\mu_S\|_2 \leq \|(I+\frac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}\hat{\mu}_R\|_2 + \|(I+\frac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}(\mu_S-\hat{\mu}_R)\|_2. \tag{99}
$$

Choose $R = n^{\frac{2}{1+\alpha}-\kappa}$ where $0 \leq \kappa < \frac{\alpha-1}{\alpha(1+\alpha)}$. In Lemma 26, (54), we showed that with probability at least $1-\delta$,

$$
\|(\sigma^2 I + \Lambda_R\Phi_R^T\Phi_R)^{-1}\mu_R\|_2 = \tilde{O}(n^{\max\{-1,\frac{1-2\beta}{2\alpha}\}}). \tag{100}
$$

The same proof holds if we replace Φ_R with Φ_S, Λ_R with Λ_S, and μ_R with $\hat{\mu}_R$. We have

$$
\|(\sigma^2 I + \Lambda_S\Phi_S^T\Phi_S)^{-1}\hat{\mu}_R\|_2 = \tilde{O}(n^{\max\{-1,\frac{1-2\beta}{2\alpha}\}}). \tag{101}
$$

Next we bound $\|(I+\frac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}(\mu_S-\hat{\mu}_R)\|_2$. By Assumption 5, we have that $\|\mu_S-\hat{\mu}_R\|_2 = O(R^{\frac{1-2\beta}{2}})$. For any $\xi \in \mathbb{R}^S$ and $\|\xi\|_2 = 1$, using the Woodbury matrix identity, with probability at least $1-2\delta$ we have

$$
\begin{aligned}
&|\xi^T(I+\frac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}(\mu_S-\hat{\mu}_R)| \\
&= |\xi^T\left(I - \frac{1}{\sigma^2}\Lambda_S\Phi_S^T(I+\frac{1}{\sigma^2}\Phi_S\Lambda_S\Phi_S^T)^{-1}\Phi_S\right)(\mu_S-\hat{\mu}_R)| \\
&= |\xi^T(\mu_S-\hat{\mu}_R) - \frac{1}{\sigma^2}\xi^T\Lambda_S\Phi_S^T(I+\frac{1}{\sigma^2}\Phi_S\Lambda_S\Phi_S^T)^{-1}\Phi_S(\mu_S-\hat{\mu}_R)| \\
&\leq \|\xi\|_2\|\mu_S-\hat{\mu}_R\|_2 + \frac{1}{\sigma^2}|\xi^T\Lambda_S\Phi_S^T(I+\frac{1}{\sigma^2}\Phi_S\Lambda_S\Phi_S^T)^{-1}\Phi_S(\mu_S-\hat{\mu}_R)| \\
&\leq O(R^{\frac{1-2\beta}{2}}) + \frac{1}{\sigma^2}\|(I+\frac{1}{\sigma^2}\Phi_S\Lambda_S\Phi_S^T)^{-1}\Phi_S\Lambda_S\xi\|_2\|\Phi_S(\mu_S-\hat{\mu}_R)\|_2 \\
&= O(R^{\frac{1-2\beta}{2}}) + O(\sqrt{\frac{1}{\delta}n}\cdot n^{-1})O(\sqrt{\frac{1}{\delta}n}R^{\frac{1-2\beta}{2}}) \\
&= O(\frac{1}{\delta}R^{\frac{1-2\beta}{2}}),
\end{aligned}
$$

where in the second to last step we used Corollary 19 to show $\|\Phi_S(\mu_S - \hat{\mu}_R)\|_2 = O(\sqrt{\frac{1}{\delta}n}R^{\frac{1-2\beta}{2}})$ with probability at least $1-\delta$, and Lemma 33 to show that $\|(I + \frac{1}{\sigma^2}\Phi_S\Lambda_S\Phi_S^T)^{-1}\Phi_S\Lambda_S\xi\|_2 = O(\sqrt{\frac{1}{\delta}n} \cdot n^{-1})$ with probability at least $1-\delta$. Choosing $R = n^{\frac{2}{1+\alpha}}$ we have

$$|\xi^T(I + \frac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}(\mu_S - \hat{\mu}_R)| = O(\frac{1}{\delta}n^{\frac{1-2\beta}{1+\alpha}}).$$

Since ξ is arbitrary, we have $\|(I + \frac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}(\mu_S - \hat{\mu}_R)\|_2 = O(\frac{1}{\delta}n^{\frac{1-2\beta}{1+\alpha}})$. From (99) and (101), we have

$$\|(I + \frac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}\mu_S\|_2 \leq \tilde{O}(n^{\max\{-1,\frac{1-2\beta}{2\alpha}\}}) + O(\frac{1}{\delta}n^{\frac{1-2\beta}{1+\alpha}}) = \tilde{O}(n^{\max\{-1,\frac{1-2\beta}{2\alpha}\}}). \quad (102)$$

Hence $G_{2,S}(D_n) \leq \frac{1}{2\sigma^2}\|(I + \frac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}\mu_S\|_2^2 = \tilde{O}(n^{\max\{-2,\frac{1-2\beta}{\alpha}\}})$. Then by (97), $G_2(D_n) = \tilde{O}(n^{\max\{-2,\frac{1-2\beta}{\alpha}\}}) + \tilde{O}\big(\frac{1}{\delta}nS^{\max\{1/2-\beta,1-\alpha\}}\big)$. Choosing $S = n^{\frac{1+\min\{2,\frac{2\beta-1}{\alpha}\}}{\min\{\beta-1/2,\alpha-1\}}+1}$, we get the result. □

Proof of Theorem 9. From Lemmas 32 and 34 we have that with probability at least $1 - 7\tilde{\delta}$, $\mathbb{E}_{\boldsymbol{\epsilon}}G(D_n) = O(n^{\max\{\frac{1}{\alpha}-1,\frac{1-2\beta}{\alpha}\}})$ because $\frac{1}{\alpha} - 1 > -2$. Letting $\delta = 7\tilde{\delta}$, the proof is complete. □

C.3 Proofs related to the expected excess mean squared generalization error

Proof of Theorem 12. For $\mu_0 = 0$, we can show that

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\epsilon}}M(D_n) &= \mathbb{E}_{\boldsymbol{\epsilon}}\mathbb{E}_{x_{n+1}}[\bar{m}(x_{n+1}) - f(x_{n+1})]^2 \\
&= \mathbb{E}_{\boldsymbol{\epsilon}}\mathbb{E}_{x_{n+1}}[K_{x_{n+1}\mathbf{x}}(K_n + \sigma^2 I_n)^{-1}\mathbf{y} - f(x_{n+1})]^2 \\
&= \mathbb{E}_{\boldsymbol{\epsilon}}\mathbb{E}_{x_{n+1}}[\eta^T\Lambda\Phi^T[\Phi\Lambda\Phi^T + \sigma^2 I_n)^{-1}(\Phi\mu + \boldsymbol{\epsilon}) - \eta^T\mu]^2 \\
&= \mathbb{E}_{\boldsymbol{\epsilon}}\mathbb{E}_{x_{n+1}}[\eta^T\Lambda\Phi^T(\Phi\Lambda\Phi^T + \sigma^2 I_n)^{-1}\boldsymbol{\epsilon}]^2 \\
&+ \mathbb{E}_{x_{n+1}}\big[\eta^T\big(\Lambda\Phi^T(\Phi\Lambda\Phi^T + \sigma^2 I_n)^{-1}\Phi - I\big)\mu\big]^2 \\
&= \operatorname{Tr}\Lambda\Phi^T(\Phi\Lambda\Phi^T + \sigma^2 I_n)^{-2}\Phi\Lambda \\
&+ \mu^T\left(I + \frac{1}{\sigma^2}\Phi^T\Phi\Lambda\right)^{-1}\left(I + \frac{1}{\sigma^2}\Lambda\Phi^T\Phi\right)^{-1}\mu \\
&= \operatorname{Tr}(I + \frac{\Lambda\Phi^T\Phi}{\sigma^2})^{-1}\Lambda - \operatorname{Tr}(I + \frac{\Lambda\Phi^T\Phi}{\sigma^2})^{-2}\Lambda + \|(I + \frac{1}{\sigma^2}\Lambda\Phi^T\Phi)^{-1}\mu\|_2^2.
\end{aligned}$$

According to (83) and (89) from the proof of Lemma 32, we have

$$\|(I + \frac{1}{\sigma^2}\Lambda\Phi^T\Phi)^{-1}\mu\|_2^2 = \tilde{O}(n^{\frac{1-\alpha}{\alpha}}).$$

According to (102) from the proof of Lemma 34, we have

$$\operatorname{Tr}(I + \frac{\Lambda\Phi^T\Phi}{\sigma^2})^{-1}\Lambda - \operatorname{Tr}(I + \frac{\Lambda\Phi^T\Phi}{\sigma^2})^{-2}\Lambda = \tilde{O}(n^{\max\{-2,\frac{1-2\beta}{\alpha}\}}).$$

Combining the above two equations we get

$$\mathbb{E}_{\boldsymbol{\epsilon}}M(D_n) = \tilde{O}(n^{\max\{\frac{1-\alpha}{\alpha},-2,\frac{1-2\beta}{\alpha}\}}) = \tilde{O}(n^{\max\{\frac{1-\alpha}{\alpha},\frac{1-2\beta}{\alpha}\}}).$$

When $\mu_0 > 0$, we have

$$\|(I + \frac{1}{\sigma^2}\Lambda\Phi^T\Phi)^{-1}\mu\|_2^2 \leq \|\mu\|_2^2 = \|f\|_2^2 = O(1).$$

Then

$$\mathbb{E}_{\boldsymbol{\epsilon}}M(D_n) = \tilde{O}(n^{\max\{\frac{1-\alpha}{\alpha}\}}) + O(1) = O(1).$$

□