Power-law asymptotics of the generalization error for GP regression under power-law priors and targets

Hui Jin[*] **Pradeep Kr. Banerjee**[†] **Guido Montúfar**[‡]

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 Bayesian generalization error behaves as $\Theta(n^{\max\{\frac{1}{\alpha}-1,\frac{1-2\beta}{\alpha}\}})$ with high probability over the draw of n input samples. Infinitely wide neural networks can be related to GPR with respect to the Neural Network Gaussian Process kernel, 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. [48]. 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. In this paper, we study the learning curves for Gaussian process regression (GPR). Our main result characterizes the asymptotics of the learning curve 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.

There is a well known correspondence between kernel methods and infinite neural networks (NNs) [27, 47, 22, 14, 29, 15, 50]. This enables *exact* Bayesian inference in the associated GP model for infinite-width NNs on regression tasks and has led to recent breakthroughs in our understanding of overparameterized NNs [19, 23, 3, 6, 13, 51, 7]. The most prominent kernels associated with infinite-width NNs are the Neural Network Gaussian Process (NNGP) kernel [22, 14] and the Neural Tangent Kernel (NTK) [19], which in several cases are known to have a power-law spectrum [7, 33, 42].

Contributions. We show that when the eigenspectrum of the prior decays with rate α and the eigenexpansion coefficients of the target function decay with rate β, the negative log-marginal likelihood behaves as $\Theta(n^{\max\{\frac{1}{\alpha},\frac{1-2\beta}{\alpha}+1\}})$ (Theorem 4) and the generalization error behaves as $\Theta(n^{\max\{\frac{1}{\alpha}-1,\frac{1-2\beta}{\alpha}\}})$ (Theorem 5) with high probability over the draw of n input samples. 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 [37] (vide Remark 6). 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 [11]. For a discussion of related work, see Appendix A.

Bayesian Deep Learning workshop, NeurIPS 2021.

[*]UCLA ..... huijin@ucla.edu
[†]MPI MiS. ..... pradeep@mis.mpg.de
[‡]UCLA & MPI MiS. ..... montufar@math.ucla.edu

2 Asymptotic analysis of generalization error of GPR with power-law priors

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 $\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_{\text{true}}^2)$. 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 $q(y|x) = \mathcal{N}(y|f(x), \sigma_{\text{true}}^2)$. 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., $f \sim \mathcal{GP}(0, k)$. By Bayes' rule, the posterior distribution of f given the training data is given by $d\Pi_n(f|D_n) = \frac{1}{Z(D_n)} \prod_{i=1}^n \mathcal{N}(y_i|f(x_i), \sigma_{\text{model}}^2) d\Pi_0(f)$, where $d\Pi_0(f)$ is the prior distribution and $Z(D_n) = \int \prod_{i=1}^n \mathcal{N}(y_i|f(x_i), \sigma_{\text{model}}^2) d\Pi_0(f)$ is the *marginal likelihood*. The form of the GPR posterior mean and covariance is given in Appendix B.

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

$$G(D_n) = \int q(x, y) \log \frac{q(y|x)}{p_n(y|x, D_n)} dx dy. \tag{1}$$

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 the normalized SC:

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

The generalization error (1) and the normalized SC relate as follows [45, Theorem 1.2]:

$$G(D_n) = \mathbb{E}_{(x_{n+1}, y_{n+1})} F^0(D_{n+1}) - F^0(D_n), \tag{3}$$

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}). Our goal is to derive the asymptotics of the expected normalized SC, $\mathbb{E}_{\boldsymbol{\epsilon}} F^0(D_n)$, and the expected generalization error, $\mathbb{E}_{\boldsymbol{\epsilon}} G(D_n)$, where $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$ is the noise vector.

By 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 $L_k\colon L^2(\Omega, \rho) \mapsto L^2(\Omega, \rho)$; $(L_k f)(x) = \int_\Omega k(x, s) f(s) \mathrm{d}\rho(s)$ and $(\lambda_p)_{p \geq 1}$ are positive eigenvalues, $\lambda_1 \geq \lambda_2 \geq \cdots > 0$ [48]. The target $f(x)$ is decomposed into the orthonormal $(\phi_p(x))_{p \geq 1}$ and its orthogonal complement as

$$f(x) = \sum_{p=1}^\infty \mu_p \phi_p(x) + \mu_0 \phi_0(x) \in L^2(\Omega, \rho), \tag{4}$$

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 $\phi_0(x) \in \{\phi_p(x) : p \geq 1\}^\perp$. We shall make the following assumptions:

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

$$\underline{C_\lambda} p^{-\alpha} \leq \lambda_p \leq \overline{C_\lambda} p^{-\alpha}, \tag{5}$$

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

This assumption is adopted in many recent works [4, 10, 21, 28]. [42] derived the exact value of α when the kernel has a homogeneous singularity on its diagonal, e.g., the arc-cosine kernel.

Assumption 2 (Power law decay of coefficients of decomposition)**.** *Let $C_\mu, \underline{C_\mu} > 0$ and $\beta > 1/2$ be positive constants and let $\{p_i\}_{i \geq 1}$ be an increasing integer sequence such that $\sup_{i \geq 1}(p_{i+1} - p_i) < \infty$. The coefficients $(\mu_p)_{p \geq 1}$ of the decomposition* (4) *of the target function follow the power law*

$$|\mu_p| \leq C_\mu p^{-\beta}, \ \forall p \geq 1 \quad \textit{and} \quad |\mu_{p_i}| \geq \underline{C_\mu} {p_i}^{-\beta}, \ \forall i \geq 1. \tag{6}$$

When $(\phi_p(x))_p$ is the Fourier or the spherical harmonics basis and $f(x)$ satisfies certain smoothness conditions, $(\mu_p)_p$ decay at least as a power law [7]. [42] gave examples satisfying Assumption 2.

Assumption 3 (Boundedness of eigenfunctions)**.** *The eigenfunctions* $(\phi_p(x))_{p\geq 0}$ *satisfy*

$$\|\phi_0\|_\infty \leq C_\phi \quad \textit{and} \quad \|\phi_p\|_\infty \leq C_\phi p^\tau,\ p \geq 1, \tag{7}$$

where C_ϕ *and* τ *are two positive constants which satisfy* $\tau < \frac{\alpha-1}{2}$.

Next we state our main results and give the proofs in Appendix D.

Theorem 4 (Asymptotics of the normalized SC)**.** *In the case that* $\mu_0 > 0$ *and* $\sigma^2_{\text{model}} = \sigma^2_{\text{true}} = \sigma^2 = \Theta(1)$*, under Assumptions 1, 2 and 3, with probability of at least* $1 - n^{-q}$ *over sample inputs* $(x_i)_{i=1}^n$*, where* $0 \leq q < \min\{\frac{(2\beta-1)(\alpha-1-2\tau)}{4\alpha^2}, \frac{\alpha-1-2\tau}{2\alpha}\}$*, the expected normalized SC* (13) *has the asymptotic behavior:*

$$\begin{aligned}\mathbb{E}_\epsilon F^0(D_n) &= \big[\tfrac{1}{2}\log\det(I + \tfrac{n}{\sigma^2}\Lambda) - \tfrac{1}{2}\operatorname{Tr}\big(I - (I + \tfrac{n}{\sigma^2}\Lambda)^{-1}\big) + \tfrac{n}{2\sigma^2}\boldsymbol{\mu}^T(I + \tfrac{n}{\sigma^2}\Lambda)^{-1}\boldsymbol{\mu}\big](1 + o(1)) \\ &= \Theta(n^{\max\{\frac{1}{\alpha}, \frac{1-2\beta}{\alpha}+1\}}).\end{aligned} \tag{8}$$

In the case that $\mu_0 > 0$*, with probability of at least* $1 - n^{-q}$ *over sample inputs* $(x_i)_{i=1}^n$ *where* $0 \leq q < \min\{\frac{2\beta-1}{2}, \alpha\} \cdot \min\{\frac{\alpha-1-2\tau}{2\alpha^2}, \frac{2\beta-1}{\alpha^2}\}$*, the expected normalized SC* (13) *has the asymptotic behavior:* $\mathbb{E}_\epsilon F^0(D_n) = \frac{1}{2\sigma^2}\mu_0^2 n + o(n)$.

Theorem 5 (Asymptotics of the Bayesian generalization error)**.** *Let Assumptions 1, 2, and 3 hold. In the case that* $\mu_0 = 0$ *and* $\sigma^2_{\text{model}} = \sigma^2_{\text{true}} = \sigma^2 = \Theta(n^t)$ *where* $1 - \frac{\alpha}{1+2\tau} < t < 1$*, with probability of at least* $1 - n^{-q}$ *over sample inputs* $(x_i)_{i=1}^n$ *where* $0 \leq q < \frac{[\alpha-(1+2\tau)(1-t)](2\beta-1)}{4\alpha^2}$*, the expectation of the Bayesian generalization error* (1) *w.r.t. the noise* $\boldsymbol{\epsilon}$ *has the asymptotic behavior:*

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\epsilon}} G(D_n) &= \tfrac{1+o(1)}{2\sigma^2}\Big(\operatorname{Tr}(I + \tfrac{n}{\sigma^2}\Lambda)^{-1}\Lambda - \|\Lambda^{1/2}(I + \tfrac{n}{\sigma^2}\Lambda)^{-1}\|_F^2 + \|(I + \tfrac{n}{\sigma^2}\Lambda)^{-1}\boldsymbol{\mu}\|_2^2\Big) \\ &= \tfrac{1}{\sigma^2}\Theta(n^{\max\{\frac{(1-\alpha)(1-t)}{\alpha}, \frac{(1-2\beta)(1-t)}{\alpha}\}}).\end{aligned} \tag{9}$$

In the case that $\mu_0 > 0$ *and under the same conditions, the expectation of the Bayesian generalization error* (1) *w.r.t. the noise* $\boldsymbol{\epsilon}$ *has the asymptotic behavior:* $\mathbb{E}_{\boldsymbol{\epsilon}} G(D_n) = \frac{1}{2\sigma^2}\mu_0^2 + o(1)$.

When $\mu_0 = 0$, the target f lies in the span of eigenfunctions with positive eigenvalues. When $\mu_0 > 0$, the generalization error remains constant when $n \to \infty$ and GPR is not able to learn f. Intuitively, in (9) the exponents $\frac{(1-\alpha)(1-t)}{\alpha}$ resp. $\frac{(1-2\beta)(1-t)}{\alpha}$ capture the rates at which the model suppresses the noise resp. learns the target function. So, larger the value of β, the easier we learn the target.

Remark 6. *For* $f \sim \mathcal{GP}(0, k)$*, using the Karhunen-Loève expansion,* $f(x) = \sum_{p=1}^\infty \sqrt{\lambda_p}\omega_p\phi_p(x)$*, where* $\omega_p \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$*. Comparing with* (4)*, 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 matches the rate obtained by [37] for the expectation of the excess mean square error w.r.t. the distribution of the target when the model is correctly specified.*

3 Experiments

We illustrate our theory on a few toy experiments. Let the input x be uniformly distributed on a unit circle, i.e., $\Omega = S^1$ and $\rho = \mathcal{U}(S^1)$, and write $x = (\cos\theta, \sin\theta)$ 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. Then Assumption 1 is satisfied with $\alpha = 4$. We consider the target functions in Table 1, which satisfy Assumption 2 with the indicated β and μ_0. For each target we conduct GPR 20 times and report the mean and standard deviation of the normalized SC and the Bayesian generalization error in Figure 1, which agree with the asymptotics predicted in Theorems 4 and 5. In Appendix F, we show more details on the experiments and more experiments confirming our theory for zero- and second- order arc-cosine kernels, with and without biases.

4 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

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

Table 1: Target functions used in the experiments for the first order arc-cosine kernel without bias $k^{(1)}_{\text{w/o bias}}$, their values of β and μ_0, and theoretical rates for the normalized SC and the Bayesian generalization error from our theorems.

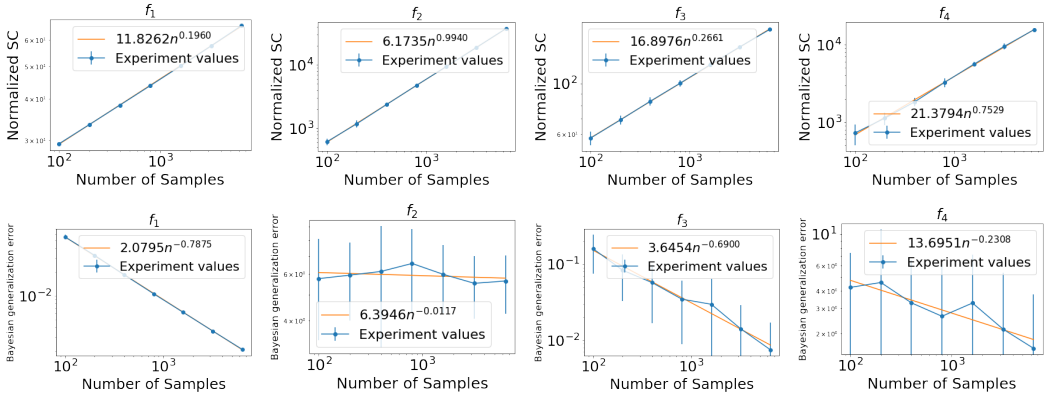

Figure 1: Normalized SC (top) and Bayesian generalization error (bottom) for GPR with the kernel $k^{(1)}_{\text{w/o bias}}$ and the target functions in Table 1. The orange curves show the linear regression fit for the experimental values (in blue) of the log Bayesian generalization error as a function of log n.

of empirical eigenfunctions $\Phi^T\Phi$ around nI. We showed that when $\beta \geq \alpha/2$, meaning 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. In future work, it would be interesting to estimate α and β for the NNGP kernel and the NTK 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.

Acknowledgment

This project has received funding from the European Research Council (ERC) under the EU's Horizon 2020 research and innovation programme (grant agreement n° 757983).

References

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

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

[3] 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*, volume 32, pages 8139–8148, 2019.

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] Y. Cho and L. K. Saul. Kernel methods for deep learning. In *Advances in Neural Information Processing Systems*, volume 22, pages 342–350, 2009.

[12] 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.

[13] 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*, volume 29, pages 2253–2261, 2016.

[14] 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.

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

[16] 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.

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

[18] 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.

[19] A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, pages 8571–8580, 2018.

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

[21] 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 *Advances in Neural Information Processing Systems*, volume 33, pages 15156–15172, 2020.

[22] 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.

[23] 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 *Advances in Neural Information Processing Systems*, volume 32, pages 8572–8583, 2019.

[24] 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.

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

[26] D. Malzahn and M. Opper. Learning curves for Gaussian processes regression: A framework for good approximations. In *Advances in Neural Information Processing Systems*, volume 13, pages 273–279, 2001.

[27] R. M. Neal. *Bayesian Learning for Neural Networks*. Springer-Verlag, Berlin, Heidelberg, 1996.

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

[29] 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.

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

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] P. Sollich. Learning curves for Gaussian processes. In *Advances in Neural Information Processing Systems*, volume 11, pages 344–350, 1999.

[36] P. Sollich. Gaussian process regression with mismatched models. In *Advances in Neural Information Processing Systems*, volume 13, pages 519–526, 2001.

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

[38] 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.

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

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

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

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

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

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

[45] S. Watanabe. *Algebraic Geometry and Statistical Learning Theory*. Cambridge University Press, 2009.

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

[47] C. K. Williams. Computing with infinite networks. In *Advances in Neural Information Processing Systems*, volume 9, pages 295–301, 1997.

[48] C. K. Williams and C. E. Rasmussen. *Gaussian processes for machine learning*. MIT press, 2006.

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

[50] G. Yang. Wide feedforward or recurrent neural networks of any architecture are gaussian processes. In *Advances in Neural Information Processing Systems*, volume 32, pages 9951–9960, 2019.

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

Appendix

A Related work

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 [18, 38]. Power-law asymptotics of learning curves have been theoretically studied in early works for the Gibbs learning algorithm [1, 2, 16] 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 [9]. In the context of GPs, approximations and bounds on learning curves have been investigated in several works [35, 37, 36, 31, 30, 49, 26, 25, 34, 41, 20], with recent extensions to kernel regression from a spectral bias perspective [8, 10]. For a review on learning curves in relation to its shape, monotonicity and some open problems, see the works [24, 44, 43].

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 [39, 48]. 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 [48]. Often, however, the asymptotic properties are known. An early work [46] 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, [32] 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, [33] 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 [4, 10, 21, 28]. [42] 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 [35, 37, 36], 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 [8, 38] and Gaussian design analysis [12]. A related but complementary line of work studies the convergence rates and posterior consistency properties of Bayesian non-parametric models [5, 34, 41].

B Posterior mean and covariance of GPR

We assume that there is a prior distribution Π_0 on the target function 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 $K_n = (k(x_i, x_j))_{i,j=1}^n \in \mathbb{R}^{n\times n}$.

We assume that the variance of Gaussian noise in GPR is σ^2_{model}. In practice, we do not know the exact value of σ_{true} and so our choice of σ_{model} can be different from σ_{true}. 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

$$\bar{m}(x) = K_{x\mathbf{x}}(K_n + \sigma^2_{\text{model}} I_n)^{-1}\mathbf{y},\ x \in \Omega \tag{10}$$

$$\bar{k}(x, x') = k(x, x') - K_{x\mathbf{x}}(K_n + \sigma^2_{\text{model}} I_n)^{-1} K_{\mathbf{x}x'},\ x, x' \in \Omega, \tag{11}$$

where $K_{x\mathbf{x}} = K^T_{\mathbf{x}x} = (k(x_1, x), \ldots, k(x_n, x))^T$ and $\mathbf{y} = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$ [48].

The distance of the posterior to the ground truth can be measured in various ways. Two such measures are the Bayesian generalization error [34, 17, 31] and the excess mean square error [37, 20, 8, 12]. In this work we focus primarily on the Bayesian generalization error defined in (1).

C The normalized stochastic complexity for GPR

Proposition 7 (Normalized stochastic complexity for GPR)**.** *Assume that $\sigma^2_{\text{model}} = \sigma^2_{\text{true}} = \sigma^2$. The normalized SC $F^0(D_n)$ (2) for GPR with prior $\mathcal{GP}(0, k)$ is given as*

$$F^0(D_n) = \frac{1}{2}\log\det(I + \tfrac{K_n}{\sigma^2}) + \frac{1}{2\sigma^2}\mathbf{y}^T(I + \tfrac{K_n}{\sigma^2})^{-1}\mathbf{y} - \frac{1}{2\sigma^2}(\mathbf{y} - f(\mathbf{x}))^T(\mathbf{y} - f(\mathbf{x})), \tag{12}$$

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

$$\mathbb{E}_{\boldsymbol{\epsilon}} F^0(D_n) = \frac{1}{2}\log\det\big(I + \tfrac{K_n}{\sigma^2}\big) - \frac{1}{2}\mathrm{Tr}\Big(I - \big(I + \tfrac{K_n}{\sigma^2}\big)^{-1}\Big) + \frac{1}{2\sigma^2} f(\mathbf{x})^T\big(I + \tfrac{K_n}{\sigma^2}\big)^{-1} f(\mathbf{x}). \tag{13}$$

This is a basic result and has applications in relation to model selection in GPR [48]. For completeness, we include a proof. [34, 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 [41] limiting the applicability of the result.

Proof of Proposition 7. 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:

$$\begin{aligned} Z_n &= \int_{\mathbb{R}^n} \left(\prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\bar{y}_i - y_i)^2}{2\sigma^2}} \right) \frac{1}{(2\pi)^{n/2} \det(K_n)^{1/2}} e^{-\frac{1}{2}\bar{\mathbf{y}}^T K_n^{-1} \bar{\mathbf{y}}} \mathrm{d}\bar{\mathbf{y}} \\ &= \frac{1}{(2\pi)^n \sigma^n \det(K_n)^{1/2}} \int_{\mathbb{R}^n} e^{-\frac{1}{2}\bar{\mathbf{y}}^T (K_n^{-1} + \frac{1}{\sigma^2} I)\bar{\mathbf{y}} + \frac{1}{\sigma^2}\bar{\mathbf{y}}^T \mathbf{y} - \frac{1}{2\sigma^2}\mathbf{y}^T \mathbf{y}} \mathrm{d}\bar{\mathbf{y}}. \end{aligned} \tag{14}$$

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

$$\begin{aligned} Z_n &= \frac{1}{(2\pi)^n \sigma^n \det(K_n)^{1/2}} \int_{\mathbb{R}^n} e^{-\frac{1}{2}(\bar{\mathbf{y}} - \mu)^T \tilde{K}_n^{-1}(\bar{\mathbf{y}} - \mu) - \frac{1}{2\sigma^2}\mathbf{y}^T \mathbf{y} + \frac{1}{2}\mu^T \tilde{K}_n^{-1} \mu} \mathrm{d}\bar{\mathbf{y}} \\ &= \frac{1}{(2\pi)^n \sigma^n \det(K_n)^{1/2}} (2\pi)^{n/2} \det(\tilde{K}_n)^{1/2} e^{-\frac{1}{2\sigma^2}\mathbf{y}^T \mathbf{y} + \frac{1}{2}\mu^T \tilde{K}_n^{-1} \mu} \\ &= \frac{\det(\tilde{K}_n)^{1/2}}{(2\pi)^{n/2} \sigma^n \det(K_n)^{1/2}} e^{-\frac{1}{2\sigma^2}\mathbf{y}^T \mathbf{y} + \frac{1}{2}\mu^T \tilde{K}_n^{-1} \mu}. \end{aligned} \tag{15}$$

The normalized evidence is

$$\begin{aligned} Z_n^0 &= \frac{Z_n}{(2\pi)^{-n/2} \sigma^{-n} e^{-\frac{1}{2\sigma^2}(\mathbf{y} - f(\mathbf{x}))^T (\mathbf{y} - f(\mathbf{x}))}} \\ &= \frac{\det(\tilde{K}_n)^{1/2}}{\det(K_n)^{1/2}} e^{-\frac{1}{2\sigma^2}\mathbf{y}^T \mathbf{y} + \frac{1}{2}\mu^T \tilde{K}_n^{-1} \mu + \frac{1}{2\sigma^2}(\mathbf{y} - f(\mathbf{x}))^T (\mathbf{y} - f(\mathbf{x}))}. \end{aligned} \tag{16}$$

So the normalized stochastic complexity is

$$\begin{aligned} F_n^0 &= -\log Z_n^0 \\ &= -\frac{1}{2}\log\det(\tilde{K}_n)^{1/2} + \frac{1}{2}\log\det(K_n)^{1/2} + \frac{1}{2\sigma^2}\mathbf{y}^T \mathbf{y} - \frac{1}{2}\mu^T \tilde{K}_n^{-1} \mu - \frac{1}{2\sigma^2}(\mathbf{y} - f(\mathbf{x}))^T (\mathbf{y} - f(\mathbf{x})) \\ &= -\frac{1}{2}\log\det(K_n^{-1} + \frac{1}{\sigma^2} I)^{-1} + \frac{1}{2}\log\det(K_n) + \frac{1}{2\sigma^2}\mathbf{y}^T \mathbf{y} - \frac{1}{2\sigma^4}\mathbf{y}^T (K_n^{-1} + \frac{1}{\sigma^2} I)^{-1} \mathbf{y} \\ &\quad - \frac{1}{2\sigma^2}(\mathbf{y} - f(\mathbf{x}))^T (\mathbf{y} - f(\mathbf{x})) \\ &= \frac{1}{2}\log\det(I + \frac{K_n}{\sigma^2}) + \frac{1}{2\sigma^2}\mathbf{y}^T (I + \frac{K_n}{\sigma^2})^{-1} \mathbf{y} - \frac{1}{2\sigma^2}(\mathbf{y} - f(\mathbf{x}))^T (\mathbf{y} - f(\mathbf{x})). \\ &= \frac{1}{2}\log\det(I + \frac{K_n}{\sigma^2}) + \frac{1}{2\sigma^2} f(\mathbf{x})^T (I + \frac{K_n}{\sigma^2})^{-1} f(\mathbf{x}) + \frac{1}{2\sigma^2}\boldsymbol{\epsilon}^T (I + \frac{K_n}{\sigma^2})^{-1} \boldsymbol{\epsilon} - \frac{1}{2\sigma^2}\boldsymbol{\epsilon}^T \boldsymbol{\epsilon} \\ &\quad + \frac{1}{2\sigma^2}\boldsymbol{\epsilon}^T (I + \frac{K_n}{\sigma^2})^{-1} f(\mathbf{x}). \end{aligned} \tag{17}$$

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

$$\mathbb{E}_{\boldsymbol{\epsilon}} F_n^0 = \frac{1}{2}\log\det(I + \frac{K_n}{\sigma^2}) + \frac{1}{2\sigma^2} f(\mathbf{x})^T (I + \frac{K_n}{\sigma^2})^{-1} f(\mathbf{x}) - \frac{1}{2}\mathrm{Tr}(I - (I + \frac{K_n}{\sigma^2})^{-1}). \tag{18}$$

This concludes the proof. □

D Proof of the main results

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 $f(\mathbf{x}) = \Phi\mu$.

D.1 Proofs related to the asymptotics of the normalized stochastic complexity

In this section we derive the asymptotics of the normalized SC (13). Define

$$T_1(D_n) = \tfrac{1}{2}\log\det\Big(I_n + \tfrac{\Phi\Lambda\Phi^T}{\sigma^2}\Big) - \tfrac{1}{2}\mathrm{Tr}\Big(I_n - \Big(I_n + \tfrac{\Phi\Lambda\Phi^T}{\sigma^2}\Big)^{-1}\Big), \tag{19}$$

$$T_2(D_n) = \tfrac{1}{2\sigma^2} f(\mathbf{x})^T \Big(I_n + \tfrac{\Phi\Lambda\Phi^T}{\sigma^2}\Big)^{-1} f(\mathbf{x}), \tag{20}$$

$$\tag{21}$$

Using (13) and (3), we have $\mathbb{E}_{\boldsymbol{\epsilon}} F^0(D_n) = T_1(D_n) + T_2(D_n)$. We first give out the proof sketch of Theorem 4 for case that $\mu_0 = 0$.

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 residual errors. Given a truncation parameter $R \in \mathbb{N}$, let $\Phi_R = (\phi_0(\mathbf{x}), \phi_1(\mathbf{x}), \ldots, \phi_R(\mathbf{x})) \in \mathbb{R}^{n\times R}$ be the truncated matrix of eigenfunctions evaluated at the data points, $\Lambda_R = \mathrm{diag}(0, \lambda_1, \ldots, \lambda_R) \in \mathbb{R}^{(R+1)\times(R+1)}$ and $\boldsymbol{\mu}_R = (\mu_0, \mu_1, \ldots, \mu_R) \in \mathbb{R}^{R+1}$. We define the truncated version of $T_1(D_n)$ as follows:

$$T_{1,R}(D_n) = \tfrac{1}{2}\log\det\Big(I_n + \tfrac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2}\Big) - \tfrac{1}{2}\mathrm{Tr}\Big(I_n - (I_n + \tfrac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}\Big). \tag{22}$$

Similarly, define $\Phi_{>R} = (\phi_{R+1}(\mathbf{x}), \phi_{R+2}(\mathbf{x}), \ldots, \phi_p(\mathbf{x}), \ldots)$, $\Lambda_{>R} = \mathrm{diag}(\lambda_{R+1}, \ldots, \lambda_p, \ldots)$, $f_R(x) = \sum_{p=1}^R \mu_p\phi_p(x)$, $f_R(\mathbf{x}) = (f_R(x_1), \ldots, f_R(x_n))^T$, $f_{>R}(x) = f(x) - f_R(x)$, 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

$$T_{2,R}(D_n) = \tfrac{1}{2\sigma^2} f_R(\mathbf{x})^T (I_n + \tfrac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1} f_R(\mathbf{x})^T. \tag{23}$$

The proof consists of three steps:

- **Approximation step:** In this step, we show that the asymptotics of $T_{1,R}$ resp. $T_{2,R}$ dominates that of the residuals, $|T_{1,R}(D_n) - T_1(D_n)|$ resp. $|T_{2,R}(D_n) - T_2(D_n)|$ (see Lemma 8). This builds upon first showing that $\|\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T\|_2 = \tilde{O}(\max\{nR^{-\alpha}, n^{\frac{1}{2}}R^{\frac{1-2\alpha}{2}}, R^{1-\alpha}\})$ (see Lemma 29) and then choosing $R = n^{\frac{1}{\alpha}+\kappa}$ where $0 < \kappa < \frac{\alpha-1-2\tau}{2\alpha^2}$ when we have $\|\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T\|_2 = o(1)$. Intuitively, the choice of the truncation parameter R is governed by the fact that $\lambda_R = \Theta(R^{-\alpha}) = n^{-1+\kappa\alpha} = o(n^{-1})$.
- **Decomposition step:** In this step, we decompose $T_{1,R}$ into a term independent of Φ_R and a series involving $\Phi_R^T\Phi_R - nI_R$, and likewise for $T_{2,R}$ (see Lemma 10). This builds upon first showing using the Woodbury matrix identity [48, §A.3] that

  $$T_{1,R}(D_n) = \tfrac{1}{2}\log\det(I_R + \tfrac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R) - \tfrac{1}{2}\mathrm{Tr}\Phi_R(\sigma^2 I_R + \Lambda_R\Phi_R^T\Phi_R)^{-1}\Lambda_R\Phi_R^T, \tag{24}$$

  $$T_{2,R}(D_n) = \tfrac{1}{2\sigma^2}\boldsymbol{\mu}_R^T\Phi_R^T\Phi_R(\sigma^2 I_R + \Lambda_R\Phi_R^T\Phi_R)^{-1}\boldsymbol{\mu}_R, \tag{25}$$

  and then Taylor expanding the matrix inverse $(\sigma^2 I_R + \Lambda_R\Phi_R^T\Phi_R)^{-1}$ in (24) and (25) to show that the Φ_R-independent terms in the decomposition of $T_{1,R}$ and $T_{2,R}$ are, respectively, $\frac{1}{2}\log\det(I_R + \frac{n}{\sigma^2}\Lambda_R) - \frac{1}{2}\mathrm{Tr}\big(I_R - (I_R + \frac{n}{\sigma^2}\Lambda_R)^{-1}\big)$, and $\frac{n}{2\sigma^2}\boldsymbol{\mu}_R^T(I_R + \frac{n}{\sigma^2}\Lambda_R)^{-1}\boldsymbol{\mu}_R$.
- **Concentration step:** Finally, we use concentration inequalities to show that these Φ_R-independent terms dominate the series involving $\Phi_R^T\Phi_R - nI_R$ (see Lemma 11) when we have

  $$T_{1,R}(D_n) = \big(\tfrac{1}{2}\log\det(I_R + \tfrac{n}{\sigma^2}\Lambda_R) - \tfrac{1}{2}\mathrm{Tr}\big(I_R - (I_R + \tfrac{n}{\sigma^2}\Lambda_R)^{-1}\big)\big)(1 + o(1)) = \Theta(n^{\frac{1}{\alpha}}),$$

  $$T_{2,R}(D_n) = \big(\tfrac{n}{2\sigma^2}\boldsymbol{\mu}_R^T(I_R + \tfrac{n}{\sigma^2}\Lambda_R)^{-1}\boldsymbol{\mu}_R\big)(1 + o(1)) = \begin{cases}\Theta(n^{\max\{0, \frac{1-2\beta}{\alpha}+1\}}), & \alpha \neq 2\beta - 1,\\ \Theta(\log n), & \alpha = 2\beta - 1.\end{cases}$$

  The key idea is to consider the matrix $\Lambda_R^{1/2}(I + \frac{n}{\sigma^2}\Lambda_R)^{-1/2}\Phi_R^T\Phi_R(I + \frac{n}{\sigma^2}\Lambda_R)^{-1/2}\Lambda_R^{1/2}$ and show that it concentrates around $n\Lambda_R(I + \frac{n}{\sigma^2})^{-1}$ (see Corollary 26). Note that an ordinary application of the matrix Bernstein inequality to $\Phi_R^T\Phi_R - nI_R$ yields $\|\Phi_R^T\Phi_R - nI\|_2 = O(R\sqrt{n})$, which is not sufficient for our purposes, since this would give $O(R\sqrt{n}) = o(n)$ only when $\alpha > 2$. In contrast, our results are valid for $\alpha > 1$ and cover cases of practical interest, e.g., the NTK of infinitely wide shallow ReLU network [42] and the arc-cosine kernels over high-dimensional hyperspheres [33] that have $\alpha = 1 + O(\frac{1}{d})$, where d is the input dimension.

Next, we flesh out the complete proof below.

Lemma 8. *Under Assumptions 1, 2 and 3, with probability of at least* $1-2\delta$ *we have, we have*

$$|T_{1,R}(D_n) - T_1(D_n)| = \tilde{O}\Big(\tfrac{1}{\sigma^2}(nR^{1-\alpha} + n^{1/2}R^{1-\alpha+\tau} + R^{1-\alpha+2\tau})\Big) \tag{26}$$

If $R = n^{\frac{1}{\alpha}+\kappa}$ *where* $\kappa > 0$*, we have* $|T_{1,R}(D_n) - T_1(D_n)| = o\Big(\tfrac{1}{\sigma^2} n^{\frac{1}{\alpha}}\Big)$*. If we further assume that* $0 < \kappa < \frac{\alpha-1-2\tau}{\alpha^2}$*,* $\mu_0 = 0$ *and* $\sigma^2 = \Theta(1)$*, then for sufficiently large* n *with probability of at least* $1-4\delta$ *we have*

$$|T_{2,R}(D_n) - T_2(D_n)| = \tilde{O}\Big((\tfrac{1}{\delta}+1)n^{\max\{(\frac{1}{\alpha}+\kappa)\frac{1-2\beta}{2}, 1+\frac{1-2\beta}{\alpha}+\frac{(1-2\beta)\kappa}{2}, -1-\kappa\alpha, 1+\frac{1-2\beta}{\alpha}-\kappa\alpha\}}\Big). \tag{27}$$

Proof of Lemma 8. Define $\Phi_{>R} = (\phi_{R+1}(\mathbf{x}), \phi_{R+2}(\mathbf{x}), \ldots, \phi_p(\mathbf{x}), \ldots)$, and $\Lambda_{>R} = \operatorname{diag}(\lambda_{R+1}, \ldots, \lambda_p, \ldots)$. We then have

$$\begin{aligned} |T_1(D_n) - T_{1,R}(D_n)| = &\left|\frac{1}{2}\log\det(I + \frac{1}{\sigma^2}\Phi\Lambda\Phi^T) - \frac{1}{2}\log\det(I + \frac{1}{\sigma^2}\Phi_R\Lambda_R\Phi_R^T)\right| \\ &+ \frac{1}{2}\left|\operatorname{Tr}(I + \frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1} - \operatorname{Tr}(I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}\right|. \end{aligned} \tag{28}$$

As for the first term in the right hand side of (28), we have

$$\begin{aligned} &\left|\frac{1}{2}\log\det(I + \frac{1}{\sigma^2}\Phi\Lambda\Phi^T) - \frac{1}{2}\log\det(I + \frac{1}{\sigma^2}\Phi_R\Lambda_R\Phi_R^T)\right| \\ =&\left|\frac{1}{2}\log\det\left((I + \frac{1}{\sigma^2}\Phi_R\Lambda_R\Phi_R^T)^{-1}(I + \frac{1}{\sigma^2}\Phi_R\Lambda_R\Phi_R^T + \frac{1}{\sigma^2}\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T)\right)\right| \\ =&\left|\frac{1}{2}\log\det\left(I + \frac{1}{\sigma^2}(I + \frac{1}{\sigma^2}\Phi_R\Lambda_R\Phi_R^T)^{-1}\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T\right)\right| \\ =&\frac{1}{2}\left|\operatorname{Tr}\log\left(I + \frac{1}{\sigma^2}(I + \frac{1}{\sigma^2}\Phi_R\Lambda_R\Phi_R^T)^{-1}\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T\right)\right|. \end{aligned} \tag{29}$$

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

$$\operatorname{Tr} h(B) = \textstyle\sum_{p=1}^n h(\zeta_i) \le nh(\frac{1}{n}\sum_{p=1}^n \zeta_i) \le nh(\frac{1}{n}\operatorname{Tr} B), \tag{30}$$

where we used Jensen's inequality. Using $h(x) = \log(1+x)$ in (30), with probability $1-\delta$, we have

$$\begin{aligned} &\left|\tfrac{1}{2}\log\det(I + \tfrac{1}{\sigma^2}\Phi\Lambda\Phi^T) - \tfrac{1}{2}\log\det(I + \tfrac{1}{\sigma^2}\Phi_R\Lambda_R\Phi_R^T)\right| \\ &\le \tfrac{n}{2}\log(1 + \tfrac{1}{n}\operatorname{Tr}(\tfrac{1}{\sigma^2}(I + \tfrac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T)) \\ &\le \tfrac{n}{2}\log(1 + \tfrac{1}{n\sigma^2}\|(I + \tfrac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}\|_2 \operatorname{Tr}(\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T)) \\ &\le \tfrac{n}{2}\log(1 + \tfrac{1}{n\sigma^2}\textstyle\sum_{p=R+1}^\infty \lambda_p\|\phi_p(\mathbf{x})\|_2^2) \le \tfrac{1}{2\sigma^2}\sum_{p=R+1}^\infty \lambda_p\|\phi_p(\mathbf{x})\|_2^2 \\ &= \tfrac{1}{2\sigma^2}\textstyle\sum_{p=R+1}^\infty \lambda_p\Big(C_\phi^2\tilde{O}\Big(\sqrt{p^{2\tau}n\|\phi_p\|_2^2} + p^{2\tau}\Big) + n\|\phi_p\|_2^2\Big) \\ &= \tilde{O}(\tfrac{1}{\sigma^2}n\textstyle\sum_{p=R+1}^\infty \lambda_p + n^{1/2}\sum_{p=R+1}^\infty \lambda_p p^\tau + \sum_{p=R+1}^\infty \lambda_p p^{2\tau}) \\ &= \tilde{O}\Big(\tfrac{1}{\sigma^2}(nR^{1-\alpha} + n^{1/2}R^{1-\alpha+\tau} + R^{1-\alpha+2\tau})\Big) = o\Big(\tfrac{1}{\sigma^2}n^{\frac{1}{\alpha}}\Big), \end{aligned} \tag{31}$$

where in the second inequality we use the fact that $\operatorname{Tr} AB \le \|A\|_2 \operatorname{Tr} B$ when A and B are symmetric positive definite matrices, and in the last inequality we use Lemma 22.

As for the second term in the right hand side of (28), let $A = (I + \frac{\Phi_R \Lambda_R \Phi_R^T}{\sigma^2})^{-1/2}$. Then we have

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

where in the first inequality we use the fact that $\|A\|_2 < 1$ and $\operatorname{Tr} ABA \le \|A\|_2^2 \operatorname{Tr} B$ when A and B are symmetric positive definite matrices, in the second inequality we use $h(x) = 1 - 1/(1+x)$ in (30) and in the last equality we use the last few steps of (31). This concludes the proof of the first statement.

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

$$
\begin{aligned}
|T_2(D_n) - T_{2,R}(D_n)| = &\Big|f(\mathbf{x})^T(I + \tfrac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f(\mathbf{x}) - f_R(\mathbf{x})^T(I + \tfrac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f_R(\mathbf{x})\Big| \\
&+ \Big|f_R(\mathbf{x})^T(I + \tfrac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f_R(\mathbf{x}) - f_R(\mathbf{x})^T(I + \tfrac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}f_R(\mathbf{x})\Big|.
\end{aligned} \tag{32}
$$

For the first term on the right-hand side of (32), we have

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

Applying Corollary 23 and Lemma 35, with probability of at least $1 - 4\delta$, we have

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

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

As for the second term on the right-hand side of (32), according to Lemma 32, Corollary 30 and Lemma 33, we have

$$\begin{aligned}
&\left|f_R(\mathbf{x})^T(I+\tfrac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f_R(\mathbf{x})-f_R(\mathbf{x})^T(I+\tfrac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}f_R(\mathbf{x})\right| \\
=&\left|\sum_{j=1}^{\infty}(-1)^j f_R(\mathbf{x})^T\Big((I+\tfrac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}\tfrac{\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T}{\sigma^2}\Big)^j(I+\tfrac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}f_R(\mathbf{x})\right| \\
\leq&\sum_{j=1}^{\infty}\|(I+\tfrac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}\|_2^{j-1}\cdot\|\frac{\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T}{\sigma^2}\|_2^j\cdot\|(I+\tfrac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}f_R(\mathbf{x})\|_2^2 \qquad (33)\\
=&\sum_{j=1}^{\infty}\tilde{O}(n^{-j\kappa\alpha})\tilde{O}((\tfrac{1}{\delta}+1)n^{1+\max\{-2,\frac{1-2\beta}{\alpha}\}}) \\
=&\tilde{O}((\tfrac{1}{\delta}+1)n^{1+\max\{-2,\frac{1-2\beta}{\alpha}\}-\kappa\alpha}).
\end{aligned}$$

By (32), we have

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

This concludes the proof of the second statement. □

In Lemma 8, we gave a bound for $|T_{2,R}(D_n)-T_2(D_n)|$ when $n^{\frac{1}{\alpha}}<R<n^{\frac{1}{\alpha}+\frac{\alpha-1-2\tau}{\alpha^2}}$. For $R>n$, we note the following lemma:

Lemma 9. *Let $R=n^C$ and $\sigma^2=n^t$. Assume that $C\geq 1$ and $C(1-\alpha+2\tau)-t<0$. Under Assumptions 1, 2 and 3, for sufficiently large n and with probability of at least $1-3\delta$ we have*

$$|T_{2,R}(D_n)-T_2(D_n)|=\tilde{O}\Big((\tfrac{1}{\delta}+1)\tfrac{1}{\sigma^2}nR^{\max\{1/2-\beta,1-\alpha+2\tau\}}\Big). \qquad (34)$$

Proof of Lemma 9. Define $\Phi_{>R}=(\phi_{R+1}(\mathbf{x}),\phi_{R+2}(\mathbf{x}),\ldots,\phi_p(\mathbf{x}),\ldots)$, and $\Lambda_{>R}=\mathrm{diag}(\lambda_{R+1},\ldots,\lambda_p,\ldots)$. Then we have

$$\begin{aligned}
|T_2(D_n)-T_{2,R}(D_n)|=&\left|f(\mathbf{x})^T(I+\frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f(\mathbf{x})-f_R(\mathbf{x})^T(I+\frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f_R(\mathbf{x})\right| \\
&+\left|f_R(\mathbf{x})^T(I+\frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f_R(\mathbf{x})-f_R(\mathbf{x})^T(I+\frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}f_R(\mathbf{x})\right|. \qquad (35)
\end{aligned}$$

For the first term on the right-hand side of (35), with probability $1-3\delta$ we have

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

where we used Corollary 23 and Lemma 21 for the last inequality.

The assumption $C(1-\alpha+2\tau)-t<0$ means that $\frac{R^{1-\alpha+2\tau}}{\sigma^2}=o(1)$. For the second term on the right-hand side of (35), by Lemmas 32 and 29, we have

$$\begin{aligned}
&\left|f_R(\mathbf{x})^T(I+\frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f_R(\mathbf{x})-f_R(\mathbf{x})^T(I+\frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}f_R(\mathbf{x})\right| \\
=&\left|\sum_{j=1}^{\infty}(-1)^j f_R(\mathbf{x})^T\left((I+\frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}\frac{\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T}{\sigma^2}\right)^j(I+\frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}f_R(\mathbf{x})\right| \\
\leq&\sum_{j=1}^{\infty}\|(I+\frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}\|_2^{j+1}\cdot\|\frac{\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T}{\sigma^2}\|_2^j\cdot\|f_R(\mathbf{x})\|_2^2 \\
=&\sum_{j=1}^{\infty}\tilde{O}(\frac{1}{\sigma^2}R^{j(1-\alpha+2\tau)})\tilde{O}((\frac{1}{\delta}+1)n\|f\|_2^2) \\
=&\tilde{O}((\frac{1}{\delta}+1)\frac{1}{\sigma^2}nR^{1-\alpha+2\tau}).
\end{aligned} \tag{36}$$

Using (35), we have

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

□

Next we consider the asympototics of $T_{1,R}(D_n)$ and $T_{2,R}(D_n)$.

Lemma 10. *Let* $A=(I+\frac{n}{\sigma^2}\Lambda_R)^{-\gamma/2}\Lambda_R^{\gamma/2}(\Phi_R^T\Phi_R-nI)\Lambda_R^{\gamma/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-\gamma/2}$. *Assume that* $\|A\|_2<1$ *where* $\frac{1+2\tau}{\alpha}<\gamma\leq 1$. *Then we have*

$$T_{2,R}(D_n)=\tfrac{n}{2\sigma^2}\boldsymbol{\mu}_R^T(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-1}\boldsymbol{\mu}_R+\tfrac{1}{2}\textstyle\sum_{j=1}^{\infty}(-1)^{j+1}E_j,$$

where

$$E_j=\boldsymbol{\mu}_R^T\tfrac{1}{\sigma^2}(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-1}(\Phi_R^T\Phi_R-nI)\big(\tfrac{1}{\sigma^2}(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R(\Phi_R^T\Phi_R-nI)\big)^{j-1}(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-1}\boldsymbol{\mu}_R.$$

Proof of Lemma 10. 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

$$\begin{aligned}
&\frac{1}{2\sigma^2}\boldsymbol{\mu}_R^T\Phi_R^T(I+\frac{1}{\sigma^2}\Phi_R\tilde{\Lambda}_{\epsilon,R}\Phi_R^T)^{-1}\Phi_R\boldsymbol{\mu}_R \\
&=\frac{1}{2\sigma^2}\boldsymbol{\mu}_R^T\Phi_R^T\Big[I-\Phi_R(\sigma^2I+\tilde{\Lambda}_{\epsilon,R}\Phi_R^T\Phi_R)^{-1}\tilde{\Lambda}_{\epsilon,R}\Phi_R^T\Big]\Phi_R\boldsymbol{\mu}_R \\
&=\frac{1}{2\sigma^2}\boldsymbol{\mu}_R^T\Phi_R^T\Phi_R\boldsymbol{\mu}_R-\frac{1}{2\sigma^2}\boldsymbol{\mu}_R^T\Phi_R^T\Phi_R(\sigma^2I+\tilde{\Lambda}_{\epsilon,R}\Phi_R^T\Phi_R)^{-1}\tilde{\Lambda}_{\epsilon,R}\Phi_R^T\Phi_R\boldsymbol{\mu}_R \\
&=\frac{1}{2}\boldsymbol{\mu}_R^T\Phi_R^T\Phi_R(\sigma^2I+\tilde{\Lambda}_{\epsilon,R}\Phi_R^T\Phi_R)^{-1}\boldsymbol{\mu}_R \\
&=\frac{1}{2}\boldsymbol{\mu}_R^T\tilde{\Lambda}_{\epsilon,R}^{-1}\tilde{\Lambda}_{\epsilon,R}\Phi_R^T\Phi_R(\sigma^2I+\tilde{\Lambda}_{\epsilon,R}\Phi_R^T\Phi_R)^{-1}\boldsymbol{\mu}_R \\
&=\frac{1}{2}\boldsymbol{\mu}_R^T\tilde{\Lambda}_{\epsilon,R}^{-1}\boldsymbol{\mu}_R-\frac{1}{2}\boldsymbol{\mu}_R^T\tilde{\Lambda}_{\epsilon,R}^{-1}(I+\frac{1}{\sigma^2}\tilde{\Lambda}_{\epsilon,R}\Phi_R^T\Phi_R)^{-1}\boldsymbol{\mu}_R.
\end{aligned} \tag{37}$$

Using Lemma 31, we have

$$\begin{aligned}
&\frac{1}{2}\boldsymbol{\mu}_R^T\tilde{\Lambda}_{\epsilon,R}^{-1}\boldsymbol{\mu}_R - \frac{1}{2}\boldsymbol{\mu}_R^T\tilde{\Lambda}_{\epsilon,R}^{-1}(I+\frac{1}{\sigma^2}\tilde{\Lambda}_{\epsilon,R}\Phi_R^T\Phi_R)^{-1}\boldsymbol{\mu}_R\\
&=\frac{1}{2}\boldsymbol{\mu}_R^T\tilde{\Lambda}_{\epsilon,R}^{-1}\boldsymbol{\mu}_R - \frac{1}{2}\boldsymbol{\mu}_R^T\tilde{\Lambda}_{\epsilon,R}^{-1}(I+\frac{n}{\sigma^2}\tilde{\Lambda}_{\epsilon,R})^{-1}\boldsymbol{\mu}_R\\
&+\frac{1}{2}\sum_{j=1}^{\infty}(-1)^{j+1}\boldsymbol{\mu}_R^T\tilde{\Lambda}_{\epsilon,R}^{-1}\bigg(\frac{1}{\sigma^2}(I+\frac{n}{\sigma^2}\tilde{\Lambda}_{\epsilon,R})^{-1}\tilde{\Lambda}_{\epsilon,R}(\Phi_R^T\Phi_R-nI)\bigg)^j(I+\frac{n}{\sigma^2}\tilde{\Lambda}_{\epsilon,R})^{-1}\boldsymbol{\mu}_R\\
&=\frac{n}{2\sigma^2}\boldsymbol{\mu}_R^T(I+\frac{n}{\sigma^2}\tilde{\Lambda}_{\epsilon,R})^{-1}\boldsymbol{\mu}_R\\
&+\frac{1}{2}\sum_{j=1}^{\infty}(-1)^{j+1}\boldsymbol{\mu}_R^T\frac{1}{\sigma^2}(I+\frac{n}{\sigma^2}\tilde{\Lambda}_{\epsilon,R})^{-1}(\Phi_R^T\Phi_R-nI)\bigg(\frac{1}{\sigma^2}(I+\frac{n}{\sigma^2}\tilde{\Lambda}_{\epsilon,R})^{-1}\tilde{\Lambda}_{\epsilon,R}(\Phi_R^T\Phi_R-nI)\bigg)^{j-1}\\
&\qquad(I+\frac{n}{\sigma^2}\tilde{\Lambda}_{\epsilon,R})^{-1}\boldsymbol{\mu}_R
\end{aligned}\tag{38}$$

Letting $\epsilon\to 0$, we get

$$\begin{aligned}
T_{2,R}(D_n) &= \frac{1}{2\sigma^2}\boldsymbol{\mu}_R^T\Phi_R^T(I+\frac{1}{\sigma^2}\Phi_R\Lambda_R\Phi_R^T)^{-1}\Phi_R\boldsymbol{\mu}_R\\
&=\frac{n}{2\sigma^2}\boldsymbol{\mu}_R^T(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\boldsymbol{\mu}_R\\
&+\frac{1}{2}\sum_{j=1}^{\infty}\bigg[(-1)^{j+1}\boldsymbol{\mu}_R^T\frac{1}{\sigma^2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}(\Phi_R^T\Phi_R-nI)\bigg(\frac{1}{\sigma^2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R(\Phi_R^T\Phi_R-nI)\bigg)^{j-1}\\
&\qquad(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\boldsymbol{\mu}_R\bigg]
\end{aligned}$$

This concludes the proof. □

Lemma 11. *Assume that $\sigma^2=\Theta(1)$. Let $R=n^{\frac{1}{\alpha}+\kappa}$ where $0<\kappa<\frac{\alpha-1-2\tau}{2\alpha^2}$. Under Assumptions 1, 2 and 3, with probability of at least $1-\delta$, we have*

$$T_{1,R}(D_n)=\big(\tfrac{1}{2}\log\det(I+\tfrac{n}{\sigma^2}\Lambda_R)-\tfrac{1}{2}\operatorname{Tr}\big(I-(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-1}\big)\big)(1+o(1))=\Theta(n^{\frac{1}{\alpha}}).\tag{39}$$

Furthermore, if we assume $\mu_0=0$, we have

$$T_{2,R}(D_n)=\big(\tfrac{n}{2\sigma^2}\boldsymbol{\mu}_R^T(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-1}\boldsymbol{\mu}_R\big)(1+o(1))=\begin{cases}\Theta(n^{\max\{0,1+\frac{1-2\beta}{\alpha}\}}), & \alpha\neq 2\beta-1,\\ \Theta(\log n), & \alpha=2\beta-1.\end{cases}\tag{40}$$

Proof of Lemma 11. Let

$$A=(I+\frac{n}{\sigma^2}\Lambda_R)^{-\gamma/2}\Lambda_R^{\gamma/2}(\Phi_R^T\Phi_R-nI)\Lambda_R^{\gamma/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-\gamma/2},\tag{41}$$

where $\frac{1+\alpha+2\tau}{2\alpha}<\gamma\leq 1$. By Corollary 26, with probability of at least $1-\delta$, we have

$$\|A\|_2=\tilde{O}(n^{\frac{1-2\gamma\alpha+\alpha+2\tau}{2\alpha}}).\tag{42}$$

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 $\|B\|_2=\frac{\sigma^{2(1-\gamma)}}{n^{1-\gamma}}\|A\|_2=\tilde{O}(n^{\frac{1-\alpha+2\tau}{2\alpha}})$. Using the Woodbury

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

$$
\begin{aligned}
T_{1,R}(D_n) &= \tfrac{1}{2}\log\det(I+\tfrac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R) - \tfrac{1}{2}\mathrm{Tr}\Phi_R(\sigma^2 I+\Lambda_R\Phi_R^T\Phi_R)^{-1}\Lambda_R\Phi_R^T \\
&= \tfrac{1}{2}\log\det(I+\tfrac{n}{\sigma^2}\Lambda_R) + \tfrac{1}{2}\log\det[I+\tfrac{1}{\sigma^2}(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-1/2}\Lambda_R^{1/2}(\Phi_R^T\Phi_R-nI)\Lambda_R^{1/2}(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-1/2}] \\
&\quad - \tfrac{1}{2}\mathrm{Tr}(\sigma^2 I+\Lambda\Phi_R^T\Phi_R)^{-1}\Lambda\Phi_R^T\Phi_R \\
&= \tfrac{1}{2}\log\det(I+\tfrac{n}{\sigma^2}\Lambda_R) + \tfrac{1}{2}\,\mathrm{Tr}\log[I+\tfrac{1}{\sigma^2}B] - \tfrac{1}{2}\mathrm{Tr}(I-\sigma^2(\sigma^2 I+\Lambda\Phi_R^T\Phi_R)^{-1})) \\
&= \tfrac{1}{2}\log\det(I+\frac{n}{\sigma^2}\Lambda_R) + \tfrac{1}{2}\,\mathrm{Tr}\sum_{j=1}^{\infty}\tfrac{(-1)^{j-1}}{j}(\tfrac{1}{\sigma^2}B)^j \\
&\quad - \tfrac{1}{2}\mathrm{Tr}\left(I-(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-1}+\sum_{j=1}^{\infty}(-1)^j\big(\tfrac{1}{\sigma^2}(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R(\Phi_R^T\Phi_R-nI)\big)^j(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-1}\right) \\
&= \big(\tfrac{1}{2}\log\det(I+\tfrac{n}{\sigma^2}\Lambda_R) - \tfrac{1}{2}\mathrm{Tr}\big(I-(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-1}\big)\big) + \tfrac{1}{2}\,\mathrm{Tr}\sum_{j=1}^{\infty}\tfrac{(-1)^{j-1}}{j}(\tfrac{1}{\sigma^2}B)^j \\
&\quad - \tfrac{1}{2}\mathrm{Tr}\left(\sum_{j=1}^{\infty}(-1)^j\tfrac{1}{\sigma^{2j}}(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-1/2}B^j(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-1/2}\right),
\end{aligned}
\tag{43}
$$

where in the last equality we apply Lemma 31.

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 (43), we have

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

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

$$
\begin{aligned}
&\tfrac{1}{2}\log\det(I+\tfrac{n}{\sigma^2}\Lambda_R) - \tfrac{1}{2}\,\mathrm{Tr}\big(I-(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-1}\big) \\
=&\tfrac{1}{2}\sum_{p=1}^{R}h(\tfrac{n}{\sigma^2}\lambda_p) \ge \tfrac{1}{2}\sum_{p=1}^{R}h(\tfrac{n}{\sigma^2}\underline{C_\lambda}p^{-\alpha}) \\
\ge&\tfrac{1}{2}\int_{[1,R+1]}h(\tfrac{n}{\sigma^2}\underline{C_\lambda}x^{-\alpha})\mathrm{d}x \\
=&\tfrac{1}{2}n^{1/\alpha}\int_{[1/n^{1/\alpha},(R+1)/n^{1/\alpha}]}h(\tfrac{1}{\sigma^2}\underline{C_\lambda}x^{-\alpha})\mathrm{d}x \\
=&\Theta(n^{1/\alpha}).
\end{aligned}
$$

Overall, we have $\frac{1}{2}\log\det(I+\frac{n}{\sigma^2}\Lambda_R) - \frac{1}{2}\,\mathrm{Tr}\big(I-(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\big)=\Theta(n^{1/\alpha})$.

As for the second term on the right hand side of (43), we have

$$\begin{aligned}\left|\mathrm{Tr}\sum_{j=1}^{\infty}\tfrac{(-1)^{j-1}}{j}(\tfrac{1}{\sigma^2}B)^j\right| &\le R\sum_{j=1}^{\infty}\|\tfrac{1}{\sigma^2}B\|_2^j\\ &= R\sum_{j=1}^{\infty}\tfrac{1}{\sigma^{2j}}\tilde{O}(n^{\frac{j(1-\alpha+2\tau)}{2\alpha}})\\ &= R\tilde{O}(n^{\frac{1-\alpha+2\tau}{2\alpha}}) = \tilde{O}(n^{\frac{1}{\alpha}+\kappa+\frac{1-\alpha+2\tau}{2\alpha}}).\end{aligned}$$

As for the third term on the right hand side of (43), we have

$$\begin{aligned}&\left|\mathrm{Tr}\left(\sum_{j=1}^{\infty}(-1)^j\tfrac{1}{\sigma^{2j}}(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-1/2}B^j(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-1/2}\right)\right|\\ \le&\sum_{j=1}^{\infty}\left|\mathrm{Tr}\Big(\tfrac{1}{\sigma^{2j}}(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-1/2}B^j(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-1/2}\Big)\right|\\ \le&R\sum_{j=1}^{\infty}\left\|\tfrac{1}{\sigma^{2j}}(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-1/2}B^j(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-1/2}\right\|_2\\ \le&R\sum_{j=1}^{\infty}\left\|\tfrac{1}{\sigma^{2j}}(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-1/2}B^j(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-1/2}\right\|_2\\ \le&R\sum_{j=1}^{\infty}\left\|\tfrac{1}{\sigma^{2j}}B^j\right\|_2 = \tilde{O}(n^{\frac{1}{\alpha}+\kappa+\frac{1-\alpha+2\tau}{2\alpha}}).\end{aligned}$$

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

$$\begin{aligned}T_{1,R}(D_n) =&\tfrac{1}{2}\log\det(I+\tfrac{n}{\sigma^2}\Lambda_R) - \tfrac{1}{2}\mathrm{Tr}\big(I-(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-1}\big) + \tilde{O}(n^{\frac{1}{\alpha}+\kappa+\frac{1-\alpha+2\tau}{2\alpha}}) + \tilde{O}(n^{\frac{1}{\alpha}+\kappa+\frac{1-\alpha+2\tau}{2\alpha}})\\ =&\Theta(n^{1/\alpha}) + \tilde{O}(n^{\frac{1}{\alpha}+\kappa+\frac{1-\alpha+2\tau}{2\alpha}})\\ =&\Theta(n^{\frac{1}{\alpha}}),\end{aligned}$$

where in the last inequality we use the assumption that $\kappa < \frac{\alpha-1-2\tau}{2\alpha}$. Since $\tilde{O}(n^{\frac{1}{\alpha}+\kappa+\frac{1-\alpha+2\tau}{2\alpha}})$ is lower order term compared to $\Theta(n^{\frac{1}{\alpha}})$, we further have

$$T_{1,R}(D_n) = \big(\tfrac{1}{2}\log\det(I+\tfrac{n}{\sigma^2}\Lambda_R) - \tfrac{1}{2}\mathrm{Tr}\big(I-(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-1}\big)\big)(1+o(1)).$$

This concludes the proof of the first statement.

Let $\Lambda_{1:R} = \mathrm{diag}\{\lambda_1,\ldots,\lambda_R\}$, $\Phi_{1:R} = (\phi_1(\mathbf{x}),\phi_1(\mathbf{x}),\ldots,\phi_R(\mathbf{x}))$ and $\boldsymbol{\mu}_{1:R} = (\mu_1,\ldots,\mu_R)$. Since $\mu_0 = 0$, we have $T_{2,R}(D_n) = \frac{1}{2\sigma^2}\boldsymbol{\mu}_{1:R}^T\Phi_{1:R}^T(I+\frac{1}{\sigma^2}\Phi_{1:R}\Lambda_{1:R}\Phi_{1:R}^T)^{-1}\Phi_{1:R}\boldsymbol{\mu}_{1:R}$. According to Lemma 10, we have

$$\begin{aligned}T_{2,R}(D_n) =& \frac{n}{2\sigma^2}\boldsymbol{\mu}_{1:R}^T(I+\frac{n}{\sigma^2}\Lambda_{1:R})^{-1}\boldsymbol{\mu}_{1:R}\\ &+\frac{1}{2}\sum_{j=1}^{\infty}(-1)^{j+1}\boldsymbol{\mu}_{1:R}^T\frac{1}{\sigma^2}(I+\frac{n}{\sigma^2}\Lambda_{1:R})^{-1}(\Phi_{1:R}^T\Phi_{1:R}-nI)\left(\frac{1}{\sigma^2}(I+\frac{n}{\sigma^2}\Lambda_{1:R})^{-1}\Lambda_{1:R}(\Phi_{1:R}^T\Phi_{1:R}-nI)\right)^{j-1}\\ =&\frac{n}{2\sigma^2}\boldsymbol{\mu}_{1:R}^T(I+\frac{n}{\sigma^2}\Lambda_{1:R})^{-1}\boldsymbol{\mu}_{1:R}\\ &+\frac{1}{2}\sum_{j=1}^{\infty}\Bigg[(-1)^{j+1}\frac{1}{\sigma^{2j}}\boldsymbol{\mu}_{1:R}^T(I+\frac{n}{\sigma^2}\Lambda_{1:R})^{-1+\gamma/2}\Lambda_{1:R}^{-\gamma/2}A\Big((I+\frac{n}{\sigma^2}\Lambda_{1:R})^{-1+\gamma}\Lambda_{1:R}^{1-\gamma}A\Big)^{j-1}\\ &\qquad(I+\frac{n}{\sigma^2}\Lambda_{1:R})^{-1+\gamma/2}\Lambda_{1:R}^{-\gamma/2}\boldsymbol{\mu}_{1:R}\Bigg]\end{aligned} \tag{44}$$

where in the second to last equality we used the definition of A (41). As for the first term on the right hand side of (44), by Lemma 19, Assumption 1 and Assumption 2, we have

$$\frac{n}{2\sigma^2}\boldsymbol{\mu}_{1:R}^T(I+\frac{n}{\sigma^2}\Lambda_{1:R})^{-1}\boldsymbol{\mu}_{1:R} \le \frac{n}{2\sigma^2}\sum_{p=1}^{R}\frac{C_\mu^2 p^{-2\beta}}{1+\frac{n}{\sigma^2}\underline{C_\lambda}p^{-\alpha}} = \begin{cases}\Theta(n^{\max\{0,1+\frac{1-2\beta}{\alpha}\}}), & \alpha\neq 2\beta-1,\\ \Theta(\log n), & \alpha=2\beta-1.\end{cases}$$

On the other hand, by Assumption 2, assuming that $\sup_{i\ge 1} p_{i+1}-p_i = h$, we have

$$\begin{aligned}\frac{n}{2\sigma^2}\boldsymbol{\mu}_{1:R}^T(I+\frac{n}{\sigma^2}\Lambda_{1:R})^{-1}\boldsymbol{\mu}_{1:R} &\ge \frac{n}{2\sigma^2}\sum_{i=1}^{\lfloor\frac{R}{h}\rfloor}\frac{\underline{C_\mu^2}p_i^{-2\beta}}{1+\frac{n}{\sigma^2}\overline{C_\lambda}p_i^{-\alpha}}\\ &\ge \frac{n}{2\sigma^2}\sum_{i=1}^{\lfloor\frac{R}{h}\rfloor}\frac{\underline{C_\mu^2}i^{-2\beta}}{1+\frac{n}{\sigma^2}\overline{C_\lambda}(hi)^{-\alpha}}\\ &= \begin{cases}\Theta(n^{\max\{0,1+\frac{1-2\beta}{\alpha}\}}), & \alpha\neq 2\beta-1,\\ \Theta(\log n), & \alpha=2\beta-1.\end{cases}\end{aligned}$$

Overall, we have

$$\frac{n}{2\sigma^2}\boldsymbol{\mu}_{1:R}^T(I+\frac{n}{\sigma^2}\Lambda_{1:R})^{-1}\boldsymbol{\mu}_{1:R} = \Theta(n^{\max\{0,1+\frac{1-2\beta}{\alpha}\}}\log^k n),$$

where $k=\begin{cases}0, & \alpha\neq 2\beta-1,\\ 1, & \alpha=2\beta-1.\end{cases}$

By Lemma 20, we have

$$\begin{aligned}\|(I+\frac{n}{\sigma^2}\Lambda_{1:R})^{-1+\gamma/2}\Lambda_{1:R}^{-\gamma/2}\boldsymbol{\mu}_{1:R}\|_2^2 \le& \sum_{p=1}^{R}\frac{C_\mu^2 p^{-2\beta}(\underline{C_\lambda}p^{-\alpha})^{-\gamma}}{(1+\frac{n}{\sigma^2}\underline{C_\lambda}p^{-\alpha})^{2-\gamma}}\\ =&\tilde{O}(\max\{n^{-2+\gamma}, R^{1-2\beta+\alpha\gamma}\})\\ =&\tilde{O}(n^{\max\{-2+\gamma,\frac{1-2\beta}{\alpha}+\gamma+\kappa(1-2\beta+\alpha\gamma)\}}).\end{aligned} \tag{45}$$

Using (42), the second term on the right hand side of (44) is computed as follows:

$$\begin{aligned}&\frac{1}{2}\sum_{j=1}^{\infty}\Bigg[(-1)^{j+1}\frac{1}{\sigma^{2j}}\boldsymbol{\mu}_{1:R}^T(I+\frac{n}{\sigma^2}\Lambda_{1:R})^{-1+\gamma/2}\Lambda_{1:R}^{-\gamma/2}A\Big((I+\frac{n}{\sigma^2}\Lambda_{1:R})^{-1+\gamma}\Lambda_{1:R}^{1-\gamma}A\Big)^{j-1}\\ &\qquad(I+\frac{n}{\sigma^2}\Lambda_{1:R})^{-1+\gamma/2}\Lambda_{1:R}^{-\gamma/2}\boldsymbol{\mu}_{1:R}\Bigg]\\ \le&\frac{1}{2}\sum_{j=1}^{\infty}\frac{1}{\sigma^{2j}}\|A\|^j\Big(\frac{n}{\sigma^2}\Big)^{(-1+\gamma)(j-1)}\|(I+\frac{n}{\sigma^2}\Lambda_{1:R})^{-1+\gamma/2}\Lambda_{1:R}^{-\gamma/2}\boldsymbol{\mu}_{1:R}\|_2^2\\ \le&\frac{1}{2}\sum_{j=1}^{\infty}\frac{1}{\sigma^{2j}}\tilde{O}(n^{\frac{j(1-2\gamma\alpha+\alpha+2\tau)}{2\alpha}})\Big(\frac{n}{\sigma^2}\Big)^{(-1+\gamma)(j-1)}\tilde{O}(n^{\max\{-2+\gamma,\frac{1-2\beta}{\alpha}+\gamma+\kappa(1-2\beta+\alpha\gamma)\}})\\ =&\tilde{O}(n^{\max\{-2+\gamma+\frac{1-2\gamma\alpha+\alpha+2\tau}{2\alpha},\frac{1-2\beta}{\alpha}+\gamma+\frac{1-2\gamma\alpha+\alpha+2\tau}{2\alpha}+\kappa(1-2\beta+\alpha\gamma)\}})\\ =&\tilde{O}(n^{\max\{-2+\frac{1+\alpha+2\tau}{2\alpha},\frac{1-2\beta}{\alpha}+\frac{1+\alpha+2\tau}{2\alpha}+\kappa(1-2\beta+\alpha\gamma)\}}).\end{aligned} \tag{46}$$

Since $\frac{1+\alpha+2\tau}{2\alpha} < \frac{1+\alpha+2\tau}{\alpha+1+2\tau} = 1$, we have $-2+\frac{1+\alpha+2\tau}{2\alpha} < 0$.Also we have

$$\begin{aligned}&\frac{1-2\beta}{\alpha}+\frac{1+\alpha+2\tau}{2\alpha}+\kappa(1-2\beta+\alpha\gamma)\\ =&\frac{1-2\beta}{\alpha}+1+\frac{1-\alpha+2\tau}{2\alpha}+\kappa(1-2\beta+\alpha\gamma)\\ \le&\frac{1-2\beta}{\alpha}+1+\frac{1-\alpha+2\tau}{2\alpha}+\kappa\alpha\gamma\\ <&\frac{1-2\beta}{\alpha}+1,\end{aligned} \tag{47}$$

where the last inequality holds because $\kappa < \frac{\alpha-1-2\tau}{2\alpha^2}$ and $\gamma \leq 1$. Hence we have

$$\begin{aligned}T_{2,R}(D_n) =&\frac{n}{2\sigma^2}\boldsymbol{\mu}_{1:R}^T(I+\frac{n}{\sigma^2}\Lambda_{1:R})^{-1}\boldsymbol{\mu}_{1:R} + \tilde{O}(n^{\max\{-2+\frac{1+\alpha+2\tau}{2\alpha}, \frac{1-2\beta}{\alpha}+\frac{1+\alpha+2\tau}{2\alpha}+\kappa(1-2\beta+\alpha\gamma)\}})\\ =&\Theta(n^{\max\{0,1+\frac{1-2\beta}{\alpha}\}}\log^k n) + \tilde{O}(n^{\max\{-2+\frac{1+\alpha+2\tau}{2\alpha}, \frac{1-2\beta}{\alpha}+\frac{1+\alpha+2\tau}{2\alpha}+\kappa(1-2\beta+\alpha\gamma)\}})\\ =&\Theta(n^{\max\{0,1+\frac{1-2\beta}{\alpha}\}}\log^k n).\end{aligned}$$

where $k = \begin{cases}0, & \alpha \neq 2\beta-1,\\ 1, & \alpha = 2\beta-1.\end{cases}$. Since $\tilde{O}(n^{\max\{-2+\frac{1+\alpha+2\tau}{2\alpha}, \frac{1-2\beta}{\alpha}+\frac{1+\alpha+2\tau}{2\alpha}+\kappa(1-2\beta+\alpha\gamma)\}})$ is lower order term compared to $\Theta(n^{\max\{0,1+\frac{1-2\beta}{\alpha}\}}\log^k n)$, we further have

$$T_{2,R}(D_n) =\Big(\frac{n}{2\sigma^2}\boldsymbol{\mu}_{1:R}^T(I+\frac{n}{\sigma^2}\Lambda_{1:R})^{-1}\boldsymbol{\mu}_{1:R}\Big)(1+o(1))$$

This concludes the proof of the second statement. □

Lemma 12. *Under Assumptions 1, 2 and 3, with probability of at least $1-5\delta$, we have*

$$T_1(D_n) = \left(\frac{1}{2}\log\det(I+\frac{n}{\sigma^2}\Lambda) - \frac{1}{2}\operatorname{Tr}\Big(I-(I+\frac{n}{\sigma^2}\Lambda)^{-1}\Big)\right)(1+o(1)) = \Theta(n^{\frac{1}{\alpha}}), \tag{48}$$

Furthermore, let $\delta = n^{-q}$ where $0 \leq q < \min\{\frac{(2\beta-1)(\alpha-1-2\tau)}{4\alpha^2}, \frac{\alpha-1-2\tau}{2\alpha}\}$. If we assume $\mu_0 = 0$, we have

$$T_2(D_n) = \Big(\frac{n}{2\sigma^2}\boldsymbol{\mu}^T(I+\frac{n}{\sigma^2}\Lambda)^{-1}\boldsymbol{\mu}\Big)(1+o(1)) = \begin{cases}\Theta(n^{\max\{0,1+\frac{1-2\beta}{\alpha}\}}), & \alpha \neq 2\beta-1,\\ \Theta(\log n), & \alpha = 2\beta-1.\end{cases} \tag{49}$$

Proof of Lemma 12. Let $R = n^{\frac{1}{\alpha}+\kappa}$ where $0 \leq \kappa < \frac{\alpha-1-2\tau}{2\alpha^2}$. By Lemmas 8 and 11, with probability of at least $1-5\delta$ we have

$$|T_{1,R}(D_n) - T_1(D_n)| = \tilde{O}(n^{\frac{1}{\alpha}+\kappa(1-\alpha)}), \tag{50}$$

and

$$|T_{2,R}(D_n) - T_2(D_n)| = \tilde{O}\left((\frac{1}{\delta}+1)n^{\max\{(\frac{1}{\alpha}+\kappa)\frac{1-2\beta}{2}, 1+\frac{1-2\beta}{\alpha}+\frac{(1-2\beta)\kappa}{2}, -1-\kappa\alpha, 1+\frac{1-2\beta}{\alpha}-\kappa\alpha\}}\right) \tag{51}$$

as well as

$$T_{1,R}(D_n) = \left(\frac{1}{2}\log\det(I+\frac{n}{\sigma^2}\Lambda_R) - \frac{1}{2}\operatorname{Tr}\Big(I-(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\Big)\right)(1+o(1)) = \Theta(n^{\frac{1}{\alpha}}), \tag{52}$$

and

$$T_{2,R}(D_n) = \Big(\frac{n}{2\sigma^2}\boldsymbol{\mu}^T(I+\frac{n}{\sigma^2}\Lambda)^{-1}\boldsymbol{\mu}\Big)(1+o(1)) = \begin{cases}\Theta(n^{\max\{0,1+\frac{1-2\beta}{\alpha}\}}), & \alpha \neq 2\beta-1,\\ \Theta(\log n), & \alpha = 2\beta-1.\end{cases} \tag{53}$$

We then have

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

Since $\tilde{O}(n^{\frac{1}{\alpha}+\kappa(1-\alpha)})$ is lower order term compared to $\Theta(n^{\frac{1}{\alpha}})$, we further have

$$T_1(D_n) = \left(\frac{1}{2}\log\det(I+\frac{n}{\sigma^2}\Lambda_R) - \frac{1}{2}\operatorname{Tr}\Big(I-(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\Big)\right)(1+o(1)) = \Theta(n^{\frac{1}{\alpha}})$$

Besides, we have

$$\begin{aligned}&\log\det(I+\frac{n}{\sigma^2}\Lambda) - \log\det(I+\frac{n}{\sigma^2}\Lambda_R)\\ &= \sum_{p=R+1}^{\infty}\log(1+\frac{n}{\sigma^2}\lambda_p) \leq \frac{n}{\sigma^2}\sum_{p=R+1}^{\infty}\lambda_p \leq \frac{n}{\sigma^2}\sum_{p=R+1}^{\infty}C_\lambda p^{-\alpha} = \frac{n}{\sigma^2}O(R^{1-\alpha})\\ &= \frac{n}{\sigma^2}O(n^{(1-\alpha)(\frac{1}{\alpha}+\kappa)})\\ &= o(n^{\frac{1}{\alpha}}).\end{aligned}$$

Then we have $\log\det(I+\frac{n}{\sigma^2}\Lambda_R) = \log\det(I+\frac{n}{\sigma^2}\Lambda)(1+o(1))$. Similarly we can prove $\mathrm{Tr}\big(I-(I+\frac{n}{\sigma^2}\Lambda)^{-1}\big) = \mathrm{Tr}\big(I-(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\big)(1+o(1))$. This concludes the proof of the first statement.

As for $T_2(D_n)$, we have

$$\begin{aligned}
T_2(D_n) &= T_{2,R}(D_n) + T_{2,R}(D_n) - T_2(D_n) \\
&= \Theta(n^{\max\{0,1+\frac{1-2\beta}{\alpha}\}}\log^k n) + \tilde{O}\bigg((\frac{1}{\delta}+1)n^{\max\{(\frac{1}{\alpha}+\kappa)\frac{1-2\beta}{2}, 1+\frac{1-2\beta}{\alpha}+\frac{(1-2\beta)\kappa}{2}, -1-\kappa\alpha, 1+\frac{1-2\beta}{\alpha}-\kappa\alpha\}}\bigg) \\
&= \Theta(n^{\max\{0,1+\frac{1-2\beta}{\alpha}\}}\log^k n) + \tilde{O}\bigg(n^{q+\max\{(\frac{1}{\alpha}+\kappa)\frac{1-2\beta}{2}, 1+\frac{1-2\beta}{\alpha}+\frac{(1-2\beta)\kappa}{2}, -1-\kappa\alpha, 1+\frac{1-2\beta}{\alpha}-\kappa\alpha\}}\bigg)
\end{aligned}$$

where we use $\delta = n^{-q}$, $k = \begin{cases} 0, & \alpha \neq 2\beta-1, \\ 1, & \alpha = 2\beta-1. \end{cases}$.

Since $0 \leq \kappa < \frac{\alpha-1-2\tau}{2\alpha^2}$ and $0 \leq q < \min\{\frac{(2\beta-1)(\alpha-1-2\tau)}{4\alpha^2}, \frac{\alpha-1-2\tau}{2\alpha}\}$, we can choose $\kappa < \frac{\alpha-1-2\tau}{2\alpha^2}$ and κ is arbitrarily close to $\frac{\alpha-1-2\tau}{2\alpha^2}$ such that $0 \leq q < \min\{\frac{(2\beta-1)\kappa}{2}, \kappa\alpha\}$. Then we have $(\frac{1}{\alpha}+\kappa)\frac{1-2\beta}{2}+q<0$, $-1-\kappa\alpha+q<0$, $\frac{(1-2\beta)\kappa}{2}+q<0$ and $-\kappa\alpha+q<0$. So we have

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

Since $\tilde{O}\Big((\frac{1}{\delta}+1)n^{\max\{(\frac{1}{\alpha}+\kappa)\frac{1-2\beta}{2}, 1+\frac{1-2\beta}{\alpha}+\frac{(1-2\beta)\kappa}{2}, -1-\kappa\alpha, 1+\frac{1-2\beta}{\alpha}-\kappa\alpha\}}\Big)$ is lower order term compared to $\Theta(n^{\max\{0,1+\frac{1-2\beta}{\alpha}\}}\log^k n)$, we further have

$$T_2(D_n) = T_{2,R}(D_n)(1+o(1)) = \Big(\frac{n}{2\sigma^2}\boldsymbol{\mu}_R^T(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\boldsymbol{\mu}_R\Big)(1+o(1)).$$

Furthermore, we have

$$\begin{aligned}
&\boldsymbol{\mu}^T(I+\frac{n}{\sigma^2}\Lambda)^{-1}\boldsymbol{\mu} - \boldsymbol{\mu}_{\boldsymbol{R}}^T(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\boldsymbol{\mu}_{\boldsymbol{R}} \\
&= \sum_{p=R+1}^{\infty}\frac{\mu_p^2}{(1+\frac{n}{\sigma^2}\lambda_p)} \leq \sum_{p=R+1}^{\infty}\mu_p^2 \leq \frac{n}{\sigma^2}\sum_{p=R+1}^{\infty}C_\mu^2 p^{-2\beta} = O(R^{1-2\beta}) \\
&= O(n^{(1-2\beta)(\frac{1}{\alpha}+\kappa)}) \\
&= o(n^{\frac{1-2\beta}{\alpha}}).
\end{aligned}$$

Then we have $\boldsymbol{\mu}^T(I+\frac{n}{\sigma^2}\Lambda)^{-1}\boldsymbol{\mu} = \boldsymbol{\mu}_R^T(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\boldsymbol{\mu}_R(1+o(1))$. This concludes the proof of the second statement. □

Proof of Theorem 4, $\mu_0 > 0$. Using Lemma 12 and noting that $\frac{1}{\alpha} > 0$, with probability of at least $1-5\tilde{\delta}$, we have

$$\begin{aligned}
\mathbb{E}_\epsilon F^0(D_n) &= T_1(D_n) + T_2(D_n) \\
&= \bigg[\frac{1}{2}\log\det(I+\frac{n}{\sigma^2}\Lambda_R) - \frac{1}{2}\mathrm{Tr}\Big(I-(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\Big) \\
&\quad + \frac{n}{2\sigma^2}\boldsymbol{\mu}_R^T(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\boldsymbol{\mu}_R\bigg](1+o(1)) \\
&= \Theta(n^{\max\{\frac{1}{\alpha}, \frac{1-2\beta}{\alpha}+1\}})
\end{aligned}$$

Letting $\delta = 5\tilde{\delta}$, we get the result. □

In the case of $\mu_0 > 0$, we have the following lemma:

Lemma 13. *Assume that $\sigma^2 = \Theta(1)$. Let $R = n^{\frac{1}{\alpha}+\kappa}$ where $0 < \kappa < \frac{\alpha-1-2\tau}{\alpha^2}$. Assume that $\mu_0 > 0$. Under Assumptions 1, 2 and 3, for sufficiently large n with probability of at least $1-4\delta$ we have*

$$|T_{2,R}(D_n) - T_2(D_n)| = \tilde{O}\bigg((\frac{1}{\delta}+1)n^{\max\{1+(\frac{1}{\alpha}+\kappa)\frac{1-2\beta}{2}, 1-\kappa\alpha\}}\bigg).. \tag{54}$$

Proof of Lemma 13. As for $|T_2(D_n) - T_{2,R}(D_n)|$, we have

$$\begin{aligned}|T_2(D_n) - T_{2,R}(D_n)| = &\left|f(\mathbf{x})^T(I + \frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f(\mathbf{x}) - f_R(\mathbf{x})^T(I + \frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f_R(\mathbf{x})\right| \\ &+ \left|f_R(\mathbf{x})^T(I + \frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f_R(\mathbf{x}) - f_R(\mathbf{x})^T(I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}f_R(\mathbf{x})\right|.\end{aligned} \tag{55}$$

For the first term on the right-hand side of (55), 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 23 and Lemma 35, with probability of at least $1 - 4\delta$, we have

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

As for the second term on the right-hand side of (32), according to Lemma 32, Corollary 30 and Lemma 34, we have

$$\begin{aligned}&\left|f_R(\mathbf{x})^T(I + \frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f_R(\mathbf{x}) - f_R(\mathbf{x})^T(I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}f_R(\mathbf{x})\right| \\ = &\left|\sum_{j=1}^{\infty}(-1)^j f_R(\mathbf{x})^T\left((I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}\frac{\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T}{\sigma^2}\right)^j(I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}f_R(\mathbf{x})\right| \\ \leq &\sum_{j=1}^{\infty}\|(I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}\|_2^{j-1}\cdot\|\frac{\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T}{\sigma^2}\|_2^j\cdot\|(I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}f_R(\mathbf{x})\|_2^2 \\ = &\sum_{j=1}^{\infty}\tilde{O}(n^{-j\kappa\alpha})\tilde{O}((\frac{1}{\delta}+1)n) \\ = &\tilde{O}((\frac{1}{\delta}+1)n^{1-\kappa\alpha}).\end{aligned} \tag{56}$$

By (32), we have

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

□

Lemma 14. *Assume that $\sigma^2 = \Theta(1)$. Let $R = n^{\frac{1}{\alpha}+\kappa}$ where $0 < \kappa < \min\{\frac{\alpha-1-2\tau}{2\alpha^2}, \frac{2\beta-1}{\alpha^2}\}$. Assume that $\mu_0 > 0$. Under Assumptions 1, 2 and 3, with probability of at least $1-\delta$, we have*

$$T_{2,R}(D_n) = \frac{n}{2\sigma^2}\mu_0^2 + \tilde{O}(n^{\max\{\frac{1+7\alpha+2\tau}{8\alpha}, 1+\frac{1-2\beta}{\alpha}\}}). \tag{57}$$

Proof of Lemma 14. Let

$$A = (I + \frac{n}{\sigma^2}\Lambda_R)^{-\gamma/2}\Lambda_R^{\gamma/2}(\Phi_R^T\Phi_R - nI)\Lambda_R^{\gamma/2}(I + \frac{n}{\sigma^2}\Lambda_R)^{-\gamma/2}, \tag{58}$$

where $\frac{1+\alpha+2\tau}{2\alpha} < \gamma \leq 1$. By Corollary 26, with probability of at least $1-\delta$, we have

$$\|A\|_2 = \tilde{O}(n^{\frac{1-2\gamma\alpha+\alpha+2\tau}{2\alpha}}). \tag{59}$$

When n is sufficiently large, $\|A\|_2$ is less than 1. Let $\boldsymbol{\mu}_{R,1} = (\mu_0, 0, \ldots, 0)$ and $\boldsymbol{\mu}_{R,2} = (0, \mu_1, \ldots, \mu_R)$. Then $\boldsymbol{\mu}_R = \boldsymbol{\mu}_{R,1} + \boldsymbol{\mu}_{R,2}$. Let $\tilde{\Lambda}_{1,R} = \mathrm{diag}\{1, \lambda_1, \ldots, \lambda_R\}$ and $I_{0,R} = (0, 1, \ldots, 1)$. Then $\Lambda_R = \tilde{\Lambda}_{1,R}I_{0,R}$. Let $B = (I + \frac{n}{\sigma^2}\Lambda_R)^{-1/2}\tilde{\Lambda}_{1,R}^{1/2}(\Phi_R^T\Phi_R - nI)\tilde{\Lambda}_{1,R}^{1/2}(I + \frac{n}{\sigma^2}\Lambda_R)^{-1/2}$. By Corollary 27, we have $\|B\|_2 = O(\sqrt{\log\frac{R}{\delta}}n^{\frac{1}{2}})$. By Lemma 10, we have

$$\begin{aligned}
T_{2,R}(D_n) = {} & \frac{n}{2\sigma^2}\boldsymbol{\mu}_R^T(I + \frac{n}{\sigma^2}\Lambda_R)^{-1}\boldsymbol{\mu}_R \\
& + \frac{1}{2}\sum_{j=1}^{\infty}\Bigg[(-1)^{j+1}\boldsymbol{\mu}_R^T\frac{1}{\sigma^2}(I + \frac{n}{\sigma^2}\Lambda_R)^{-1}(\Phi_R^T\Phi_R - nI)\bigg(\frac{1}{\sigma^2}(I + \frac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R(\Phi_R^T\Phi_R - nI)\bigg)^{j-1} \\
& \qquad (I + \frac{n}{\sigma^2}\Lambda_R)^{-1}\boldsymbol{\mu}_R\Bigg]
\end{aligned} \tag{60}$$

As for the first term on the right hand side of (60), by Lemma 19, we have

$$\frac{n}{2\sigma^2}\mu^T(I + \frac{n}{\sigma^2}\Lambda)^{-1}\mu \leq \frac{n}{2\sigma^2}\left(\mu_0^2 + \sum_{p=1}^{R}\frac{C_\mu^2 p^{-2\beta}}{1 + \frac{n}{\sigma^2}\underline{C_\lambda}p^{-\alpha}}\right) = \frac{n}{2\sigma^2}\mu_0^2 + \tilde{O}(n^{\max\{0, 1+\frac{1-2\beta}{\alpha}\}}).$$

We define $Q_{1,j}$, $Q_{2,j}$ and $Q_{3,j}$ by

$$\begin{aligned}
Q_{1,j} = {} & \boldsymbol{\mu}_{R,1}^T\frac{1}{\sigma^2}(I + \frac{n}{\sigma^2}\Lambda_R)^{-1}(\Phi_R^T\Phi_R - nI)\bigg(\frac{1}{\sigma^2}(I + \frac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R(\Phi_R^T\Phi_R - nI)\bigg)^{j-1} \\
& (I + \frac{n}{\sigma^2}\Lambda_R)^{-1}\boldsymbol{\mu}_{R,1} \\
Q_{2,j} = {} & \boldsymbol{\mu}_{R,1}^T\frac{1}{\sigma^2}(I + \frac{n}{\sigma^2}\Lambda_R)^{-1}(\Phi_R^T\Phi_R - nI)\bigg(\frac{1}{\sigma^2}(I + \frac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R(\Phi_R^T\Phi_R - nI)\bigg)^{j-1} \\
& (I + \frac{n}{\sigma^2}\Lambda_R)^{-1}\boldsymbol{\mu}_{R,2} \\
Q_{3,j} = {} & \boldsymbol{\mu}_{R,2}^T\frac{1}{\sigma^2}(I + \frac{n}{\sigma^2}\Lambda_R)^{-1}(\Phi_R^T\Phi_R - nI)\bigg(\frac{1}{\sigma^2}(I + \frac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R(\Phi_R^T\Phi_R - nI)\bigg)^{j-1} \\
& (I + \frac{n}{\sigma^2}\Lambda_R)^{-1}\boldsymbol{\mu}_{R,2}
\end{aligned} \tag{61}$$

The quantity $Q_{3,j}$ actually shows up in the case of $\mu_0 = 0$ in the proof of Lemma 11. By (44), (46) and (47), we have that

$$|\sum_{j=1}^{\infty}(-1)^{j+1}Q_{3,j}| = |\sum_{j=1}^{\infty}(-1)^{j+1}\tilde{O}(n^{\frac{(j-1)(1-\alpha+2\tau)}{2\alpha}})o(n^{\max\{0,1+\frac{1-2\beta}{\alpha}\}})| = o(n^{\max\{0,1+\frac{1-2\beta}{\alpha}\}}). \tag{62}$$

For $Q_{1,j}$, we have

$$\begin{aligned}
Q_{1,1} &= \frac{1}{\sigma^{2j}}\boldsymbol{\mu}_{R,1}^T(I + \frac{n}{\sigma^2}\Lambda_R)^{-1+\frac{\gamma}{2}}B(I + \frac{n}{\sigma^2}\Lambda_R)^{-1+\frac{\gamma}{2}}\boldsymbol{\mu}_{R,1} \\
&\leq \frac{1}{\sigma^{2j}}\|\boldsymbol{\mu}_{R,1}\|_2^2\|(I + \frac{n}{\sigma^2}\Lambda_R)^{-1+\frac{\gamma}{2}}\|_2^2\|B\|_2 \\
&= O(\sqrt{\log\frac{R}{\delta}}n^{\frac{1}{2}}),
\end{aligned}$$

where in the last equality we use $\|B\|_2 = O(\sqrt{\log \frac{R}{\delta}} n^{\frac{1}{2}})$. For $j \geq 2$, we have

$$\begin{aligned}
Q_{1,j} &= \frac{1}{\sigma^{2j}} \boldsymbol{\mu}_{R,1}^T (I + \frac{n}{\sigma^2}\Lambda_R)^{-1+\frac{\gamma}{2}} B \Big((I + \frac{n}{\sigma^2}\Lambda_R)^{-1+\gamma} \Lambda_R^{1-\gamma} A \Big)^{j-2} (I + \frac{n}{\sigma^2}\Lambda_R)^{-1+\gamma} \Lambda_R^{1-\gamma} \\
&\quad B (I + \frac{n}{\sigma^2}\Lambda_R)^{-1+\frac{\gamma}{2}} \boldsymbol{\mu}_{R,1} \\
&\leq \frac{1}{\sigma^{2j}} \|\boldsymbol{\mu}_{R,1}\|_2^2 \|(I + \frac{n}{\sigma^2}\Lambda_R)^{-1+\frac{\gamma}{2}}\|_2^2 \|B\|_2^2 \|A\|_2^{j-2} \|(I + \frac{n}{\sigma^2}\Lambda_R)^{-1+\gamma} \Lambda_R^{1-\gamma}\|_2^{j-1} \\
&= O(\log \frac{R}{\delta} n \cdot n^{\frac{(j-2)(1-2\gamma\alpha+\alpha+2\tau)}{2\alpha}} \cdot n^{-(1-\gamma)(j-1)}) \\
&= O(\log \frac{R}{\delta} n^{\gamma} \cdot n^{\frac{(j-2)(1-\alpha+2\tau)}{2\alpha}}).
\end{aligned}$$

Then we have

$$|\sum_{j=1}^{\infty} (-1)^{j+1} Q_{1,j}| \leq O(\sqrt{\log \frac{R}{\delta}} n^{\frac{1}{2}}) + \sum_{j=2}^{\infty} O(\log \frac{R}{\delta} n^{\gamma} \cdot n^{\frac{(j-2)(1-\alpha+2\tau)}{2\alpha}}) = O(\log \frac{R}{\delta} n^{\gamma}) \tag{63}$$

For $Q_{2,j}$, we have

$$\begin{aligned}
Q_{2,j} &= \frac{1}{\sigma^{2j}} \boldsymbol{\mu}_{R,1}^T (I + \frac{n}{\sigma^2}\Lambda_R)^{-1+\frac{\gamma}{2}} B \Big((I + \frac{n}{\sigma^2}\Lambda_R)^{-1+\gamma} \Lambda_R^{1-\gamma} A \Big)^{j-1} (I + \frac{n}{\sigma^2}\Lambda)^{-1+\frac{\gamma}{2}} \tilde{\Lambda}_{1,R}^{-\frac{\gamma}{2}} \boldsymbol{\mu}_{R,2} \\
&\leq \frac{1}{\sigma^{2j}} \|\boldsymbol{\mu}_{R,1}\|_2 \|B\|_2 \|A\|_2^{j-1} \|(I + \frac{n}{\sigma^2}\Lambda_R)^{-1+\gamma} \Lambda_R^{1-\gamma}\|_2^{j-1} \|(I + \frac{n}{\sigma^2}\Lambda)^{-1+\frac{\gamma}{2}} \tilde{\Lambda}_{1,R}^{-\frac{\gamma}{2}} \boldsymbol{\mu}_{R,2}\|_2 \\
&= O(\sqrt{\log \frac{R}{\delta}} n^{\frac{1}{2}} \cdot n^{\frac{(j-1)(1-\alpha+2\tau)}{2\alpha}}) \|(I + \frac{n}{\sigma^2}\Lambda)^{-1+\frac{\gamma}{2}} \tilde{\Lambda}_{1,R}^{-\frac{\gamma}{2}} \boldsymbol{\mu}_{R,2}\|_2.
\end{aligned}$$

Since $\|(I + \frac{n}{\sigma^2}\Lambda)^{-1+\frac{\gamma}{2}} \tilde{\Lambda}_{1,R}^{-\frac{\gamma}{2}} \boldsymbol{\mu}_{R,2}\|_2$ is actually the case of $\mu_0 = 0$, we can use (45) in the proof of Lemma 11 and get

$$\begin{aligned}
\|(I + \frac{n}{\sigma^2}\Lambda)^{-1+\frac{\gamma}{2}} \tilde{\Lambda}_{1,R}^{-\frac{\gamma}{2}} \boldsymbol{\mu}_{R,2}\|_2^2 =& \|(I + \frac{n}{\sigma^2}\Lambda_{1:R})^{-1+\gamma/2} \Lambda_{1:R}^{-\gamma/2} \boldsymbol{\mu}_{1:R}\|_2^2 \\
=& \tilde{O}(n^{\max\{-2+\gamma, \frac{1-2\beta}{\alpha}+\gamma+\kappa(1-2\beta+\alpha\gamma)\}} \\
=& \tilde{O}(n^{\max\{-2+\gamma, \frac{1-2\beta}{\alpha}+\gamma+\kappa(1-2\beta+\alpha\gamma)\}}) \\
=& o(n^{\gamma}),
\end{aligned} \tag{64}$$

where in the last equality we use $\kappa < \frac{2\beta-1}{\alpha^2}$. Then we have

$$|\sum_{j=1}^{\infty} (-1)^{j+1} Q_{2,j}| \leq \sum_{j=1}^{\infty} o(\sqrt{\log \frac{R}{\delta}} n^{\frac{1+\gamma}{2}} \cdot n^{\frac{(j-1)(1-\alpha+2\tau)}{2\alpha}}) = o(\sqrt{\log \frac{R}{\delta}} n^{\frac{1+\gamma}{2}}) \tag{65}$$

Choosing $\gamma = \frac{1}{2}(1 + \frac{1+\alpha+2\tau}{2\alpha}) = \frac{1+3\alpha+2\tau}{4\alpha} < 1$, we have

$$\begin{aligned}
T_{2,R}(D_n) &= \frac{n}{2\sigma^2} \boldsymbol{\mu}_R^T (I + \frac{n}{\sigma^2}\Lambda_R)^{-1} \boldsymbol{\mu}_R + \sum_{j=1}^{\infty} (-1)^{j+1} (Q_{1,j} + Q_{2,j} + Q_{3,j}) \\
&= \frac{n}{2\sigma^2}\mu_0^2 + \tilde{O}(n^{\max\{0, 1+\frac{1-2\beta}{\alpha}\}}) + o(n^{\max\{0, 1+\frac{1-2\beta}{\alpha}\}}) + O(\log \frac{R}{\delta} n^{\gamma}) + o(\sqrt{\log \frac{R}{\delta}} n^{\frac{1+\gamma}{2}}) \\
&= \frac{n}{2\sigma^2}\mu_0^2 + \tilde{O}(n^{\max\{\frac{1+\gamma}{2}, 1+\frac{1-2\beta}{\alpha}\}}) \\
&= \frac{n}{2\sigma^2}\mu_0^2 + \tilde{O}(n^{\max\{\frac{1+7\alpha+2\tau}{8\alpha}, 1+\frac{1-2\beta}{\alpha}\}}).
\end{aligned}$$

□

Proof of Theorem 4, $\mu_0 > 0$. Let $R = n^{\frac{1}{\alpha}+\kappa}$ where $0 < \kappa < \min\{\frac{\alpha-1-2\tau}{2\alpha^2}, \frac{2\beta-1}{\alpha^2}\}$. Since $0 \leq q < \min\{\frac{2\beta-1}{2}, \alpha\} \cdot \min\{\frac{\alpha-1-2\tau}{2\alpha^2}, \frac{2\beta-1}{\alpha^2}\}$, we can choose $\kappa < \min\{\frac{\alpha-1-2\tau}{2\alpha^2}, \frac{2\beta-1}{\alpha^2}\}$ and κ is

arbitrarily close to $\kappa < \min\{\frac{\alpha-1-2\tau}{2\alpha^2}, \frac{2\beta-1}{\alpha^2}\}$ such that $0 \le q < \min\{\frac{(2\beta-1)\kappa}{2}, \kappa\alpha\}$. Then we have $(\frac{1}{\alpha}+\kappa)\frac{1-2\beta}{2} + q < 0$, and $-\kappa\alpha + q < 0$. As for $T_2(D_n)$, we have

$$\begin{aligned}
T_2(D_n) &\le T_{2,R}(D_n) + |T_{2,R}(D_n) - T_2(D_n)| \\
&= \frac{n}{2\sigma^2}\mu_0^2 + \tilde{O}(n^{\max\{\frac{1+7\alpha+2\tau}{8\alpha}, 1+\frac{1-2\beta}{\alpha}\}}) + \tilde{O}\Big((\frac{1}{\delta}+1)n^{\max\{1+(\frac{1}{\alpha}+\kappa)\frac{1-2\beta}{2}, 1-\kappa\alpha\}}\Big) \\
&= \frac{n}{2\sigma^2}\mu_0^2 + \tilde{O}(n^{\max\{\frac{1+7\alpha+2\tau}{8\alpha}, 1+\frac{1-2\beta}{\alpha}\}}) + \tilde{O}\Big(n^{q+\max\{1+(\frac{1}{\alpha}+\kappa)\frac{1-2\beta}{2}, 1-\kappa\alpha\}}\Big) \\
&= \frac{n}{2\sigma^2}\mu_0^2 + o(n).
\end{aligned}$$

By Lemma 12, we have $T_1(D_n) = O(n^{\frac{1}{\alpha}})$. Hence $\mathbb{E}_\epsilon F^0(D_n) = T_1(D_n) + T_2(D_n) = \frac{n}{2\sigma^2}\mu_0^2 + o(n)$. □

D.2 Proofs related to the asymptotics of the generalization error

In this section we derive the asymptotics of the Bayesian generalization error (1). Define

$$G_1(D_n) = \mathbb{E}_{(x_{n+1},y_{n+1})}(T_1(D_{n+1}) - T_1(D_n)), \tag{66}$$
$$G_2(D_n) = \mathbb{E}_{(x_{n+1},y_{n+1})}(T_2(D_{n+1}) - T_2(D_n)). \tag{67}$$

Using (3), we have

$$\mathbb{E}_{\boldsymbol{\epsilon}} G(D_n) = G_1(D_n) + G_2(D_n). \tag{68}$$

We derive the asymptotics of the expected generalization error (68) by analyzing the asymptotics of the components $G_1(D_n)$ and $G_2(D_n)$.

Next, we flesh out the complete proof.

Lemma 15. *Assume $\sigma^2 = \Theta(n^t)$ where $1 - \frac{\alpha}{1+2\tau} < t < 1$. Let $R = n^{(\frac{2\alpha-1}{\alpha(\alpha-1)}+1)(1-t)}$. Under Assumptions 1, 2 and 3, with probability of at least $1-\delta$ over sample inputs $(x_i)_{i=1}^n$, we have*

$$G_1(D_n) = \frac{1+o(1)}{2\sigma^2}\Big(\mathrm{Tr}(I + \tfrac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R - \|\Lambda_R^{1/2}(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-1}\|_F^2\Big) = \frac{1}{\sigma^2}\Theta\Big(n^{\frac{(1-\alpha)(1-t)}{\alpha}}\Big). \tag{69}$$

Proof of Lemma 15. Let $G_{1,R}(D_n) = \mathbb{E}_{(x_{n+1},y_{n+1})}(T_{1,R}(D_{n+1}) - T_{1,R}(D_n))$, where $R = n^C$ for some constant C. By Lemma 8, we have that

$$\begin{aligned}
|G_1(D_n) - G_{1,R}(D_n)| &= \big|\mathbb{E}_{(x_{n+1},y_{n+1})}[T_1(D_{n+1}) - T_{1,R}(D_{n+1})] - [T_1(D_n) - T_{1,R}(D_n)]\big| \\
&= \big|\mathbb{E}_{(x_{n+1},y_{n+1})}O((n+1)R^{1-\alpha})\big| + \big|O(nR^{1-\alpha})\big| \\
&= O(\tfrac{1}{\sigma^2}nR^{1-\alpha}).
\end{aligned} \tag{70}$$

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

$$\begin{aligned}
G_{1,R}(D_n) &= \mathbb{E}_{(x_{n+1},y_{n+1})}(T_{1,R}(D_{n+1}) - T_{1,R}(D_n)) \\
&= \mathbb{E}_{(x_{n+1},y_{n+1})}\left(\frac{1}{2}\log\det(I + \frac{\widetilde{\Phi}_R\Lambda_R\widetilde{\Phi}_R^T}{\sigma^2}) - \frac{1}{2}\mathrm{Tr}(I - (I + \frac{\widetilde{\Phi}_R\Lambda_R\widetilde{\Phi}_R^T}{\sigma^2})^{-1})\right) \\
&- \left(\frac{1}{2}\log\det(I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2}) - \frac{1}{2}\mathrm{Tr}(I - (I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1})\right) \\
&= \frac{1}{2}\left(\mathbb{E}_{(x_{n+1},y_{n+1})}\log\det(I + \frac{\widetilde{\Phi_R}\Lambda_R\widetilde{\Phi_R}^T}{\sigma^2}) - \log\det(I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})\right) \\
&- \frac{1}{2}\left(\mathbb{E}_{(x_{n+1},y_{n+1})}\mathrm{Tr}(I - (I + \frac{\widetilde{\Phi}_R\Lambda_R\widetilde{\Phi}_R^T}{\sigma^2})^{-1}) - \mathrm{Tr}(I - (I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1})\right).
\end{aligned} \tag{71}$$

As for the first term in the right hand side (71), we have

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

Let

$$
A=(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}\Lambda_R^{1/2}(\Phi_R^T\Phi_R-nI)\Lambda_R^{1/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}. \tag{72}
$$

According to Corollary 26, with probability of at least $1-\delta$, we have $\|\frac{1}{\sigma^2}A\|_2 = O(\sqrt{\log\frac{R}{\delta}}n^{\frac{1-\alpha+2\tau}{2\alpha}-\frac{(1+2\tau)t}{2\alpha}}) = o(1)$. When n is sufficiently large, $\|\frac{1}{\sigma^2}A\|_2$ is less than 1. By Lemma 31, we have

$$
\begin{aligned}
&\eta_R^T(I+\frac{\Lambda_R\Phi_R^T\Phi_R}{\sigma^2})^{-1}\Lambda_R\eta_R\\
=&\eta_R^T(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R\eta_R+\sum_{j=1}^{\infty}(-1)^j\eta_R^T\left(\frac{1}{\sigma^2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R(\Phi_R^T\Phi_R-nI)\right)^j(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R\eta_R\\
=&\eta_R^T(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R\eta_R+\sum_{j=1}^{\infty}(-1)^j\frac{1}{\sigma^{2j}}\eta_R^T(I+\frac{n}{\sigma^{2j}}\Lambda_R)^{-1/2}\Lambda_R^{1/2}A^j(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}\Lambda_R^{1/2}\eta_R\\
\le&\eta_R^T(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R\eta_R+\sum_{j=1}^{\infty}\|\frac{1}{\sigma^2}A\|_2^j\|(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}\Lambda_R^{1/2}\eta_R\|_2^2\\
\le&\sum_{p=1}^{R}\phi_p^2(x_{n+1})\frac{\overline{C_\lambda}p^{-\alpha}}{1+n\underline{C_\lambda}p^{-\alpha}/\sigma^2}+\sum_{j=1}^{\infty}\|\frac{1}{\sigma^2}A\|_2^j\sum_{p=1}^{R}\phi_p^2(x_{n+1})\frac{\overline{C_\lambda}p^{-\alpha}}{1+n\underline{C_\lambda}p^{-\alpha}/\sigma^2}\\
\le&\sum_{p=1}^{R}\frac{\overline{C_\lambda}p^{-\alpha}p^{2\tau}}{1+n\underline{C_\lambda}p^{-\alpha}/\sigma^2}+\sum_{j=1}^{\infty}\|\frac{1}{\sigma^2}A\|_2^j\sum_{p=1}^{R}\frac{\overline{C_\lambda}p^{-\alpha}p^{2\tau}}{1+n\underline{C_\lambda}p^{-\alpha}/\sigma^2}\\
\le&O(n^{\frac{(1-\alpha+2\tau)(1-t)}{\alpha}})+\sum_{j=1}^{\infty}\|\frac{1}{\sigma^2}A\|_2^jO(n^{\frac{(1-\alpha+2\tau)(1-t)}{\alpha}})\\
=&O(n^{\frac{(1-\alpha+2\tau)(1-t)}{\alpha}})=o(1),
\end{aligned} \tag{73}
$$

where we use Lemma 19 in the last inequality. Next we have

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

where in the last equality we use the fact that $\mathbb{E}_{(x_{n+1},y_{n+1})}\eta_R\eta_R^T=I$. By Lemma 31, we have

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

By Lemma 19, we have

$$\begin{aligned}
\mathrm{Tr}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R&\le\sum_{p=1}^{R}\frac{\overline{C_\lambda}p^{-\alpha}}{1+n\underline{C_\lambda}p^{-\alpha}/\sigma^2}=\Theta(n^{\frac{(1-\alpha)(1-t)}{\alpha}})\\
\mathrm{Tr}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R&\ge\sum_{p=1}^{R}\frac{\underline{C_\lambda}p^{-\alpha}}{1+n\overline{C_\lambda}p^{-\alpha}/\sigma^2}=\Theta(n^{\frac{(1-\alpha)(1-t)}{\alpha}}).
\end{aligned}$$

Overall,

$$\mathrm{Tr}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R=\Theta(n^{\frac{(1-\alpha)(1-t)}{\alpha}}). \tag{74}$$

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

$$\begin{aligned}
&\bigg|\mathrm{Tr}\,\frac{1}{\sigma^{2j}}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}\Lambda_R^{1/2}A^j(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}\Lambda_R^{1/2}\bigg|\\
&=\bigg|\sum_{p=1}^{R}\frac{\lambda_p\frac{1}{\sigma^{2j}}(A^j)_{p,p}}{1+n\lambda_p/\sigma^2}\bigg|\le\sum_{p=1}^{R}\frac{\lambda_p\|\frac{1}{\sigma^{2j}}A\|_2^j}{1+n\lambda_p/\sigma^2}=\Theta(n^{\frac{(1-\alpha)(1-t)}{\alpha}})\tilde{O}(n^{\frac{j(1-\alpha+2\tau-(1+2\tau)t)}{2\alpha}}(\log R)^{j/2}),
\end{aligned} \tag{75}$$

where in the last step we used (74). According to (74) and (75), we have

$$\begin{aligned}
&\frac{1}{2}\bigg(\mathbb{E}_{(x_{n+1},y_{n+1})}\log\det(I+\frac{\widetilde{\Phi}_R\Lambda_R\widetilde{\Phi}_R^T}{\sigma^2})-\log\det(I+\frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})\bigg)\\
&=\frac{1}{2\sigma^2}\bigg(\mathrm{Tr}(I+\frac{\Lambda_R\Phi_R^T\Phi_R}{\sigma^2})^{-1}\Lambda_R\bigg)(1+o(1))=\tfrac{1}{\sigma^2}\Theta(n^{\frac{(1-\alpha)(1-t)}{\alpha}})+\tfrac{1}{\sigma^2}\sum_{j=1}^{\infty}\Theta(n^{\frac{(1-\alpha)(1-t)}{\alpha}})\tilde{O}(n^{\frac{j(1-\alpha+2\tau-(1+2\tau)t)}{2\alpha}}(\log R)^{j/2}\\
&=\tfrac{1}{\sigma^2}\Theta(n^{\frac{(1-\alpha)(1-t)}{\alpha}})+\tfrac{1}{\sigma^2}\Theta(n^{\frac{(1-\alpha)(1-t)}{\alpha}})o(1)=\tfrac{1}{\sigma^2}\Theta(n^{\frac{(1-\alpha)(1-t)}{\alpha}})\\
&=\frac{1}{2\sigma^2}\Big(\mathrm{Tr}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R\Big)(1+o(1)).
\end{aligned} \tag{76}$$

Using the Woodbury matrix identity, the second term in the right hand side (71) is given by

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

where the last equality uses the Sherman–Morrison formula. According to (73), we get

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

where in the penultimate equality we use $\mathrm{Tr}(BB^T)=\|B\|_F^2$, $\|B\|_F$ is the Frobenius norm of A, and in the last equality we use the definition of A (72). Then we have

$$
\begin{aligned}
&\frac{1+o(1)}{2\sigma^2}\|\Lambda_R^{1/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}(I+\frac{1}{\sigma^2}A)^{-1}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}\|_F^2\\
&=\frac{1+o(1)}{2\sigma^2}\|\Lambda_R^{1/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}(I+\sum_{j=1}^{\infty}(-1)^j\frac{1}{\sigma^{2j}}A^j)(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}\|_F^2\\
&=\frac{1+o(1)}{2\sigma^2}\|\Lambda_R^{1/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}+\sum_{j=1}^{\infty}(-1)^j\frac{1}{\sigma^{2j}}\Lambda_R^{1/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}A^j(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}\|_F^2.
\end{aligned}
\tag{77}
$$

By Lemma 19, we have

$$
\begin{aligned}
\|\Lambda_R^{1/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\|_F&\leq\sqrt{\sum_{p=1}^{R}\frac{\overline{C_\lambda}p^{-\alpha}}{(1+n\underline{C_\lambda}p^{-\alpha}/\sigma^2)^2}}=\Theta(n^{\frac{(1-\alpha)(1-t)}{2\alpha}})\\
\|\Lambda_R^{1/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\|_F&\geq\sqrt{\sum_{p=1}^{R}\frac{\underline{C_\lambda}p^{-\alpha}}{(1+n\overline{C_\lambda}p^{-\alpha}/\sigma^2)^2}}=\Theta(n^{\frac{(1-\alpha)(1-t)}{2\alpha}}).
\end{aligned}
$$

Overall, we have

$$\|\Lambda_R^{1/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\|_F=\Theta(n^{\frac{(1-\alpha)(1-t)}{2\alpha}}). \tag{78}$$

Since $\|\frac{1}{\sigma^2}A\|_2=O(\sqrt{\log\frac{R}{\delta}}n^{\frac{1-\alpha+2\tau}{2\alpha}-\frac{(1+2\tau)t}{2\alpha}})=o(1)$, we have

$$\begin{aligned}&\|\frac{1}{\sigma^{2j}}\Lambda_R^{1/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}A^j(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}\|_F\\&\le\|\Lambda_R^{1/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}\|_F\|\frac{1}{\sigma^2}A\|_2^j\|(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}\|_2\\&=O(n^{\frac{(1-\alpha)(1-t)}{2\alpha}})\tilde{O}(n^{\frac{j(1-\alpha+2\tau-(1+2\tau)t)}{2\alpha}}(\log R)^{j/2}),\end{aligned} \tag{79}$$

where in the first inequality we use the fact that $\|AB\|_F\le\|A\|_F\|B\|_2$ when B is symmetric. By Lemma 19, we have

$$\begin{aligned}&\frac{1}{\sigma^{2j}}\left|\operatorname{Tr}\Lambda_R^{1/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R^{1/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}A^j(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}\right|\\&=\left|\sum_{p=1}^R\frac{\lambda_p((\frac{1}{\sigma^2}A)^j)_{p,p}}{(1+n\lambda_p/\sigma^2)^2}\right|\le\sum_{p=1}^R\frac{\lambda_p\|\frac{1}{\sigma^2}A\|_2^j}{(1+n\lambda_p/\sigma^2)^2}=\Theta(n^{\frac{(1-\alpha)(1-t)}{\alpha}})\tilde{O}(n^{\frac{j(1-\alpha+2\tau-(1+2\tau)t)}{2\alpha}}(\log R)^{j/2}),\end{aligned} \tag{80}$$

According to (78), (79) and (80), we have

$$\begin{aligned}&\frac{1}{2}\left(\mathbb{E}_{(x_{n+1},y_{n+1})}\operatorname{Tr}(I-(I+\frac{\widetilde{\Phi}_R\Lambda_R\widetilde{\Phi}_R^T}{\sigma^2})^{-1}-\operatorname{Tr}(I-(I+\frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}\right)\\=&\frac{1+o(1)}{2\sigma^2}\operatorname{Tr}(I+\frac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}\Lambda_R(I+\frac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}\\=&\frac{1+o(1)}{2\sigma^2}\|\Lambda_R^{1/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}+\sum_{j=1}^\infty(-1)^j\frac{1}{\sigma^{2j}}\Lambda_R^{1/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}A^j(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}\|_F^2\\=&\frac{1+o(1)}{2\sigma^2}\left(\|\Lambda_R^{1/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\|_F^2+\sum_{j=1}^\infty\left\|\frac{1}{\sigma^{2j}}\Lambda_R^{1/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}A^j(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}\right\|_F^2\right.\\&\left.+2\operatorname{Tr}\Lambda_R^{1/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\sum_{j=1}^\infty(-1)^j\frac{1}{\sigma^{2j}}\Lambda_R^{1/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}A^j(I+\frac{n}{\sigma^2}\Lambda_R)^{-1/2}\right)\\=&\frac{1+o(1)}{2\sigma^2}\left(\Theta(n^{\frac{(1-\alpha)(1-t)}{\alpha}})+\sum_{j=1}^\infty\frac{1}{\sigma^{2j}}O(n^{\frac{(1-\alpha)(1-t)}{\alpha}})\tilde{O}(n^{\frac{j(1-\alpha+2\tau-(1+2\tau)t)}{2\alpha}}(\log R)^{j/2})\right.\\&\left.+2\sum_{j=1}^\infty\frac{1}{\sigma^{2j}}\Theta(n^{\frac{(1-\alpha)(1-t)}{\alpha}})\tilde{O}(n^{\frac{j(1-\alpha+2\tau-(1+2\tau)t)}{2\alpha}}(\log R)^{j/2})\right)\\=&\tfrac{1}{\sigma^2}\Theta(n^{\frac{(1-\alpha)(1-t)}{\alpha}})=\frac{1+o(1)}{2\sigma^2}\|\Lambda_R^{1/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\|_F^2.\end{aligned} \tag{81}$$

Combining (76) and (81) we get that $G_{1,R}(D_n)=\frac{1+o(1)}{2\sigma^2}(\operatorname{Tr}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R+\|\Lambda_R^{1/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\|_F^2)=\frac{1}{\sigma^2}\Theta(n^{\frac{(1-\alpha)(1-t)}{\alpha}})$. From (70) we have that $G_1(D_n)\le G_{1,R}(D_n)+|G_1(D_n)-G_{1,R}(D_n)|=\frac{1}{\sigma^2}\Theta(n^{\frac{(1-\alpha)(1-t)}{\alpha}})+O(n\frac{1}{\sigma^2}R^{1-\alpha})$. Choosing $R=n^{(\frac{2\alpha-1}{\alpha(\alpha-1)}+1)(1-t)}$ we conclude the proof. □

Lemma 16. *Assume $\sigma^2=\Theta(n^t)$ where $1-\frac{\alpha}{1+2\tau}<t<1$. Let $S=n^D$. Assume that $\|\xi\|_2=1$. When n is sufficiently large, with probability of at least $1-2\delta$ we have*

$$\|(I+\tfrac{1}{\sigma^2}\Phi_S\Lambda_S\Phi_S^T)^{-1}\Phi_S\Lambda_S\xi\|_2=O(\sqrt{(\tfrac{1}{\delta}+1)n}\cdot n^{-(1-t)}). \tag{82}$$

Proof of Lemma 16. Using the Woodbury matrix identity, we have that

$$\begin{aligned}((I+\frac{1}{\sigma^2}\Phi_S\Lambda_S\Phi_S^T)^{-1}\Phi_S\Lambda_S\xi =&\big[I-\Phi_S(\sigma^2 I+\Lambda_S\Phi_S^T\Phi_S)^{-1}\Lambda_S\Phi_S^T\big]\Phi_S\Lambda_S\xi\\ =&\Phi_S\Lambda_S\xi-\Phi_S(\sigma^2 I+\Lambda_S\Phi_S^T\Phi_S)^{-1}\Lambda_S\Phi_S^T\Phi_S\Lambda_S\xi\\ =&\Phi_S(I+\tfrac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}\Lambda_S\xi.\end{aligned} \tag{83}$$

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 $\gamma>\frac{1+\alpha+2\tau-(1+2\tau+2\alpha)t}{2\alpha(1-t)}$. By Corollary 26, with probability of at least $1-\delta$, we have $\|\frac{1}{\sigma^2}A\|_2=\tilde{O}(n^{\frac{1+\alpha+2\tau-(1+2\tau+2\alpha)t}{2\alpha}-\gamma(1-t)})$. When n is sufficiently large, $\|\frac{1}{\sigma^2}A\|_2$ is less than 1. By Lemma 31, 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

$$\begin{aligned}&\|(I+\tfrac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}\Lambda_S\xi\|_2\\ =&\left\|\left((I+\frac{n}{\sigma^2}\Lambda_S)^{-1}+\sum_{j=1}^{\infty}(-1)^j\left(\frac{1}{\sigma^2}(I+\frac{n}{\sigma^2}\Lambda_S)^{-1}\Lambda_S(\Phi_S^T\Phi_S-nI)\right)^j(I+\frac{n}{\sigma^2}\Lambda_S)^{-1}\right)\Lambda_S\xi\right\|_2\\ \le&\left(\|(I+\frac{n}{\sigma^2}\Lambda_S)^{-1}\Lambda_S\xi\|_2+\sum_{j=1}^{\infty}\left\|\left(\frac{1}{\sigma^2}(I+\frac{n}{\sigma^2}\Lambda_S)^{-1}\Lambda_S(\Phi_S^T\Phi_S-nI)\right)^j(I+\frac{n}{\sigma^2}\Lambda_S)^{-1}\Lambda_S\xi\right\|_2\right).\end{aligned} \tag{84}$$

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

$$\|(I+\frac{n}{\sigma^2}\Lambda_S)^{-1}\Lambda_S\xi\|_2\le\|(I+\frac{n}{\sigma^2}\Lambda_S)^{-1}\Lambda_S\|_2\|\xi\|_2\le\frac{\sigma^2}{n}=O(n^{-(1-t)}). \tag{85}$$

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

$$\begin{aligned}&\left\|\left(\frac{1}{\sigma^2}(I+\frac{n}{\sigma^2}\Lambda_S)^{-1}\Lambda_S(\Phi_S^T\Phi_S-nI)\right)^j(I+\frac{n}{\sigma^2}\Lambda_S)^{-1}\Lambda_S\xi\right\|_2\\ =&\frac{1}{\sigma^{2j}}\left\|(I+\frac{n}{\sigma^2}\Lambda_S)^{-1+\frac{\gamma}{2}}\Lambda_S^{1-\frac{\gamma}{2}}\left(A(I+\frac{n}{\sigma^2}\Lambda_S)^{-1+\gamma}\Lambda_S^{1-\gamma}\right)^{j-1}A(I+\frac{n}{\sigma^2}\Lambda_S)^{-1+\frac{\gamma}{2}}\Lambda_S^{-\frac{\gamma}{2}}\Lambda_S\xi\right\|_2\\ \le&n^{(1-t)(-1+\frac{\gamma}{2}+(-1+\gamma)(j-1))}\tilde{O}(n^{\frac{j(1+\alpha+2\tau-(1+2\tau+2\alpha)t)}{2\alpha}-j\gamma(1-t)})\|(I+\frac{n}{\sigma^2}\Lambda_S)^{-1+\frac{\gamma}{2}}\Lambda_S^{1-\frac{\gamma}{2}}\xi\|_2\\ =&\tilde{O}(n^{-\frac{\gamma}{2}(1-t)+\frac{(1-\alpha+2\tau-(1+2\tau)t)j}{2\alpha}})\|(I+\frac{n}{\sigma^2}\Lambda_S)^{-1+\frac{\gamma}{2}}\Lambda_S^{1-\frac{\gamma}{2}}\|_2\|\xi\|_2\\ =&\tilde{O}(n^{-\frac{\gamma}{2}(1-t)+\frac{(1-\alpha+2\tau-(1+2\tau)t)j}{2\alpha}})O(n^{(-1+\gamma/2)(1-t)})\\ =&\tilde{O}(n^{-(1-t)+\frac{(1-\alpha+2\tau-(1+2\tau)t)j}{2\alpha}}).\end{aligned} \tag{86}$$

Using (84), (85) and (86), we have

$$\begin{aligned}&\|(I+\tfrac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}\Lambda_S\xi\|_2\\ =&\left(\tilde{O}(n^{-(1-t)})+\sum_{j=1}^{\infty}\tilde{O}(n^{-1+\frac{(1-\alpha+2\tau-(1+2\tau)t)j}{2\alpha}})\right)\\ =&\left(\tilde{O}(n^{-(1-t)})+\tilde{O}(n^{-1+\frac{1-\alpha+2\tau-(1+2\tau)t}{2\alpha}})\right)\\ =&\tilde{O}(n^{-(1-t)}).\end{aligned} \tag{87}$$

By Corollary 24, with probability of at least $1-\delta$, we have

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

From (83) we get $\|(I+\frac{1}{\sigma^2}\Phi_S\Lambda_S\Phi_S^T)^{-1}f_S(\mathbf{x})\|_2=\tilde{O}(\sqrt{(\frac{1}{\delta}+1)n}\cdot n^{-(1-t)})$. This concludes the proof. □

Lemma 17. *Assume $\sigma^2=\Theta(n^t)$ where $1-\frac{\alpha}{1+2\tau}<t<1$. Let $\delta=n^{-q}$ where $0\le q<\frac{[\alpha-(1+2\tau)(1-t)](2\beta-1)}{4\alpha^2}$. Under Assumptions 1, 2 and 3, assume that $\mu_0=0$. Let $R=n^{(\frac{1}{\alpha}+\kappa)(1-t)}$ where $0<\kappa<\frac{\alpha-1-2\tau+(1+2\tau)t}{2\alpha^2(1-t)}$. Then with probability of at least $1-6\delta$ over sample inputs $(x_i)_{i=1}^n$, we have $G_2(D_n)=\frac{(1+o(1))}{2\sigma^2}\|(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\boldsymbol{\mu}_R\|_2^2=\frac{1}{\sigma^2}\Theta(n^{\max\{-2(1-t),\frac{(1-2\beta)(1-t)}{\alpha}\}}\log^{k/2}n)$, where $k=\begin{cases}0, & 2\alpha\neq 2\beta-1,\\ 1, & 2\alpha=2\beta-1.\end{cases}$*

Proof of Lemma 17. Let $S=n^D$. 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 9, when $S>n^{\max\{1,\frac{-t}{(\alpha-1-2\tau)}\}}$ with probability of at least $1-3\delta$ we have that

$$\begin{aligned}&|G_2(D_n)-G_{2,S}(D_n)|=|\mathbb{E}_{(x_{n+1},y_{n+1})}[T_2(D_{n+1})-T_{2,S}(D_{n+1})]-[T_2(D_n)-T_{2,S}(D_n)]|\\ &=\left|\mathbb{E}_{(x_{n+1},y_{n+1})}\tilde{O}\Big((\tfrac{1}{\delta}+1)\tfrac{1}{\sigma^2}(n+1)S^{\max\{1/2-\beta,1-\alpha+2\tau\}}\Big)-\tilde{O}\Big((\tfrac{1}{\delta}+1)\tfrac{1}{\sigma^2}nS^{\max\{1/2-\beta,1-\alpha+2\tau\}}\Big)\right|\\ &=\tilde{O}\Big((\tfrac{1}{\delta}+1)\tfrac{1}{\sigma^2}nS^{\max\{1/2-\beta,1-\alpha+2\tau\}}\Big) \qquad (88)\end{aligned}$$

(89)

Let $\Lambda_{1:S}=\mathrm{diag}\{\lambda_1,\dots,\lambda_S\}$, $\Phi_{1:S}=(\phi_1(\mathbf{x}),\phi_1(\mathbf{x}),\dots,\phi_S(\mathbf{x}))$ and $\boldsymbol{\mu}_{1:S}=(\mu_1,\dots,\mu_S)$. Since $\mu_0=0$, we have $T_{2,S}(D_n)=\frac{1}{2\sigma^2}\boldsymbol{\mu}_{1:S}^T\Phi_{1:S}^T(I+\frac{1}{\sigma^2}\Phi_{1:S}\Lambda_{1:S}\Phi_{1:S}^T)^{-1}\Phi_{1:S}\boldsymbol{\mu}_{1:S}$. Define $\eta_{1:S}=(\phi_1(x_{n+1}),\dots,\phi_S(x_{n+1}))^T$ and $\widetilde{\Phi}_{1:S}=(\Phi_{1:S}^T,\eta_{1:S})^T$. In the proof of Lemma 10, we showed that

$$\begin{aligned}T_{2,S}(D_n)&=\frac{1}{2\sigma^2}\boldsymbol{\mu}_{1:S}^T\Phi_{1:S}^T(I+\frac{1}{\sigma^2}\Phi_{1:S}\Lambda_{1:S}\Phi_{1:S}^T)^{-1}\Phi_{1:S}\boldsymbol{\mu}_{1:S}\\ &=\frac{1}{2}\boldsymbol{\mu}_{1:S}^T\Lambda_{1:S}^{-1}\boldsymbol{\mu}_{1:S}-\frac{1}{2}\boldsymbol{\mu}_{1:S}^T\Lambda_{1:S}^{-1}(I+\frac{1}{\sigma^2}\Lambda_{1:S}\Phi_{1:S}^T\Phi_{1:S})^{-1}\boldsymbol{\mu}_{1:S}.\end{aligned}$$

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}\boldsymbol{\mu}_{1:S}^T\Lambda_{1:S}^{-1}\boldsymbol{\mu}_{1:S}-\frac{1}{2}\boldsymbol{\mu}_{1:S}^T\Lambda_{1:S}^{-1}(I+\frac{1}{\sigma^2}\Lambda_{1:S}\widetilde{\Phi}_S^T\widetilde{\Phi}_S)^{-1}\boldsymbol{\mu}_{1:S}\right)\\ &-\left(\frac{1}{2}\boldsymbol{\mu}_{1:S}^T\Lambda_{1:S}^{-1}\boldsymbol{\mu}_{1:S}-\frac{1}{2}\boldsymbol{\mu}_{1:S}^T\Lambda_{1:S}^{-1}(I+\frac{1}{\sigma^2}\Lambda_{1:S}\Phi_{1:S}^T\Phi_{1:S})^{-1}\boldsymbol{\mu}_{1:S})\right)\\ &=\mathbb{E}_{(x_{n+1},y_{n+1})}\left(\frac{1}{2}\boldsymbol{\mu}_{1:S}^T\Lambda_{1:S}^{-1}(I+\frac{1}{\sigma^2}\Lambda_{1:S}\Phi_{1:S}^T\Phi_{1:S})^{-1}\boldsymbol{\mu}_{1:S}-\frac{1}{2}\boldsymbol{\mu}_{1:S}^T\Lambda_{1:S}^{-1}(I+\frac{1}{\sigma^2}\Lambda_{1:S}\widetilde{\Phi}_S^T\widetilde{\Phi}_S)^{-1}\boldsymbol{\mu}_{1:S}\right)\\ &=\mathbb{E}_{(x_{n+1},y_{n+1})}\left(\frac{1}{2\sigma^2}\boldsymbol{\mu}_{1:S}^T\Lambda_{1:S}^{-1}\frac{(I+\frac{1}{\sigma^2}\Lambda_{1:S}\Phi_{1:S}^T\Phi_{1:S})^{-1}\Lambda_{1:S}\eta_{1:S}\eta_{1:S}^T(I+\frac{1}{\sigma^2}\Lambda_{1:S}\Phi_{1:S}^T\Phi_{1:S})^{-1}}{1+\frac{1}{\sigma^2}\eta_{1:S}^T(I+\frac{1}{\sigma^2}\Lambda_{1:S}\Phi_{1:S}^T\Phi_{1:S})^{-1}\Lambda_{1:S}\eta_{1:S}}\boldsymbol{\mu}_{1:S})\right)\\ &=\mathbb{E}_{(x_{n+1},y_{n+1})}\left(\frac{1}{2\sigma^2}\frac{\boldsymbol{\mu}_{1:S}^T(I+\frac{1}{\sigma^2}\Phi_{1:S}^T\Phi_{1:S}\Lambda_{1:S})^{-1}\eta_{1:S}\eta_{1:S}^T(I+\frac{1}{\sigma^2}\Lambda_{1:S}\Phi_{1:S}^T\Phi_{1:S})^{-1}\boldsymbol{\mu}_{1:S}}{1+\frac{1}{\sigma^2}\eta_{1:S}^T(I+\frac{1}{\sigma^2}\Lambda_{1:S}\Phi_{1:S}^T\Phi_{1:S})^{-1}\Lambda_{1:S}\eta_{1:S}})\right)\\ &=\mathbb{E}_{(x_{n+1},y_{n+1})}\left(\frac{1+o(1)}{2\sigma^2}\boldsymbol{\mu}_{1:S}^T(I+\frac{1}{\sigma^2}\Phi_{1:S}^T\Phi_{1:S}\Lambda_{1:S})^{-1}\eta_{1:S}\eta_{1:S}^T(I+\frac{1}{\sigma^2}\Lambda_{1:S}\Phi_{1:S}^T\Phi_{1:S})^{-1}\boldsymbol{\mu}_{1:S}\right)\\ &=\frac{1+o(1)}{2\sigma^2}\boldsymbol{\mu}_{1:S}^T(I+\frac{1}{\sigma^2}\Phi_{1:S}^T\Phi_{1:S}\Lambda_{1:S})^{-1}(I+\frac{1}{\sigma^2}\Lambda_{1:S}\Phi_{1:S}^T\Phi_{1:S})^{-1}\boldsymbol{\mu}_{1:S}\\ &=\frac{1+o(1)}{2\sigma^2}\|(I+\frac{1}{\sigma^2}\Lambda_{1:S}\Phi_{1:S}^T\Phi_{1:S})^{-1}\boldsymbol{\mu}_{1:S}\|_2^2,\end{aligned} \qquad (90)$$

where in the fourth to last equality we used the Sherman–Morrison formula, in the third inequality we used (73) , and in the last equality we used the fact that $\mathbb{E}_{(x_{n+1},y_{n+1})}\eta_{1:S}\eta_{1:S}^T = I$.

Let $\hat{\boldsymbol{\mu}}_{1:R} = (\mu_1, \dots, \mu_R, 0, \dots, 0) \in \mathbb{R}^S$. Then we have

$$\begin{aligned}
\|(I + \frac{1}{\sigma^2}\Lambda_{1:S}\Phi_{1:S}^T\Phi_{1:S})^{-1}\boldsymbol{\mu}_{1:S}\|_2 &\le \|(I + \frac{1}{\sigma^2}\Lambda_{1:S}\Phi_{1:S}^T\Phi_{1:S})^{-1}\hat{\boldsymbol{\mu}}_{1:R}\|_2 + \|(I + \frac{1}{\sigma^2}\Lambda_{1:S}\Phi_{1:S}^T\Phi_{1:S})^{-1}(\boldsymbol{\mu}_{1:S} - \hat{\boldsymbol{\mu}}_{1:R})\|_2, \\
\|(I + \frac{1}{\sigma^2}\Lambda_{1:S}\Phi_{1:S}^T\Phi_{1:S})^{-1}\boldsymbol{\mu}_{1:S}\|_2 &\ge \|(I + \frac{1}{\sigma^2}\Lambda_{1:S}\Phi_{1:S}^T\Phi_{1:S})^{-1}\hat{\boldsymbol{\mu}}_{1:R}\|_2 - \|(I + \frac{1}{\sigma^2}\Lambda_{1:S}\Phi_{1:S}^T\Phi_{1:S})^{-1}(\boldsymbol{\mu}_{1:S} - \hat{\boldsymbol{\mu}}_{1:R})\|_2.
\end{aligned} \tag{91}$$

Let $R = n^{(\frac{1}{\alpha}+\kappa)(1-t)}$ where $0 < \kappa < \frac{\alpha-1-2\tau+(1+2\tau)t}{2\alpha^2(1-t)}$. In Lemma 33, (136), we showed that with probability of at least $1-\delta$,

$$\begin{aligned}
\|(I + \tfrac{1}{\sigma^2}\Lambda_{1:R}\Phi_{1:R}^T\Phi_{1:R})^{-1}\boldsymbol{\mu}_{1:R}\|_2 &= \Theta(n^{(1-t)\max\{-1, \frac{1-2\beta}{2\alpha}\}}\log^{k/2} n) \\
&= (1+o(1))\|(I + \frac{n}{\sigma^2}\Lambda_{1:R})^{-1}\boldsymbol{\mu}_{1:R}\|_2,
\end{aligned} \tag{92}$$

where $k = \begin{cases} 0, & 2\alpha \neq 2\beta - 1, \\ 1, & 2\alpha = 2\beta - 1. \end{cases}$. The same proof holds if we replace $\Phi_{1:R}$ with $\Phi_{1:S}$, $\Lambda_{1:R}$ with $\Lambda_{1:S}$, and $\boldsymbol{\mu}_{1:R}$ with $\hat{\boldsymbol{\mu}}_{1:R}$. We have

$$\begin{aligned}
\|(I + \tfrac{1}{\sigma^2}\Lambda_{1:S}\Phi_{1:S}^T\Phi_{1:S})^{-1}\hat{\boldsymbol{\mu}}_{1:R}\|_2 &= \Theta(n^{(1-t)\max\{-1, \frac{1-2\beta}{2\alpha}\}}\log^{k/2} n) \\
&= (1+o(1))\|(I + \frac{n}{\sigma^2}\Lambda_{1:S})^{-1}\hat{\boldsymbol{\mu}}_{1:R}\|_2.
\end{aligned} \tag{93}$$

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

$$\begin{aligned}
&|\xi^T(I + \frac{1}{\sigma^2}\Lambda_{1:S}\Phi_{1:S}^T\Phi_{1:S})^{-1}(\boldsymbol{\mu}_{1:S} - \hat{\boldsymbol{\mu}}_{1:R})| \\
&= |\xi^T\left(I - \frac{1}{\sigma^2}\Lambda_{1:S}\Phi_{1:S}^T(I + \frac{1}{\sigma^2}\Phi_{1:S}\Lambda_{1:S}\Phi_{1:S}^T)^{-1}\Phi_{1:S}\right)(\boldsymbol{\mu}_{1:S} - \hat{\boldsymbol{\mu}}_{1:R})| \\
&= |\xi^T(\boldsymbol{\mu}_{1:S} - \hat{\boldsymbol{\mu}}_{1:R}) - \frac{1}{\sigma^2}\xi^T\Lambda_{1:S}\Phi_{1:S}^T(I + \frac{1}{\sigma^2}\Phi_{1:S}\Lambda_{1:S}\Phi_{1:S}^T)^{-1}\Phi_{1:S}(\boldsymbol{\mu}_{1:S} - \hat{\boldsymbol{\mu}}_{1:R})| \\
&\le \|\xi\|_2\|\boldsymbol{\mu}_{1:S} - \hat{\boldsymbol{\mu}}_{1:R}\|_2 + \frac{1}{\sigma^2}|\xi^T\Lambda_{1:S}\Phi_{1:S}^T(I + \frac{1}{\sigma^2}\Phi_{1:S}\Lambda_{1:S}\Phi_{1:S}^T)^{-1}\Phi_{1:S}(\boldsymbol{\mu}_{1:S} - \hat{\boldsymbol{\mu}}_{1:R})| \\
&\le O(R^{\frac{1-2\beta}{2}}) + \frac{1}{\sigma^2}\|(I + \frac{1}{\sigma^2}\Phi_{1:S}\Lambda_{1:S}\Phi_{1:S}^T)^{-1}\Phi_{1:S}\Lambda_{1:S}\xi\|_2\|\Phi_{1:S}(\boldsymbol{\mu}_{1:S} - \hat{\boldsymbol{\mu}}_{1:R})\|_2 \\
&= O(R^{\frac{1-2\beta}{2}}) + \frac{1}{\sigma^2}O(\sqrt{(\frac{1}{\delta}+1)n}\cdot n^{-(1-t)})O(\sqrt{(\frac{1}{\delta}+1)n}R^{\frac{1-2\beta}{2}}) \\
&= O((\frac{1}{\delta}+1)R^{\frac{1-2\beta}{2}}),
\end{aligned}$$

where in the second to last step we used Corollary 24 to show $\|\Phi_{1:S}(\boldsymbol{\mu}_{1:S} - \hat{\boldsymbol{\mu}}_{1:R})\|_2 = O(\sqrt{(\frac{1}{\delta}+1)n}R^{\frac{1-2\beta}{2}})$ with probability of at least $1-\delta$, and Lemma 16 to show that $\|(I + \frac{1}{\sigma^2}\Phi_{1:S}\Lambda_{1:S}\Phi_{1:S}^T)^{-1}\Phi_{1:S}\Lambda_{1:S}\xi\|_2 = O(\sqrt{(\frac{1}{\delta}+1)n}\cdot n^{-1})$ with probability of at least $1-\delta$. Since $R = n^{(\frac{1}{\alpha}+\kappa)(1-t)}$, we have

$$|\xi^T(I + \frac{1}{\sigma^2}\Lambda_{1:S}\Phi_{1:S}^T\Phi_{1:S})^{-1}(\boldsymbol{\mu}_{1:S} - \hat{\boldsymbol{\mu}}_{1:R})| = O((\frac{1}{\delta}+1)n^{\frac{(1-2\beta)(1-t)}{2\alpha} + \frac{(1-2\beta)(1-t)\kappa}{2}}).$$

Since ξ is arbitrary, we have $\|(I + \frac{1}{\sigma^2}\Lambda_{1:S}\Phi_{1:S}^T\Phi_{1:S})^{-1}(\boldsymbol{\mu}_{1:S} - \hat{\boldsymbol{\mu}}_{1:R})\|_2 = O((\frac{1}{\delta}+1)n^{\frac{(1-2\beta)(1-t)}{2\alpha} + \frac{(1-2\beta)(1-t)\kappa}{2}})$. Since $0 \le q < \frac{[\alpha-(1+2\tau)(1-t)](2\beta-1)}{4\alpha^2}$ and $0 < \kappa < \frac{\alpha-1-2\tau+(1+2\tau)t}{2\alpha^2(1-t)}$, we can choose $\kappa < \frac{\alpha-1-2\tau+(1+2\tau)t}{2\alpha^2(1-t)}$ and κ is arbitrarily close to $\kappa < \frac{\alpha-1-2\tau+(1+2\tau)t}{2\alpha^2(1-t)}$ such that

$0 \le q < \frac{(2\beta-1)(1-t)\kappa}{2}$. Then we have $\frac{(1-2\beta)(1-t)\kappa}{2} + q < 0$. From (91) and (93), we have

$$\begin{aligned}
\|(I + \frac{1}{\sigma^2}\Lambda_{1:S}\Phi_{1:S}^T\Phi_{1:S})^{-1}\boldsymbol{\mu}_{1:S}\|_2 =& \Theta(n^{\max\{-(1-t), \frac{(1-2\beta)(1-t)}{2\alpha}\}}\log^{k/2} n) + O((\frac{1}{\delta}+1)n^{\frac{(1-2\beta)(1-t)}{2\alpha}+\frac{(1-2\beta)(1-t)\kappa}{2}}) \\
=& \Theta(n^{\max\{-(1-t), \frac{(1-2\beta)(1-t)}{2\alpha}\}}\log^{k/2} n) + O((n^{q+\frac{(1-2\beta)(1-t)}{2\alpha}+\frac{(1-2\beta)(1-t)\kappa}{2}}) \\
=& \Theta(n^{\max\{-(1-t), \frac{(1-2\beta)(1-t)}{2\alpha}\}}\log^{k/2} n) \\
=& (1+o(1))\|(I + \frac{n}{\sigma^2}\Lambda_{1:S})^{-1}\hat{\boldsymbol{\mu}}_{1:R}\|_2 \\
=& (1+o(1))\|(I + \frac{n}{\sigma^2}\Lambda_R)^{-1}\boldsymbol{\mu}_R\|_2.
\end{aligned} \tag{94}$$

Hence $G_{2,S}(D_n) = \frac{1+o(1)}{2\sigma^2}\|(I + \frac{1}{\sigma^2}\Lambda_{1:S}\Phi_{1:S}^T\Phi_{1:S})^{-1}\boldsymbol{\mu}_{1:S}\|_2^2 = \frac{1}{\sigma^2}\Theta(n^{(1-t)\max\{-2, \frac{1-2\beta}{\alpha}\}}\log^{k/2} n)$. Then by (88), we have

$$G_2(D_n) = \frac{1}{\sigma^2}\Theta(n^{\max\{-2(1-t), \frac{(1-2\beta)(1-t)}{\alpha}\}}\log^{k/2} n) + \tilde{O}\bigg((\frac{1}{\delta}+1)\frac{n}{\sigma^2}S^{\max\{1/2-\beta, 1-\alpha+2\tau\}}\bigg).$$

Choosing $S = n^{\max\left\{1, \frac{-t}{(\alpha-1-2\tau)}, \left(\frac{1+q+\min\{2, \frac{2\beta-1}{\alpha}\}}{\min\{\beta-1/2, \alpha-1-2\tau\}}+1\right)(1-t)\right\}}$, we get the result. □

Proof of Theorem 5, $\mu_0 = 0$. From Lemmas 15 and 17 and $\frac{1}{\alpha} - 1 > -2$, we have that with probability of at least $1 - 7\tilde{\delta}$,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\epsilon}} G(D_n) =& \frac{1+o(1)}{2\sigma^2}(\mathrm{Tr}(I + \frac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R - \|\Lambda_R^{1/2}(I + \frac{n}{\sigma^2}\Lambda_R)^{-1}\|_F^2 + \|(I + \frac{n}{\sigma^2}\Lambda_R)^{-1}\boldsymbol{\mu}_R\|_2^2) \\
=& \frac{1}{\sigma^2}\Theta(n^{\frac{(1-\alpha)(1-t)}{\alpha}}) + \frac{1}{\sigma^2}\Theta(n^{\max\{-2(1-t), \frac{(1-2\beta)(1-t)}{\alpha}\}}\log^{k/2} n) \\
=& \frac{1}{\sigma^2}\Theta(n^{\max\{\frac{(1-\alpha)(1-t)}{\alpha}, \frac{(1-2\beta)(1-t)}{\alpha}\}})
\end{aligned} \tag{95}$$

where $k = \begin{cases} 0, & 2\alpha \neq 2\beta - 1 \\ 1, & 2\alpha = 2\beta - 1 \end{cases}$, and $R = n^{(\frac{1}{\alpha}+\kappa)(1-t)}$, $\kappa > 0$.

Furthermore, we have

$$\begin{aligned}
&\mathrm{Tr}(I + \frac{n}{\sigma^2}\Lambda)^{-1}\Lambda - \mathrm{Tr}(I + \frac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R \\
&= \sum_{p=R+1}^{\infty}\frac{\lambda_p}{1+\frac{n}{\sigma^2}\lambda_p} \le \sum_{p=R+1}^{\infty}\frac{C_\lambda p^{-\alpha}}{1+\frac{n}{\sigma^2}C_\lambda p^{-\alpha}} \le \sum_{p=R+1}^{\infty} C_\lambda p^{-\alpha} = \frac{n}{\sigma^2}O(R^{1-\alpha}) \\
&= O(n^{(1-\alpha)(1-t)(\frac{1}{\alpha}+\kappa)}) \\
&= o(n^{\frac{(1-\alpha)(1-t)}{\alpha}}).
\end{aligned}$$

Then we have

$$\mathrm{Tr}(I + \frac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R = \mathrm{Tr}(I + \frac{n}{\sigma^2}\Lambda)^{-1}\Lambda(1+o(1)). \tag{96}$$

Similarly we can prove

$$\|\Lambda_R^{1/2}(I + \frac{n}{\sigma^2}\Lambda_R)^{-1}\|_F^2 = \|\Lambda^{1/2}(I + \frac{n}{\sigma^2}\Lambda)^{-1}\|_F^2(1+o(1)) \tag{97}$$

$$\|(I + \frac{n}{\sigma^2}\Lambda_R)^{-1}\boldsymbol{\mu}_R\|_2^2 = \|(I + \frac{n}{\sigma^2}\Lambda)^{-1}\boldsymbol{\mu}\|_2^2(1+o(1)) \tag{98}$$

Letting $\delta = 7\tilde{\delta}$, the proof is complete. □

In the case of $\mu_0 > 0$, we have the following lemma:

Lemma 18. *Let $\delta = n^{-q}$ where $0 \le q < \frac{[\alpha-(1+2\tau)(1-t)](2\beta-1)}{4\alpha^2}$. Under Assumptions 1, 2 and 3, assume that $\mu_0 > 0$. Then with probability of at least $1 - 6\delta$ over sample inputs $(x_i)_{i=1}^n$, we have $G_2(D_n) = \frac{1}{2\sigma^2}\mu_0^2 + o(1)$.*

Proof of Lemma 18. Let $S = n^D$. 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 9, when $S > n^{\max\{1,\frac{-t}{(\alpha-1-2\tau)}\}}$, with probability of at least $1-3\delta$ we have that

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

Let $\Lambda_S = \mathrm{diag}\{\lambda_1,\ldots,\lambda_S\}$, $\Phi_S = (\phi_1(\mathbf{x}),\phi_1(\mathbf{x}),\ldots,\phi_S(\mathbf{x}))$ and $\boldsymbol{\mu}_S = (\mu_1,\ldots,\mu_S)$. Define $\eta_S = (\phi_0(x_{n+1}),\phi_1(x_{n+1}),\ldots,\phi_S(x_{n+1}))^T$ and $\widetilde{\Phi}_S = (\Phi_S^T,\eta_S)^T$. By the same technique as in the proof of Lemma 10, we replace Λ_R by $\tilde{\Lambda}_{\epsilon,R} = \mathrm{diag}\{\epsilon,\lambda_1,\ldots,\lambda_R\}$, let $\epsilon\to 0$ and show the counterpart of the result (90) in the proof of Lemma 17:

$$
\begin{aligned}
G_{2,S}(D_n) &= \mathbb{E}_{(x_{n+1},y_{n+1})}(T_{2,S}(D_{n+1}) - T_{2,S}(D_n)) \\
&= \mathbb{E}_{(x_{n+1},y_{n+1})}\left(\frac{1}{2\sigma^2}\frac{\boldsymbol{\mu}_S^T(I+\frac{1}{\sigma^2}\Phi_S^T\Phi_S\Lambda_S)^{-1}\eta_S\eta_S^T(I+\frac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}\boldsymbol{\mu}_S}{1+\frac{1}{\sigma^2}\eta_S^T(I+\frac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}\Lambda_S\eta_S})\right) \\
&= \mathbb{E}_{(x_{n+1},y_{n+1})}\left(\frac{1+o(1)}{2\sigma^2}\boldsymbol{\mu}_S^T(I+\frac{1}{\sigma^2}\Phi_S^T\Phi_S\Lambda_S)^{-1}\eta_S\eta_S^T(I+\frac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}\boldsymbol{\mu}_S\right) \\
&= \frac{1+o(1)}{2\sigma^2}\boldsymbol{\mu}_S^T(I+\frac{1}{\sigma^2}\Phi_S^T\Phi_S\Lambda_S)^{-1}(I+\frac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}\boldsymbol{\mu}_S \\
&= \frac{1+o(1)}{2\sigma^2}\|(I+\frac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}\boldsymbol{\mu}_S\|_2^2,
\end{aligned}
\tag{99}
$$

where in the fourth to last equality we used the Sherman–Morrison formula, in the third inequality we used (73) , and in the last equality we used the fact that $\mathbb{E}_{(x_{n+1},y_{n+1})}\eta_{1:S}\eta_{1:S}^T = I$.

Let $\hat{\boldsymbol{\mu}}_R = (\mu_0,\mu_1,\ldots,\mu_R,0,\ldots,0)\in\mathbb{R}^S$. Then we have

$$
\begin{aligned}
\|(I+\frac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}\boldsymbol{\mu}_S\|_2 &\le \|(I+\frac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}\hat{\boldsymbol{\mu}}_R\|_2 + \|(I+\frac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}(\boldsymbol{\mu}_S-\hat{\boldsymbol{\mu}}_R)\|_2, \\
\|(I+\frac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}\boldsymbol{\mu}_S\|_2 &\ge \|(I+\frac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}\hat{\boldsymbol{\mu}}_R\|_2 - \|(I+\frac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}(\boldsymbol{\mu}_S-\hat{\boldsymbol{\mu}}_R)\|_2.
\end{aligned}
\tag{100}
$$

Choose $R = n^{(\frac{1}{\alpha}+\kappa)(1-t)}$ where $0<\kappa<\frac{\alpha-1-2\tau+(1+2\tau)t}{\alpha^2(1-t)}$. In Lemma 33, (136), we showed that with probability of at least $1-\delta$,

$$
\begin{aligned}
\|(I+\tfrac{1}{\sigma^2}\Lambda_{1:R}\Phi_{1:R}^T\Phi_{1:R})^{-1}\boldsymbol{\mu}_{1:R}\|_2 &= \Theta(n^{(1-t)\max\{-1,\frac{1-2\beta}{2\alpha}\}}\log^{k/2}n) \\
&= (1+o(1))\|(I+\frac{n}{\sigma^2}\Lambda_{1:R})^{-1}\boldsymbol{\mu}_{1:R}\|_2,
\end{aligned}
\tag{101}
$$

where $k = \begin{cases}0, & 2\alpha\neq 2\beta-1,\\ 1, & 2\alpha = 2\beta-1.\end{cases}$. The same proof holds if we replace $\Phi_{1:R}$ with $\Phi_{1:S}$, $\Lambda_{1:R}$ with $\Lambda_{1:S}$, and $\boldsymbol{\mu}_{1:R}$ with $\hat{\boldsymbol{\mu}}_{1:R}$. We have

$$
\begin{aligned}
\|(I+\tfrac{1}{\sigma^2}\Lambda_{1:S}\Phi_{1:S}^T\Phi_{1:S})^{-1}\hat{\boldsymbol{\mu}}_{1:R}\|_2 &= \Theta(n^{(1-t)\max\{-1,\frac{1-2\beta}{2\alpha}\}}\log^{k/2}n) \\
&= (1+o(1))\|(I+\frac{n}{\sigma^2}\Lambda_{1:S})^{-1}\hat{\boldsymbol{\mu}}_{1:R}\|_2.
\end{aligned}
\tag{102}
$$

So we have

$$
\begin{aligned}
\|(I+\tfrac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}\hat{\boldsymbol{\mu}}_R\|_2 &= \mu_0 + \Theta(n^{(1-t)\max\{-1,\frac{1-2\beta}{2\alpha}\}}\log^{k/2}n) \\
&= \mu_0 + o(1).
\end{aligned}
\tag{103}
$$

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

at least $1-2\delta$ we have

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

where in the second to last step we used Corollary 24 to show $\|\Phi_S(\boldsymbol{\mu}_S-\hat{\boldsymbol{\mu}}_R)\|_2 = O(\sqrt{(\frac{1}{\delta}+1)n}R^{\frac{1-2\beta}{2}})$ with probability of at least $1-\delta$, and Lemma 16 to show that $\|(I+\frac{1}{\sigma^2}\Phi_S\Lambda_S\Phi_S^T)^{-1}\Phi_S\Lambda_S\xi\|_2 = O(\sqrt{(\frac{1}{\delta}+1)n}\cdot n^{-(1-t)})$ with probability of at least $1-\delta$. Since $R=n^{(\frac{1}{\alpha}+\kappa)(1-t)}$, we have

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

Since ξ is arbitrary, we have $\|(I+\frac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}(\boldsymbol{\mu}_S-\hat{\boldsymbol{\mu}}_R)\|_2 = O((\frac{1}{\delta}+1)n^{\frac{(1-2\beta)(1-t)}{2\alpha}+\frac{(1-2\beta)(1-t)\kappa}{2}})$. Since $0\le q<\frac{[\alpha-(1+2\tau)(1-t)](2\beta-1)}{4\alpha^2}$ and $0<\kappa<\frac{\alpha-1-2\tau+(1+2\tau)t}{2\alpha^2(1-t)}$, we can choose $\kappa<\frac{\alpha-1-2\tau+(1+2\tau)t}{2\alpha^2(1-t)}$ and κ is arbitrarily close to $\kappa<\frac{\alpha-1-2\tau+(1+2\tau)t}{2\alpha^2(1-t)}$ such that $0\le q<\frac{(2\beta-1)(1-t)\kappa}{2}$. Then we have $\frac{(1-2\beta)(1-t)\kappa}{2}+q<0$. From (100) and (103), we have

$$\begin{aligned}
\|(I+\tfrac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}\boldsymbol{\mu}_S\|_2 =&\mu_0+\Theta(n^{(1-t)\max\{-1,\frac{1-2\beta}{2\alpha}\}}\log^{k/2}n)+O((\frac{1}{\delta}+1)n^{\frac{(1-2\beta)(1-t)}{2\alpha}+\frac{(1-2\beta)(1-t)\kappa}{2}})\\
=&\mu_0+\Theta(n^{(1-t)\max\{-1,\frac{1-2\beta}{2\alpha}\}}\log^{k/2}n)\\
=&\mu_0+o(1).
\end{aligned}\tag{104}$$

Hence $G_{2,S}(D_n)=\frac{1+o(1)}{2\sigma^2}\|(I+\frac{1}{\sigma^2}\Lambda_S\Phi_S^T\Phi_S)^{-1}\boldsymbol{\mu}_S\|_2^2=\frac{1}{2\sigma^2}\mu_0^2+o(1)$. Then by (99), $G_2(D_n)=\frac{1}{2\sigma^2}\mu_0^2+o(1)+\tilde{O}\big((\frac{1}{\delta}+1)nS^{\max\{1/2-\beta,1-\alpha\}}\big)$. Choosing $S=n^{\max\Big\{1,\frac{-t}{(\alpha-1-2\tau)},\Big(\frac{1+q+\min\{2,\frac{2\beta-1}{\alpha}\}}{\min\{\beta-1/2,\alpha-1-2\tau\}}+1\Big)(1-t)\Big\}}$, we get the result. □

Proof of Theorem 5, $\mu_0>0$. According to Lemma 18, $G_2(D_n)=\frac{1}{2\sigma^2}\mu_0^2+o(1)$. By Lemma 15, we have $G_1(D_n)=\Theta(n^{\frac{(1-\alpha)(1-t)}{\alpha}})$. Then $\mathbb{E}_\epsilon G(D_n)=G_1(D_n)+G_2(D_n)=\frac{1}{2\sigma^2}\mu_0^2+o(1)$. □

E Helper lemmas

Lemma 19. *Assume that $m\to\infty$ as $n\to\infty$. Given constants $a_1,a_2,s_1,s_2>0$, if $s_1>1$ and $s_2s_3>s_1-1$, we have that*

$$\sum_{i=1}^{R}\frac{a_1i^{-s_1}}{(1+a_2mi^{-s_2})^{s_3}}=\Theta(m^{\frac{1-s_1}{s_2}}).\tag{105}$$

If $s_1>1$ and $s_2s_3=s_1-1$, we have that

$$\sum_{i=1}^{R}\frac{a_1i^{-s_1}}{(1+a_2mi^{-s_2})^{s_3}}=\Theta(m^{-s_3}\log m).\tag{106}$$

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

$$\sum_{i=1}^{R} \frac{a_1 i^{-s_1}}{(1 + a_2 m i^{-s_2})^{s_3}} = \Theta(m^{-s_3}). \tag{107}$$

Overall, if $s_1 > 1$ and $m \to \infty$,

$$\sum_{i=1}^{R} \frac{a_1 i^{-s_1}}{(1 + a_2 m i^{-s_2})^{s_3}} = \begin{cases} \Theta(m^{\max\{-s_3, \frac{1-s_1}{s_2}\}}), & s_2 s_3 \neq s_1 - 1, \\ \Theta(m^{\frac{1-s_1}{s_2}} \log m), & s_2 s_3 = s_1 - 1. \end{cases} \tag{108}$$

Proof of Lemma 19. First, when $s_1 > 1$ and $s_2 s_3 > s_1 - 1$, we have that

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

On the other hand, we have

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

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

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

On the other hand, we have

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

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

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

On the other hand, we have

$$\begin{aligned}\sum_{i=1}^{R} \frac{a_1 i^{-s_1}}{(1+a_2 m i^{-s_2})^{s_3}} &\le \frac{a_1}{(1+a_2 m)^{s_3}} + m^{\frac{1-s_1}{s_2}} \int_{[2/m^{1/s_2}, (R+1)/m^{1/s_2}]} \frac{a_1 x^{-s_1}}{(1+a_2 x^{-s_2})^{s_3}} \mathrm{d}x \\ &\le \frac{a_1}{(1+a_2 m)^{s_3}} + m^{\frac{1-s_1}{s_2}} \Theta(m^{(-1/s_2)(1-s_1+s_2 s_3)}) \\ &= \Theta(m^{-s_3}).\end{aligned}$$

Overall, if $s_1 > 1$,

$$\sum_{i=1}^{R} \frac{a_1 i^{-s_1}}{(1+a_2 m i^{-s_2})^{s_3}} = \begin{cases} \Theta(m^{\max\{-s_3, \frac{1-s_1}{s_2}\}}), & s_2 s_3 \neq s_1 - 1, \\ \Theta(m^{-s_3} \log n), & s_2 s_3 = s_1 - 1. \end{cases} \tag{109}$$

□

Lemma 20. *Assume that $R = m^{\frac{1}{s^2}+\kappa}$ for $\kappa > 0$. Given constants $a_1, a_2, s_1, s_2 > 0$, if $s_1 \le 1$, we have that*

$$\sum_{i=1}^{R} \frac{a_1 i^{-s_1}}{(1+a_2 m i^{-s_2})^{s_3}} = \tilde{O}(\max\{m^{-s_3}, R^{1-s_1}\}). \tag{110}$$

Proof of Lemma 20. First, when $s_1 \le 1$ and $s_2 s_3 > s_1 - 1$, we have that

$$\begin{aligned}\sum_{i=1}^{R} \frac{a_1 i^{-s_1}}{(1+a_2 m i^{-s_2})^{s_3}} &\le \frac{a_1}{(1+a_2 m)^{s_3}} + \int_{[1,R]} \frac{a_1 x^{-s_1}}{(1+a_2 m x^{-s_2})^{s_3}} \mathrm{d}x \\ &= \frac{a_1}{(1+a_2 m)^{s_3}} + m^{\frac{1-s_1}{s_2}} \int_{[1,R]} \frac{a_1 (\frac{x}{m^{1/s_2}})^{-s_1}}{(1+a_2 (\frac{x}{m^{1/s_2}})^{-s_2})^{s_3}} \mathrm{d} \frac{x}{m^{1/s_2}} \\ &= \frac{a_1}{(1+a_2 m)^{s_3}} + m^{\frac{1-s_1}{s_2}} \int_{[1/m^{1/s_2}, R/m^{1/s_2}]} \frac{a_1 x^{-s_1}}{(1+a_2 x^{-s_2})^{s3}} \mathrm{d}x \\ &= \frac{a_1}{(1+a_2 m)^{s_3}} + \tilde{O}(m^{\frac{1-s_1}{s_2}} (\frac{R}{m^{1/s_2}})^{1-s_1}) \\ &= \tilde{O}(\max\{m^{-s_3}, R^{1-s_1}\}).\end{aligned}$$

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

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

Overall, if $s_1 \le 1$,

$$\sum_{i=1}^{R} \frac{a_1 i^{-s_1}}{(1+a_2 m i^{-s_2})^{s_3}} = \tilde{O}(\max\{m^{-s_3}, R^{1-s_1}\}). \tag{111}$$

□

Lemma 21. *Assume that $f \in L^2(\Omega, \rho)$. Consider the random vector $f(\mathbf{x}) = (f(x_1), \ldots, f(x_n))^T$, where $x_1, \ldots, x_n$ are drawn i.i.d from ρ. Then with probability of at least $1 - \delta_1$, we have*

$$\|f(\mathbf{x})\|_2^2 = \sum_{i=1}^n f^2(x_i) = \tilde{O}\Big((\tfrac{1}{\delta_1} + 1)n\|f\|_2^2\Big),$$

where $\|f\|_2^2 = \int_{x\in\Omega} f^2(x)\mathrm{d}\rho(x)$.

Proof of Lemma 21. Given a positive number $C \geq \|f\|_2^2$, applying Markov's inequality we have

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

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

$$\mathbb{P}(A) \geq 1 - n\mathbb{P}(f^2(X) > C) \geq 1 - \frac{1}{C}n\|f\|_2^2. \tag{112}$$

Define $\bar{f}^2(x) = \min\{f^2(x), C\}$. Then $\mathbb{E}\bar{f}^2(X) \leq \mathbb{E}f^2(X) = \|f\|_2^2$. So $|\bar{f}^2(X) - \mathbb{E}\bar{f}^2(X)| \leq \max\{C, \|f\|_2^2\} = C$ Since $0 \leq \bar{f}^2(x) \leq C$, we have

$$\mathbb{E}(\bar{f}^4(X)) \leq C\mathbb{E}(\bar{f}^2(X)) \leq C\|f\|_2^2. \tag{113}$$

So we have

$$\mathbb{E}|\bar{f}^2(X) - \mathbb{E}\bar{f}^2(X)|^2 \leq \mathbb{E}(\bar{f}^4(X)) \leq C\|f\|_2^2. \tag{114}$$

Applying Bernstein's inequality, we have

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

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

$$\begin{aligned}
\sum_{i=1}^n \bar{f}^2(x_i) &\leq \max\left\{\sqrt{4C\log\frac{2}{\delta_1}n\|f\|_2^2}, \frac{4C}{3}\log\frac{2}{\delta_1}\right\} + n\mathbb{E}\bar{f}^2(X) \\
&\leq \max\left\{\sqrt{4C\log\frac{2}{\delta_1}n\|f\|_2^2}, \frac{4C}{3}\log\frac{2}{\delta_1}\right\} + n\|f\|_2^2.
\end{aligned} \tag{115}$$

When event A happens, $f^2(x_i) = \bar{f}^2(x_i)$ for all sample inputs. According to (112) and (115), with probability at least $1 - \frac{1}{C}n\|f\|_2^2 - \delta_1/2$, we have

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

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

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

□

Lemma 22. *Assume that $f \in L^2(\Omega, \rho)$. Consider the random vector $f(\mathbf{x}) = (f(x_1), \ldots, f(x_n))^T$, where $x_1, \ldots, x_n$ are drawn i.i.d from ρ. Assume that $\|f\|_\infty = \sup_{x\in\Omega} f(x) \leq C$. With probability of at least $1 - \delta_1$, we have*

$$\|f(\mathbf{x})\|_2^2 = \tilde{O}\left(\sqrt{C^2n\|f\|_2^2} + C^2\right) + n\|f\|_2^2,$$

where $\|f\|_2^2 = \int_{x\in\Omega} f^2(x)\mathrm{d}\rho(x)$.

Proof of Lemma 22. We have $|f^2(X) - \mathbb{E}f^2(X)| \leq \max\{C^2, \|f\|_2^2\} = C^2$ Since $0 \leq f^2(x) \leq C$, we have

$$\mathbb{E}(f^4(X)) \leq C^2\mathbb{E}(f^2(X)) \leq C^2\|f\|_2^2. \tag{116}$$

So we have

$$\mathbb{E}|f^2(X) - \mathbb{E}f^2(X)|^2 \leq \mathbb{E}(f^4(X)) \leq C^2\|f\|_2^2. \tag{117}$$

Applying Bernstein's inequality, we have

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

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

$$\begin{aligned}
\sum_{i=1}^{n} f^2(x_i) \leq & \max\left\{\sqrt{4C^2 \log\frac{1}{\delta_1} n\|f\|_2^2}, \frac{4C^2}{3}\log\frac{1}{\delta_1}\right\} + n\mathbb{E}f^2(X) \\
\leq & \tilde{O}\left(\max\left\{\sqrt{C^2 n\|f\|_2^2}, C^2\right\}\right) + n\|f\|_2^2 \\
\leq & \tilde{O}\left(\sqrt{C^2 n\|f\|_2^2} + C^2\right) + n\|f\|_2^2.
\end{aligned} \tag{118}$$

□

For the proofs in the reminder of this section, the definitions of the relevant quantities are given in Section 2 and the beginning of the Appendix D and D.1.

Corollary 23. *With probability of at least $1 - \delta_1$, we have*

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

Proof of Corollary 23. 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 21 we get the result. □

Corollary 24. *For any $\nu \in \mathbb{R}^R$, with probability of at least $1 - \delta_1$ we have*

$$\|\Phi_R\nu\|_2^2 = \tilde{O}\Big((\tfrac{1}{\delta_1} + 1)n\|\nu\|_2^2\Big).$$

Proof of Corollary 24. 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 21 we get the result. □

Next we consider the quantity, $\Phi_R^T\Phi_R - nI$. The key tool that we use is the matrix Bernstein inequality that describes the upper tail of a sum of independent zero-mean random matrices.

Lemma 25. *Let $D = \mathrm{diag}\{d_1, \ldots, d_R\}$, $d_1, \ldots, d_R > 0$ and $d_{\max} = \max\{d_1, \ldots, d_R\}$. Let $M = \max\{\sum_{p=0}^{R} d_p^2\|\phi_p\|_\infty^2, d_{\max}^2\}$. Then with probability of at least $1 - \delta$, we have*

$$\|D(\Phi_R^T\Phi_R - nI)D\|_2 \leq \max\left\{\sqrt{nd_{\max}^2 M\log\tfrac{R}{\delta}}, M\log\tfrac{R}{\delta})\right\}. \tag{119}$$

Proof of Lemma 25. Let $Y_j = (\phi_1(x_j), \ldots, \phi_R(x_j))^T$ and $Z_j = DY_j$. It is easy to verify that $\mathbb{E}(Z_j Z_j^T) = D^2$. Then the left hand side of (119) is $\sum_{j=1}^{n}[Z_j Z_j^T - \mathbb{E}(Z_j Z_j^T)]$. We note that

$$\|Z_j Z_j^T - \mathbb{E}(Z_j Z_j^T)\|_2 \leq \max\{\|Z_j Z_j^T\|_2, \|\mathbb{E}(Z_j Z_j^T)\|_2\} \leq \max\{\|Z_j\|_2^2, d_{\max}^2\}.$$

For $\|Z_j\|_2^2$, we have

$$\|Z_j\|_2^2 = \sum_{p=0}^{R} d_p^2 \phi_p^2(x_j) \le \sum_{p=0}^{R} d_p^2 \|\phi_p\|_\infty^2, \tag{120}$$

we have

$$\|Z_j Z_j^T - \mathbb{E}(Z_j Z_j^T)\|_2 \le \max\{\textstyle\sum_{p=0}^{R} d_p^2 \|\phi_p\|_\infty^2, d_{\max}^2\}.$$

On the other hand,

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

Since

$$\begin{aligned}\mathbb{E}[\|Z_j\|_2^2 Z_j Z_j^T] &\preccurlyeq \mathbb{E}[\sum_{p=0}^{R} d_p^2 \|\phi_p\|_\infty^2 Z_j Z_j^T], \quad \text{(by (120))} \\ &= \sum_{p=0}^{R} d_p^2 \|\phi_p\|_\infty^2 \mathbb{E}[Z_j Z_j^T],\end{aligned}$$

we have

$$\begin{aligned}\|\mathbb{E}[(Z_j Z_j^T - \mathbb{E}(Z_j Z_j^T))^2]\|_2 &\le \max\{\textstyle\sum_{p=0}^{R} d_p^2 \|\phi_p\|_\infty^2 \|\mathbb{E}[Z_j Z_j^T]\|_2, d_{\max}^4\} \\ &\le \max\{\textstyle\sum_{p=0}^{R} d_p^2 \|\phi_p\|_\infty^2 d_{\max}^2, d_{\max}^4\} \\ &\le d_{\max}^2 \max\{\textstyle\sum_{p=0}^{R} d_p^2 \|\phi_p\|_\infty^2, d_{\max}^2\}.\end{aligned}$$

Using the matrix Bernstein inequality [40, Theorem 6.1], we have

$$\begin{aligned}&\mathbb{P}(\|\sum_{j=1}^{n}[Z_j Z_j^T - \mathbb{E}(Z_j Z_j^T)]\|_2 > t) \\ \le& R \exp\left(\frac{-t^2}{2(n\|\mathbb{E}[(Z_j Z_j^T - \mathbb{E}(Z_j Z_j^T))^2]\|_2 + \frac{t \max_j \|Z_j Z_j^T - \mathbb{E}(Z_j Z_j^T)\|_2}{3})}\right) \\ \le& R \exp\left(\frac{-t^2}{2(n d_{\max}^2 \max\{\sum_{p=0}^{R} d_p^2 \|\phi_p\|_\infty^2, d_{\max}^2\} + \frac{t \max\{\sum_{p=0}^{R} d_p^2 \|\phi_p\|_\infty^2, d_{\max}^2\}}{3})}\right) \\ =& R \exp\left(\frac{-t^2}{O(\max\{n d_{\max}^2 \max\{\sum_{p=0}^{R} d_p^2 \|\phi_p\|_\infty^2, d_{\max}^2\}, t \max\{\sum_{p=0}^{R} d_p^2 \|\phi_p\|_\infty^2, d_{\max}^2\}\})}\right).\end{aligned}$$

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

$$\begin{aligned}&\|\sum_{j=1}^{n}[Z_j Z_j^T - \mathbb{E}(Z_j Z_j^T)]\|_2 \\ &\le \max\left\{\sqrt{n d_{\max}^2 \max\{\textstyle\sum_{p=0}^{R} d_p^2 \|\phi_p\|_\infty^2,\, d_{\max}^2\} \log\frac{R}{\delta}}, \max\left\{\textstyle\sum_{p=0}^{R} d_p^2 \|\phi_p\|_\infty^2,\, d_{\max}^2\right\} \log\frac{R}{\delta}\right\}.\end{aligned}$$

□

Corollary 26. *Suppose that the eigenvalues $(\lambda_p)_{p\ge1}$ satisfy Assumption 1, and the eigenfunctions satisfy Assumption 3. Assume $\sigma^2 = \Theta(n^t)$ where $1 - \frac{\alpha}{1+2\tau} < t < 1$ Let γ be a positive number such that $\frac{1+\alpha+2\tau-(1+2\tau+2\alpha)t}{2\alpha(1-t)} < \gamma \le 1$. Then with probability of at least $1-\delta$, we have*

$$\begin{aligned}&\|\tfrac{1}{\sigma^2}(I + \tfrac{n}{\sigma^2}\Lambda_R)^{-\gamma/2}\Lambda_R^{\gamma/2}(\Phi_R^T\Phi_R - nI)\Lambda_R^{\gamma/2}(I + \tfrac{n}{\sigma^2}\Lambda_R)^{-\gamma/2}\|_2 \\ &\le O\left(n^{\frac{1+\alpha+2\tau-(1+2\tau+2\alpha)t}{2\alpha} - \gamma(1-t)} \sqrt{\log\tfrac{R}{\delta}}\right).\end{aligned} \tag{121}$$

Proof of Corollary 26. Use the same notation as in Lemma 25. Let $D = (I + \frac{n}{\sigma^2}\Lambda_R)^{-\gamma/2}\Lambda_R^{\gamma/2}$. Then $d_{\max}^2 \le \frac{\sigma^{2\gamma}}{n^\gamma}$ and $\sum_{p=0}^R d_p^2\|\phi_p\|_\infty^2 \le \sum_{p=0}^R C_\phi^2 \frac{\lambda_p^\gamma p^{2\tau}}{(1+\frac{n}{\sigma^2}\lambda_p)^\gamma} = O((\frac{n}{\sigma^2})^{\frac{1-\gamma\alpha+2\tau}{\alpha}})$, where the first inequality follows from Assumptions 1 and 3 and the last equality from Lemma 19. Then $M = \max\{\sum_{p=0}^R d_p^2\|\phi_p\|_\infty^2, d_{\max}^2\} = O((\frac{n}{\sigma^2})^{\frac{1-\gamma\alpha+2\tau}{\alpha}})$. Applying Lemma 25, we have

$$\begin{aligned}
&\|\tfrac{1}{\sigma^2}(I + \tfrac{n}{\sigma^2}\Lambda_R)^{-\gamma/2}\Lambda_R^{\gamma/2}(\Phi_R^T\Phi_R - nI)\Lambda_R^{\gamma/2}(I + \tfrac{n}{\sigma^2}\Lambda_R)^{-\gamma/2}\|_2 \\
&\le \tfrac{1}{\sigma^2}\max\left\{\sqrt{n\tfrac{\sigma^{2\gamma}}{n^\gamma}O((\tfrac{n}{\sigma^2})^{\frac{1-\gamma\alpha+2\tau}{\alpha}})\log\tfrac{R}{\delta}}, O((\tfrac{n}{\sigma^2})^{\frac{1-\gamma\alpha+2\tau}{\alpha}})\log\tfrac{R}{\delta}\right\} \\
&= O(\tfrac{1}{\sigma^2}(\tfrac{n}{\sigma^2})^{\frac{1-2\gamma\alpha+2\tau}{2\alpha}}n^{\frac{1}{2}}) = O(\sqrt{\log\tfrac{R}{\delta}}n^{\frac{(1-2\gamma\alpha+2\tau)(1-t)}{2\alpha}+\frac{1}{2}-t}) \\
&= O\bigg(\sqrt{\log\tfrac{R}{\delta}}n^{\frac{1+\alpha+2\tau}{2\alpha}-\frac{(1+2\tau+2\alpha)t}{2\alpha}-\gamma(1-t)}\bigg).
\end{aligned} \tag{122}$$

□

Corollary 27. *Suppose that the eigenvalues $(\lambda_p)_{p\ge1}$ satisfy Assumption 1, and the eigenfunctions satisfy Assumption 3. Let $\tilde\Lambda_{1,R} = \mathrm{diag}\{1, \lambda_1, \ldots, \lambda_R\}$. Assume $\sigma^2 = \Theta(n^t)$ where $t < 1$ Let γ be a positive number such that $\frac{1+2\tau}{\alpha} < \gamma \le 1$. Then with probability of at least $1-\delta$, we have*

$$\|(I + \tfrac{n}{\sigma^2}\Lambda_R)^{-\gamma/2}\tilde\Lambda_{1,R}^{\gamma/2}(\Phi_R^T\Phi_R - nI)\tilde\Lambda_{1,R}^{\gamma/2}(I + \frac{n}{\sigma^2}\Lambda_R)^{-\gamma/2}\|_2 \le O\bigg(\sqrt{\log\tfrac{R}{\delta}}n^{\frac{1}{2}}\bigg). \tag{123}$$

Proof of Corollary 27. Use the same notation as in Lemma 25. Let $D = (I+\frac{n}{\sigma^2}\Lambda_R)^{-\gamma/2}\tilde\Lambda_{1,R}^{\gamma/2}$. Then $d_{\max}^2 \le 1$ and $\sum_{p=0}^R d_p^2\|\phi_p\|_\infty^2 \le C_\phi^2 + \sum_{p=1}^R C_\phi^2\frac{\lambda_p^\gamma p^{2\tau}}{(1+\frac{n}{\sigma^2}\lambda_p)^\gamma} = C_\phi^2 + O(n^{\frac{(1-\gamma\alpha+2\tau)(1-t)}{\alpha}}) = O(1)$ where the first inequality follows from Assumptions 1 and 3 and the second equality from Lemma 19. Then $M = \max\{\sum_{p=0}^R d_p^2\|\phi_p\|_\infty^2, d_{\max}^2\} = O(1)$. Applying Lemma 25, we have

$$\begin{aligned}
&\|(I + \tfrac{n}{\sigma^2}\Lambda_R)^{-\gamma/2}\Lambda_R^{\gamma/2}(\Phi_R^T\Phi_R - nI)\Lambda_R^{\gamma/2}(I + \tfrac{n}{\sigma^2}\Lambda_R)^{-\gamma/2}\|_2 \\
&\le \max\left\{\sqrt{\log\tfrac{R}{\delta}nO(1)}, \log\tfrac{R}{\delta}O(1)\right\} \\
&= O\bigg(\sqrt{\log\tfrac{R}{\delta}}n^{\frac{1}{2}}\bigg).
\end{aligned} \tag{124}$$

□

Corollary 28. *Suppose that the eigenvalues $(\lambda_p)_{p\ge1}$ satisfy Assumption 1, and the eigenfunctions satisfy Assumption 3. Let $\Phi_{R+1:S} = (\phi_{R+1}(\mathbf{x}), \ldots, \phi_S(\mathbf{x}))$, and $\Lambda_{R+1:S} = (\lambda_{R+1}, \ldots, \lambda_S)$. Then with probability of at least $1-\delta$, we have*

$$\|\Lambda_{R+1:S}^{1/2}(\Phi_{R+1:S}^T\Phi_{R+1:S} - nI)\Lambda_{R+1:S}^{1/2}\|_2 \le O\big(\log\tfrac{S-R}{\delta}\max\{n^{\frac{1}{2}}R^{\frac{1-2\alpha+2\tau}{2}}, R^{1-\alpha+2\tau}\}\big). \tag{125}$$

Proof of Corollary 28. Use the same notation as in Lemma 25. Let $D = \Lambda_{R+1:S}^{1/2}$. Then $d_{\max}^2 \le \overline{C_\lambda}R^{-\alpha} = O(R^{-\alpha})$ and $\sum_{p=R+1}^S C_\phi^2 d_p^2 p^{2\tau} \le \sum_{p=R+1}^S C_\phi^2\overline{C_\lambda}p^{-\alpha}p^{2\tau} = O(R^{1-\alpha+2\tau})$, where the first inequality follows from Assumptions 1 and 3. Then $M = \max\{\sum_{p=R+1}^S C_\phi^2 d_p^2 p^{2\tau}, d_{\max}^2\} = O(R^{1-\alpha+2\tau})$. Applying Lemma 25, we have

$$\begin{aligned}
&\|(I + \frac{n}{\sigma^2}\Lambda_R)^{-\gamma/2}\Lambda_R^{\gamma/2}(\Phi_R^T\Phi_R - nI)\Lambda_R^{\gamma/2}(I + \frac{n}{\sigma^2}\Lambda_R)^{-\gamma/2}\|_2 \\
&\le \max\left\{\sqrt{\log\tfrac{S-R}{\delta}nO(R^{-\alpha})O(R^{1-\alpha+2\tau})}, \log\tfrac{S-R}{\delta}O(R^{1-\alpha+2\tau}))\right\} \\
&= O\big(\log\tfrac{S-R}{\delta}\max\{n^{\frac{1}{2}}R^{\frac{1-2\alpha+2\tau}{2}}, R^{1-\alpha+2\tau}\}\big).
\end{aligned} \tag{126}$$

□

Lemma 29. *Under the assumptions of Corollary 28, with probability of at least $1-\delta$, we have*

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

Proof of Lemma 29. For $S \in \mathbb{N}$, we have

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

Let $S = R^{\frac{\alpha}{\alpha-1-2\tau}}$. Then we get $\|\Phi_{>S}\Lambda_{>S}\Phi_{>S}^T\|_2 = O(nR^{-\alpha})$.

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

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

where in the fourth inequality we use Corollary 28. □

Corollary 30. *Assume that $\sigma^2 = \Theta(1)$. If $R = n^{\frac{1}{\alpha}+\kappa}$ where $0 < \kappa < \frac{\alpha-1-2\tau}{\alpha(1+2\tau)}$, then with probability of at least $1-\delta$, we have*

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

Proof of Corollary 30. By Lemma 29 and the assumption $R = n^{\frac{1}{\alpha}+\kappa}$, we have

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

□

Lemma 31. *Assume that $\|\frac{1}{\sigma^2}(I + \frac{n}{\sigma^2}\Lambda_R)^{-\gamma/2}\Lambda_R^{\gamma/2}(\Phi_R^T\Phi_R - nI)\Lambda_R^{\gamma/2}(I + \frac{n}{\sigma^2}\Lambda_R)^{-\gamma/2}\|_2 < 1$ where $\frac{1+2\tau}{\alpha} < \gamma \leq 1$. We then have*

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

Proof of Lemma 31. First note that

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

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

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

Letting $\epsilon \to 0$, we get

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

This concludes the proof. □

Lemma 32. *If* $\|(I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}\frac{\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T}{\sigma^2}\|_2 < 1$*, then we have*

$$
(I + \tfrac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1} - (I + \tfrac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1} = \sum_{j=1}^{\infty}(-1)^j\left((I + \tfrac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}\tfrac{\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T}{\sigma^2}\right)^j(I + \tfrac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}. \tag{127}
$$

In particular, assume that $\sigma^2 = \Theta(1)$*. Let* $R = n^{\frac{1}{\alpha}+\kappa}$ *where* $0 < \kappa < \frac{\alpha-1-2\tau}{\alpha(1+2\tau)}$*. Then with probability of at least* $1 - \delta$*, for sufficiently large* n*, we have* $\|(I + \frac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}\frac{\Phi_{>R}\Lambda_{>R}\Phi_{>R}^T}{\sigma^2}\|_2 < 1$ *and* (127) *holds.*

Proof of Lemma 32. Define $\Phi_{>R} = (\phi_{R+1}(\mathbf{x}), \phi_{R+2}(\mathbf{x}), \ldots)$, $\Lambda_{>R} = \mathrm{diag}(\lambda_{R+1}, \lambda_{R+2}, \ldots)$. Then we have

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

By Corollary 30, 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 of at least $1 - \delta$. Hence

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

□

Lemma 33. *Assume that $\mu_0 = 0$ and $\sigma^2 = \Theta(n^t)$ where $1 - \frac{\alpha}{1+2\tau} < t < 1$. Let $R = n^{(\frac{1}{\alpha}+\kappa)(1-t)}$ where $0 < \kappa < \frac{\alpha-1-2\tau+(1+2\tau)t}{\alpha^2(1-t)}$. Then when n is sufficiently large, with probability of at least $1-2\delta$ we have*

$$\|(I + \tfrac{1}{\sigma^2}\Phi_R\Lambda_R\Phi_R^T)^{-1} f_R(\mathbf{x})\|_2 = \tilde{O}\bigg(\sqrt{(\tfrac{1}{\delta}+1)n}\cdot n^{\max\{-(1-t),\frac{(1-2\beta)(1-t)}{2\alpha}\}}\bigg). \tag{128}$$

Proof of Lemma 33. Let $\Lambda_{1:R} = \mathrm{diag}\{\lambda_1,\ldots,\lambda_R\}$, $\Phi_{1:R} = (\phi_1(\mathbf{x}),\phi_1(\mathbf{x}),\ldots,\phi_R(\mathbf{x}))$ and $\boldsymbol{\mu}_{1:R} = (\mu_1,\ldots,\mu_R)$. Since $\mu_0 = 0$, we have $(I + \frac{1}{\sigma^2}\Phi_R\Lambda_R\Phi_R^T)^{-1} f_R(\mathbf{x}) = (I + \frac{1}{\sigma^2}\Phi_{1:R}\Lambda_{1:R}\Phi_{1:R}^T)^{-1}\Phi_{1:R}\boldsymbol{\mu}_{1:R}$. Using the Woodbury matrix identity, we have that

$$\begin{aligned}(I + \tfrac{1}{\sigma^2}\Phi_{1:R}\Lambda_{1:R}\Phi_{1:R}^T)^{-1}\Phi_{1:R}\boldsymbol{\mu}_{1:R} =& \big[I - \Phi_{1:R}(\sigma^2 I + \Lambda_{1:R}\Phi_{1:R}^T\Phi_{1:R})^{-1}\Lambda_{1:R}\Phi_{1:R}^T\big]\Phi_{1:R}\boldsymbol{\mu}_{1:R}\\ =&\Phi_{1:R}\boldsymbol{\mu}_{1:R} - \Phi_{1:R}(\sigma^2 I + \Lambda_{1:R}\Phi_{1:R}^T\Phi_{1:R})^{-1}\Lambda_{1:R}\Phi_{1:R}^T\Phi_{1:R}\boldsymbol{\mu}_{1:R}\\ =&\Phi_{1:R}(I + \tfrac{1}{\sigma^2}\Lambda_{1:R}\Phi_{1:R}^T\Phi_{1:R})^{-1}\boldsymbol{\mu}_{1:R}.\end{aligned} \tag{129}$$

Let $A = (I + \frac{n}{\sigma^2}\Lambda_{1:R})^{-1/2}\Lambda_{1:R}^{1/2}(\Phi_{1:R}^T\Phi_{1:R} - nI)\Lambda_{1:R}^{1/2}(I + \frac{n}{\sigma^2}\Lambda_{1:R})^{-1/2}$.By Corollary 26, with probability of at least $1-\delta$, we have $\|\frac{1}{\sigma^2}A\|_2 = \sqrt{\log\frac{R}{\delta}}n^{\frac{1-\alpha+2\tau}{2\alpha}-\frac{(1+2\tau)t}{2\alpha}}$. When n is sufficiently large, $\|\frac{1}{\sigma^2}A\|_2 = o(1)$ is less than 1 because $1 - \frac{\alpha}{1+2\tau} < t < 1$. By Lemma 31, we have

$$\begin{aligned}&(I + \tfrac{1}{\sigma^2}\Lambda_{1:R}\Phi_{1:R}^T\Phi_{1:R})^{-1}\\ =&(I + \tfrac{n}{\sigma^2}\Lambda_{1:R})^{-1} + \sum_{j=1}^{\infty}(-1)^j\big(\tfrac{1}{\sigma^2}(I + \tfrac{n}{\sigma^2}\Lambda_{1:R})^{-1}\Lambda_{1:R}(\Phi_{1:R}^T\Phi_{1:R} - nI)\big)^j(I + \tfrac{n}{\sigma^2}\Lambda_{1:R})^{-1}.\end{aligned}$$

We then have

$$\begin{aligned}&\|(I + \frac{1}{\sigma^2}\Lambda_{1:R}\Phi_{1:R}^T\Phi_{1:R})^{-1}\boldsymbol{\mu}_{1:R}\|_2\\ =&\bigg\|\bigg((I + \tfrac{n}{\sigma^2}\Lambda_{1:R})^{-1} + \sum_{j=1}^{\infty}(-1)^j\big(\tfrac{1}{\sigma^2}(I + \tfrac{n}{\sigma^2}\Lambda_{1:R})^{-1}\Lambda_{1:R}(\Phi_{1:R}^T\Phi_{1:R} - nI)\big)^j(I + \tfrac{n}{\sigma^2}\Lambda_{1:R})^{-1}\bigg)\boldsymbol{\mu}_{1:R}\bigg\|_2\\ \le&\bigg(\|(I + \tfrac{n}{\sigma^2}\Lambda_{1:R})^{-1}\boldsymbol{\mu}_{1:R}\|_2 + \sum_{j=1}^{\infty}\Big\|\big(\tfrac{1}{\sigma^2}(I + \tfrac{n}{\sigma^2}\Lambda_{1:R})^{-1}\Lambda_{1:R}(\Phi_{1:R}^T\Phi_{1:R} - nI)\big)^j(I + \tfrac{n}{\sigma^2}\Lambda_{1:R})^{-1}\boldsymbol{\mu}_{1:R}\Big\|_2\bigg).\end{aligned} \tag{130}$$

By Lemma 19 and Assumption 2, assuming that $\sup_{i\ge1} p_{i+1} - p_i = h$, we have

$$\begin{aligned}\|(I + \tfrac{n}{\sigma^2}\Lambda_{1:R})^{-1}\boldsymbol{\mu}_{1:R}\|_2 \le& \sqrt{\sum_{p=1}^{R}\frac{C_\mu^2 p^{-2\beta}}{(1 + n\underline{C_\lambda}p^{-\alpha}/\sigma^2)^2}} = \Theta(n^{\max\{-(1-t),\frac{(1-2\beta)(1-t)}{2\alpha}\}}\log^{k/2}n),\\ \|(I + \tfrac{n}{\sigma^2}\Lambda_{1:R})^{-1}\boldsymbol{\mu}_{1:R}\|_2 \ge& \sqrt{\sum_{i=1}^{\lfloor\frac{R}{h}\rfloor}\frac{\underline{C_\mu^2} i^{-2\beta}}{(1 + \frac{n}{\sigma^2}\overline{C_\lambda}(hi)^{-\alpha})^2}} = \Theta(n^{\max\{-(1-t),\frac{(1-2\beta)(1-t)}{2\alpha}\}}\log^{k/2}n)\end{aligned}$$

where $k = \begin{cases}0, & 2\alpha \ne 2\beta - 1,\\ 1, & 2\alpha = 2\beta - 1.\end{cases}$. Overall we have

$$\|(I + \tfrac{n}{\sigma^2}\Lambda_{1:R})^{-1}\boldsymbol{\mu}_{1:R}\|_2 = \Theta(n^{(1-t)\max\{-1,\frac{1-2\beta}{2\alpha}\}}\log^{k/2}n). \tag{131}$$

Using the fact that $\|\frac{1}{\sigma^2}A\|_2 = \sqrt{\log\frac{R}{\delta}}n^{\frac{1-\alpha+2\tau}{2\alpha}-\frac{(1+2\tau)t}{2\alpha}}$ and $\|(I + \frac{n}{\sigma^2}\Lambda_{1:R})^{-1}\Lambda_{1:R}\|_2 \le n^{-1}$, we have

$$\begin{aligned}&\Big\|\big(\tfrac{1}{\sigma^2}(I + \tfrac{n}{\sigma^2}\Lambda_{1:R})^{-1}\Lambda_{1:R}(\Phi_{1:R}^T\Phi_{1:R} - nI)\big)^j(I + \tfrac{n}{\sigma^2}\Lambda_{1:R})^{-1}\boldsymbol{\mu}_{1:R}\Big\|_2\\ =&\Big\|(I + \tfrac{n}{\sigma^2}\Lambda_{1:R})^{-\frac{1}{2}}\Lambda_{1:R}^{\frac{1}{2}}(\tfrac{1}{\sigma^2}A)^j(I + \tfrac{n}{\sigma^2}\Lambda_{1:R})^{-\frac{1}{2}}\Lambda_{1:R}^{\frac{1}{2}}\boldsymbol{\mu}_{1:R}\Big\|_2\\ \le&\tilde{O}(n^{-\frac{1-t}{2}})\|\tfrac{1}{\sigma^2}A\|_2^j\|(I + \tfrac{n}{\sigma^2}\Lambda_{1:R})^{-\frac{1}{2}}\Lambda_{1:R}^{-\frac{1}{2}}\boldsymbol{\mu}_{1:R}\|_2\end{aligned} \tag{132}$$

By Lemma 20 and the assumption $R = n^{(\frac{1}{\alpha}+\kappa)(1-t)}$,

$$\begin{aligned}\|(I+\tfrac{n}{\sigma^2}\Lambda_{1:R})^{-\frac{1}{2}}\Lambda_{1:R}^{-\frac{1}{2}}\boldsymbol{\mu}_{1:R}\|_2 \leq& \sqrt{\sum_{p=1}^{R}\frac{(\underline{C_\lambda}p^{-\alpha})^{-1}C_\mu^2 p^{-2\beta}}{(1+n\underline{C_\lambda}p^{-\alpha}/\sigma^2)^1}}\\ =&\tilde{O}(\max\{n^{-(1-t)/2}, R^{1/2-\beta+\alpha/2}\})\\ =&\tilde{O}(\max\{n^{-(1-t)/2}, n^{(\frac{1}{2}+\frac{1-2\beta}{2\alpha}+\kappa(1/2-\beta+\alpha/2))(1-t)}\})\end{aligned} \tag{133}$$

We then have

$$\begin{aligned}&\left\|\left(\tfrac{1}{\sigma^2}(I+\tfrac{n}{\sigma^2}\Lambda_{1:R})^{-1}\Lambda_{1:R}(\Phi_{1:R}^T\Phi_{1:R}-nI)\right)^j(I+\tfrac{n}{\sigma^2}\Lambda_{1:R})^{-1}\boldsymbol{\mu}_{1:R}\right\|_2\\ =&\|\tfrac{1}{\sigma^2}A\|_2^j\tilde{O}(\max\{n^{-(1-t)}, n^{(\frac{1-2\beta}{2\alpha}+\kappa(1/2-\beta+\alpha/2))(1-t)}\})\end{aligned} \tag{134}$$

By (130), (131) and (134), we have

$$\begin{aligned}&\|(I+\tfrac{1}{\sigma^2}\Lambda_{1:R}\Phi_{1:R}^T\Phi_{1:R})^{-1}\boldsymbol{\mu}_{1:R}\|_2\\ =&\Theta(n^{(1-t)\max\{-1,\frac{1-2\beta}{2\alpha}\}}\log^{k/2}n)+\sum_{j=1}^{\infty}\|\frac{1}{\sigma^2}A\|_2^j\tilde{O}(\max\{n^{-(1-t)}, n^{(1-t)\frac{1-2\beta}{2\alpha}+\kappa(1-t)(1/2-\beta+\alpha/2)}\})\\ =&\Theta(n^{(1-t)\max\{-1,\frac{1-2\beta}{2\alpha}\}}\log^{k/2}n)+\tilde{O}(n^{\frac{1-\alpha+2\tau}{2\alpha}-\frac{(1+2\tau)t}{2\alpha}})\tilde{O}(\max\{n^{-(1-t)}, n^{(1-t)\frac{1-2\beta}{2\alpha}+\kappa(1-t)(1/2-\beta+\alpha/2)}\}).\end{aligned} \tag{135}$$

By assumption $\kappa < \frac{\alpha-1-2\tau+(1+2\tau)t}{\alpha^2(1-t)}$, we have that

$$\kappa(1-t)(1/2-\beta+\alpha/2)+\frac{1-\alpha+2\tau}{2\alpha}-\frac{(1+2\tau)t}{2\alpha}<\kappa\alpha(1-t)/2+\frac{1-\alpha+2\tau}{2\alpha}-\frac{(1+2\tau)t}{2\alpha}<0.$$

Using (135), we then get

$$\begin{aligned}\|(I+\tfrac{1}{\sigma^2}\Lambda_{1:R}\Phi_{1:R}^T\Phi_{1:R})^{-1}\boldsymbol{\mu}_{1:R}\|_2 &= \Theta(n^{(1-t)\max\{-1,\frac{1-2\beta}{2\alpha}\}}\log^{k/2}n)\\ &= \frac{1+o(1)}{\sigma^2}\|(I+\frac{n}{\sigma^2}\Lambda_{1:R})^{-1}\boldsymbol{\mu}_{1:R}\|_2.\end{aligned} \tag{136}$$

By Corollary 24, with probability of at least $1-\delta$, we have

$$\begin{aligned}\|\Phi_{1:R}(I+\tfrac{1}{\sigma^2}\Lambda_{1:R}\Phi_{1:R}^T\Phi_{1:R})^{-1}\boldsymbol{\mu}_{1:R}\|_2 =&\tilde{O}(\sqrt{(\tfrac{1}{\delta}+1)n}\|(I+\tfrac{1}{\sigma^2}\Lambda_{1:R}\Phi_{1:R}^T\Phi_{1:R})^{-1}\boldsymbol{\mu}_{1:R}\|_2)\\ =&\tilde{O}(\sqrt{(\tfrac{1}{\delta}+1)n}\cdot n^{(1-t)\max\{-1,\frac{1-2\beta}{2\alpha}\}}).\end{aligned} \tag{137}$$

From (129), we get $\|(I+\frac{1}{\sigma^2}\Phi_{1:R}\Lambda_{1:R}\Phi_{1:R}^T)^{-1}\Phi_{1:R}\boldsymbol{\mu}_{1:R}\|_2 = \tilde{O}(\sqrt{(\frac{1}{\delta}+1)n}\cdot n^{(1-t)\max\{-1,\frac{1-2\beta}{2\alpha}\}})$. This concludes the proof. □

Lemma 34. *Assume that $\mu_0 > 0$ and $\sigma^2 = \Theta(n^t)$ where $1-\frac{\alpha}{1+2\tau} < t < 1$. Let $R = n^{\frac{1}{\alpha}+\kappa}$ where $0 < \kappa < \frac{\alpha-1-2\tau+(1+2\tau)t}{\alpha^2}$. Then when n is sufficiently large, with probability of at least $1-2\delta$, we have*

$$\|(I+\tfrac{1}{\sigma^2}\Phi_R\Lambda_R\Phi_R^T)^{-1}f_R(\mathbf{x})\|_2 = \tilde{O}\left(\sqrt{(\tfrac{1}{\delta}+1)n}\right). \tag{138}$$

Proof of Lemma 34. Using the Woodbury matrix identity, we have that

$$\begin{aligned}(I+\tfrac{1}{\sigma^2}\Phi_R\Lambda_R\Phi_R^T)^{-1}f_R(\mathbf{x}) =&\left[I-\Phi_R(\sigma^2 I+\Lambda_R\Phi_R^T\Phi_R)^{-1}\Lambda_R\Phi_R^T\right]\Phi_R\boldsymbol{\mu}_R\\ =&\Phi_R\boldsymbol{\mu}_R-\Phi_R(\sigma^2 I+\Lambda_R\Phi_R^T\Phi_R)^{-1}\Lambda_R\Phi_R^T\Phi_R\boldsymbol{\mu}_R\\ =&\Phi_R(I+\tfrac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}\boldsymbol{\mu}_R.\end{aligned} \tag{139}$$

Let $\boldsymbol{\mu}_{R,1} = (\mu_0, 0, \ldots, 0)$ and $\boldsymbol{\mu}_{R,2} = (0, \mu_1, \ldots, \mu_R)$. Then $\boldsymbol{\mu}_R = \boldsymbol{\mu}_{R,1} + \boldsymbol{\mu}_{R,2}$. Then we have

$$\|(I+\tfrac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}\boldsymbol{\mu}_R\|_2 = \|(I+\tfrac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}\boldsymbol{\mu}_{R,1}\|_2 + \|(I+\tfrac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}\boldsymbol{\mu}_{R,2}\|_2. \tag{140}$$

According to (136) in the proof of Lemma 33, we have $\|(I+\frac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}\boldsymbol{\mu}_{R,2}\|_2 = \tilde{O}(n^{\max\{-(1-t),\frac{(1-t)(1-2\beta)}{2\alpha}\}})$. Next we estimate $\|(I+\frac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}\boldsymbol{\mu}_{R,1}\|_2$.

Let

$$A=(I+\frac{n}{\sigma^2}\Lambda_{1:R})^{-\gamma/2}\Lambda_{1:R}^{\gamma/2}(\Phi_{1:R}^T\Phi_{1:R}-nI)\Lambda_{1:R}^{\gamma/2}(I+\frac{n}{\sigma^2}\Lambda_{1:R})^{-\gamma/2}$$

where $\frac{1}{1-t}(\frac{1+\alpha+2\tau}{2\alpha}-\frac{(1+2\tau+2\alpha)t}{2\alpha})<\gamma<1$. Since $1-\frac{\alpha}{1+2\tau}<t<1$, $\frac{1}{1-t}(\frac{1+\alpha+2\tau}{2\alpha}-\frac{(1+2\tau+2\alpha)t}{2\alpha})<1$ so the range for γ is well-defined.By Corollary 26, with probability of at least $1-\delta$, we have $\|\frac{1}{\sigma^2}A\|_2=\tilde{O}(\sqrt{\log\frac{R}{\delta}}n^{\frac{1+\alpha+2\tau}{2\alpha}-\frac{(1+2\tau+2\alpha)t}{2\alpha}-\gamma(1-t)})=o(1)$. When n is sufficiently large, $\|\frac{1}{\sigma^2}A\|_2$ is less than 1 because $1-\frac{\alpha}{1+2\tau}<t<1$. By Lemma 31, we have

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

We then have

$$\begin{aligned}&\|(I+\tfrac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}\boldsymbol{\mu}_{R,1}\|_2\\ =&\left\|\left((I+\tfrac{n}{\sigma^2}\Lambda_R)^{-1}+\sum_{j=1}^{\infty}(-1)^j\big(\tfrac{1}{\sigma^2}(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R(\Phi_R^T\Phi_R-nI)\big)^j(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-1}\right)\boldsymbol{\mu}_{R,1}\right\|_2\\ \le&\left(\|(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-1}\boldsymbol{\mu}_{R,1}\|_2+\sum_{j=1}^{\infty}\left\|\big(\tfrac{1}{\sigma^2}(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R(\Phi_R^T\Phi_R-nI)\big)^j(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-1}\boldsymbol{\mu}_{R,1}\right\|_2\right).\end{aligned}\tag{141}$$

By Lemma 19,

$$\|(I+\frac{n}{\sigma^2}\Lambda_R)^{-1}\boldsymbol{\mu}_{R,1}\|_2\le\sqrt{\mu_0^2+\sum_{p=1}^{R}\frac{C_\mu^2p^{-2\beta}}{(1+n\underline{C_\lambda}p^{-\alpha}/\sigma^2)^2}}=O(1).\tag{142}$$

Let $\tilde{\Lambda}_{1,R}=\mathrm{diag}\{1,\lambda_1,\ldots,\lambda_R\}$ and $I_{0,R}=(0,1,\ldots,1)$. Then $\Lambda_R=\tilde{\Lambda}_{1,R}I_{0,R}$. Let $B=(I+\frac{n}{\sigma^2}\Lambda_R)^{-\gamma/2}\tilde{\Lambda}_{1,R}^{\gamma/2}(\Phi_R^T\Phi_R-nI)\tilde{\Lambda}_{1,R}^{\gamma/2}(I+\frac{n}{\sigma^2}\Lambda_R)^{-\gamma/2}$. According to Corollary 27, we have $\|B\|_2=O\big(\sqrt{\log\frac{R}{\delta}}n^{\frac{1}{2}}\big)$. Using the fact that $\|\frac{1}{\sigma^2}A\|_2=\tilde{O}\big(\sqrt{\log\frac{R}{\delta}}n^{\frac{1+\alpha+2\tau}{2\alpha}-\frac{(1+2\tau+2\alpha)t}{2\alpha}-\gamma(1-t)}\big)$, we have

$$\begin{aligned}&\left\|\big(\tfrac{1}{\sigma^2}(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-1}\Lambda_R(\Phi_R^T\Phi_R-nI)\big)^j(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-1}\boldsymbol{\mu}_{R,1}\right\|_2\\ =&\frac{1}{\sigma^{2j}}\left\|(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-1+\frac{\gamma}{2}}\Lambda_R^{1-\frac{\gamma}{2}}\Big(A(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-1+\gamma}\Lambda_R^{1-\gamma}\Big)^{j-1}B(I+\tfrac{n}{\sigma^2}\Lambda_R)^{-1+\frac{\gamma}{2}}\boldsymbol{\mu}_{R,1}\right\|_2\\ \le&\frac{1}{\sigma^2}(n^{(-1+\frac{\gamma}{2}+(-1+\gamma)(j-1))(1-t)}\tilde{O}(\sqrt{\log\tfrac{R}{\delta}}n^{(j-1)(\frac{1+\alpha+2\tau}{2\alpha}-\frac{(1+2\tau+2\alpha)t}{2\alpha}-\gamma(1-t))})\sqrt{\log\tfrac{R}{\delta}}n^{\frac{1}{2}}\|\boldsymbol{\mu}_{R,1}\|_2\\ \le&n^{(-1+\frac{\gamma}{2})(1-t)+\frac{1}{2}-t}\tilde{O}(n^{\frac{[1-\alpha+2\tau-(1+2\tau)t](j-1)}{2\alpha}})\sqrt{\log\tfrac{R}{\delta}}\|\boldsymbol{\mu}_{R,1}\|_2\\ =&\tilde{O}(n^{-\frac{1}{2}+\frac{\gamma}{2}(1-t)+\frac{[1-\alpha+2\tau-(1+2\tau)t](j-1)}{2\alpha}}).\end{aligned}\tag{143}$$

Since $\frac{1}{1-t}(\frac{1+\alpha+2\tau}{2\alpha}-\frac{(1+2\tau+2\alpha)t}{2\alpha})<\gamma<1$ and $-\frac{1}{2}+\frac{1}{1-t}(\frac{1+\alpha+2\tau}{2\alpha}-\frac{(1+2\tau+2\alpha)t}{2\alpha})\frac{1-t}{2}<0$, we can let γ be a little bit larger than $\frac{1}{1-t}(\frac{1+\alpha+2\tau}{2\alpha}-\frac{(1+2\tau+2\alpha)t}{2\alpha})$ and make $-\frac{1}{2}+\frac{\gamma}{2}(1-t)<0$ holds. By (141), (142), (143), we have

$$\begin{aligned}&\|(I+\tfrac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}\boldsymbol{\mu}_{R,1}\|_2\\ \le&O(1)+\sum_{j=1}^{\infty}\tilde{O}(n^{-\frac{1}{2}+\frac{\gamma}{2}(1-t)+\frac{[1-\alpha+2\tau-(1+2\tau)t](j-1)}{2\alpha}})\\ \le&O(1)+o(1)=O(1).\end{aligned}\tag{144}$$

According to (140), we have $\|(I+\frac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}\boldsymbol{\mu}_R\|_2=\tilde{O}(n^{\max\{-(1-t),\frac{(1-t)(1-2\beta)}{2\alpha}\}})+O(1)=O(1)$. By Corollary 24, with probability of at least $1-\delta$, we have

$$\begin{aligned}\|\Phi_R(I+\tfrac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}\boldsymbol{\mu}_R\|_2=&\tilde{O}(\sqrt{(\tfrac{1}{\delta}+1)n}\|(I+\tfrac{1}{\sigma^2}\Lambda_R\Phi_R^T\Phi_R)^{-1}\boldsymbol{\mu}_R\|_2)\\=&\tilde{O}\bigg(\sqrt{(\tfrac{1}{\delta}+1)n}\bigg).\end{aligned}$$

From (139), we get $\|(I+\frac{1}{\sigma^2}\Phi_R\Lambda_R\Phi_R^T)^{-1}f_R(\mathbf{x})\|_2=\tilde{O}\Big(\sqrt{(\frac{1}{\delta}+1)n}\Big)$. This concludes the proof. □

Lemma 35. *Assume that $\sigma^2=\Theta(1)$. Let $R=n^{\frac{1}{\alpha}+\kappa}$ where $0<\kappa<\frac{\alpha-1-2\tau}{\alpha^2}$. Assume that $\mu_0=0$. Then when n is sufficiently large, with probability of at least $1-3\delta$ we have*

$$\|(I+\tfrac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f_R(\mathbf{x})\|_2=\tilde{O}(\sqrt{(\tfrac{1}{\delta}+1)n}\cdot n^{\max\{-1,\frac{1-2\beta}{2\alpha}\}}).\tag{145}$$

Assume that $\mu_0>0$. Then when n is sufficiently large, with probability of at least $1-3\delta$ we have

$$\|(I+\tfrac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f_R(\mathbf{x})\|_2=\tilde{O}(\sqrt{(\tfrac{1}{\delta}+1)n}).\tag{146}$$

Proof of Lemma 35. We have

$$\begin{aligned}&(I+\tfrac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f_R(\mathbf{x})\\=&(I+\tfrac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}f_R(\mathbf{x})+\Big((I+\tfrac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}-(I+\tfrac{\Phi_R\Lambda_R\Phi_R^T}{\sigma^2})^{-1}\Big)f_R(\mathbf{x}).\end{aligned}\tag{147}$$

When $\mu_0=0$, by Lemma 33, with probability of at least $1-2\delta$, we have

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

Since $\frac{\alpha-1-2\tau}{\alpha^2}<\frac{\alpha-1-2\tau}{\alpha(1+2\tau)}$, we apply Lemma 32 and Corollary 30 and get that with probability of at least $1-\delta$, the second term in the right hand side of (147) is estimated as follows:

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

Overall, from (147), we have that with probability $1-3\delta$,

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

When $\mu_0>0$, using the same approach and Lemma 34, we can prove that $\|(I+\frac{\Phi\Lambda\Phi^T}{\sigma^2})^{-1}f_R(\mathbf{x})\|_2=\tilde{O}(\sqrt{(\frac{1}{\delta}+1)n})$. This concludes the proof. □

F Details on the experiments

In our experiment, the input space and input distribution are $\Omega=S^1$ and $\rho=\mathcal{U}(S^1)$, and we use the first order arc-cosine kernel function. [11] 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 [22]. Under this setting, the eigenvalues and eigenfunctions 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 1 is satisfied with $\alpha = 4$, and the second part of Assumption 3 is satisfied with $\|\phi_p\| \leq \frac{\sqrt{2}}{2}$, $p \geq 1$.

The training and test data are generated as follows: We independently sample training inputs $x_1, \ldots, x_n$ and test input x_{n+1} from $\mathcal{U}(S^1)$ and training outputs y_i, $i = 1, \ldots, n$ from $\mathcal{N}(f(x_i), \sigma^2)$, where we choose $\sigma = 0.1$. The Bayesian predictive distribution conditioned on the test point x_{n+1} $\mathcal{N}(\bar{m}(x_{n+1}), \bar{k}(x_{n+1}, x_{n+1}))$ is obtained by (10) and (11). We compute the normalized SC by (12) and the Bayesian generalization error by the Kullback-Leibler divergence between $\mathcal{N}(f(x_{n+1}), \sigma^2)$ and $\mathcal{N}(\bar{m}(x_{n+1}), \bar{k}(x_{n+1}, x_{n+1}))$.

Consider the first order arc-cosine kernel function with biases,

$$k^{(1)}_{\text{w/ bias}}(x_1, x_2) = \frac{1}{\pi}(\sin\bar{\psi} + (\pi - \bar{\psi})\cos\bar{\psi}), \text{ where } \bar{\psi} = \arccos\big(\tfrac{1}{2}(\langle x_1, x_2\rangle + 1)\big). \tag{148}$$

[33] showed that this kernel is the conjugate kernel of an infinitely wide shallow ReLU network with two inputs and one hidden layer with biases, whose eigenvalues satisfy Assumption 1 with $\alpha = 4$. The eigenfunctions of this kernel are the same as that of the first-order arc-cosine kernel without biases, $k^{(1)}_{\text{w/o bias}}$ in Section 3. We consider the target functions in Table 3, which satisfy Assumption 5 with the indicated β, and μ_0 indicates whether the function lies in the span of eigenfunctions of the kernel. For each target we conduct GPR 20 times and report the mean and standard deviation of the normalized SC and the Bayesian generalization error in Figure 3, which agree with the asymptotics predicted in Theorems 4 and 5.

Table 2 summarizes all the different kernel functions that we consider in our experiments with pointers to the corresponding tables and figures.

	kernel function	α	activation function	bias	pointer
$k^{(1)}_{\text{w/o bias}}$	$\frac{1}{\pi}(\sin\psi + (\pi - \psi)\cos\psi)$	4	$\max\{0, x\}$	no	Table 1/Figure 1
$k^{(1)}_{\text{w/ bias}}$	$\frac{1}{\pi}(\sin\bar{\psi} + (\pi - \bar{\psi})\cos\bar{\psi})$	4	$\max\{0, x\}$	yes	Table 3/Figure 3
$k^{(2)}_{\text{w/o bias}}$	$\frac{1}{\pi}(3\sin\psi\cos\psi + (\pi - \psi)(1 + 2\cos^2\psi))$	6	$(\max\{0, x\})^2$	no	Table 4/Figure 4
$k^{(2)}_{\text{w/ bias}}$	$\frac{1}{\pi}(3\sin\bar{\psi}\cos\bar{\psi} + (\pi - \bar{\psi})(1 + 2\cos^2\bar{\psi}))$	6	$(\max\{0, x\})^2$	yes	Table 5/Figure 5
$k^{(0)}_{\text{w/o bias}}$	$\frac{1}{\pi}(\sin\psi + (\pi - \psi)\cos\psi)$	2	$\frac{1}{2}(1 + \text{sign}(x))$	no	Table 6/Figure 6
$k^{(0)}_{\text{w/ bias}}$	$\frac{1}{\pi}(\sin\bar{\psi} + (\pi - \psi)\cos\bar{\psi})$	2	$\frac{1}{2}(1 + \text{sign}(x))$	yes	Table 7/Figure 7

Table 2: The different kernel functions used in our experiments, their values of α, the corresponding neural network activation function along with a pointer to the tables showing the target functions used for the kernels and the corresponding figures.

Summarizing the observations from these experiments, we see that the smoothness of the activation function (which is controlled by the order of the arc-cosine kernel) influences the decay rate α of the eigenvalues. In general, when the activation function is smoother, the decay rate α is larger. Theorem 5 then implies that smooth activation functions are more capable in suppressing noise but slower in learning the target. We also observe that networks with biases are more capable at learning functions compared to networks without bias. For example, the function $\cos(2\theta)$ cannot be learned by the zero order arc-cosine kernel without biases (see Table 6 and Figure 6), but it can be learned by the zero order arc-cosine kernel with biases (see Table 7 and Figure 7).

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

Table 3: Target functions used in the experiments for the first order arc-cosine kernel with bias, $k^{(1)}_{\text{w/ bias}}$, their values of β and μ_0, and theoretical rates for the normalized SC and the Bayesian generalization error from our theorems.

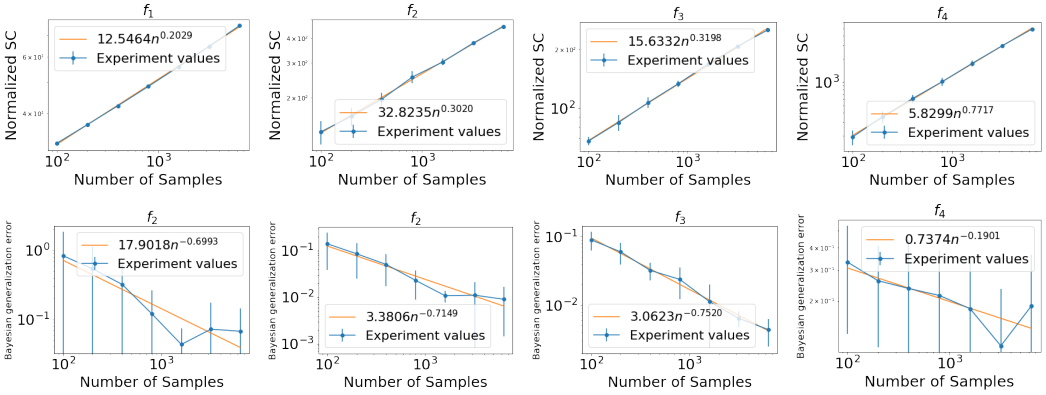

Figure 3: Normalized SC (top) and Bayesian generalization error (bottom) for GPR with kernel $k^{(1)}_{\text{w/ bias}}$ and the target functions in Table 3. The orange curves show the linear regression fit for the experimental values (in blue) of the log Bayesian generalization error as a function of log n.

	function value	β	μ_0	$\mathbb{E}_{\boldsymbol{\epsilon}} F^0(D_n)$	$\mathbb{E}_{\boldsymbol{\epsilon}} G(D_n)$
f_1	$\cos 2\theta$	$+\infty$	0	$\Theta(n^{1/6})$	$\Theta(n^{-5/6})$
f_2	$\text{sign}(\theta)$	1	0	$\Theta(n^{5/6})$	$\Theta(n^{-1/6})$
f_3	$\pi/2 - \lvert\theta\rvert$	2	0	$\Theta(n^{1/2})$	$\Theta(n^{-1/2})$
f_4	$\begin{cases} \pi/2 - \theta, & \theta \in [0, \pi) \\ -\pi/2 - \theta, & \theta \in [-\pi, 0) \end{cases}$	1	> 0	$\Theta(n)$	$\Theta(1)$

Table 4: Target functions used in the experiments for the second order arc-cosine kernel without bias, $k^{(2)}_{\text{w/o bias}}$, their values of β and μ_0, and theoretical rates for the normalized SC and the Bayesian generalization error from our theorems.

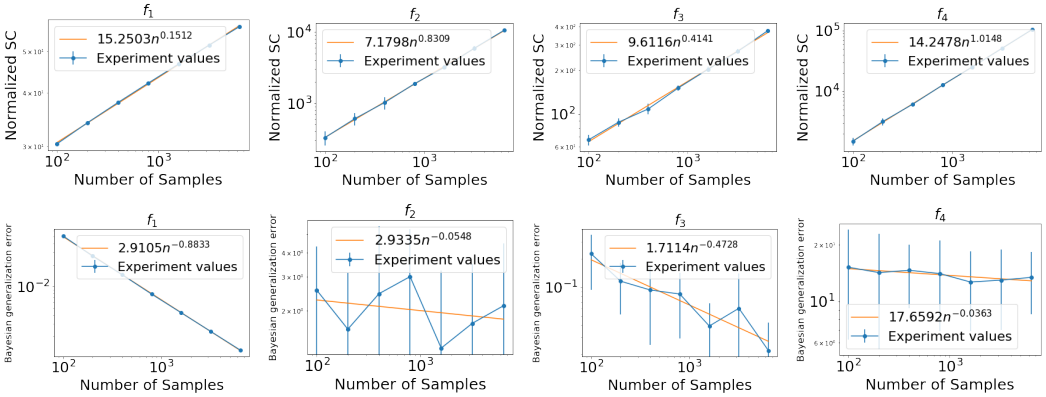

Figure 4: Normalized SC (top) and Bayesian generalization error (bottom) for GPR with kernel $k^{(2)}_{\text{w/o bias}}$ and the target functions in Table 4.

	function value	β	μ_0	$\mathbb{E}_{\boldsymbol{\epsilon}} F^0(D_n)$	$\mathbb{E}_{\boldsymbol{\epsilon}} G(D_n)$
f_1	$\cos 2\theta$	$+\infty$	0	$\Theta(n^{1/6})$	$\Theta(n^{-5/6})$
f_2	θ^2	2	0	$\Theta(n^{1/2})$	$\Theta(n^{-1/2})$
f_3	$(\lvert\theta\rvert - \pi/2)^2$	2	0	$\Theta(n^{1/2})$	$\Theta(n^{-1/2})$
f_4	$\begin{cases} \pi/2 - \theta, & \theta \in [0, \pi) \\ -\pi/2 - \theta, & \theta \in [-\pi, 0) \end{cases}$	1	0	$\Theta(n^{5/6})$	$\Theta(n^{-1/6})$

Table 5: Target functions used in the experiments for the second order arc-cosine kernel with bias, $k^{(2)}_{\text{w/ bias}}$, their values of β and μ_0, and theoretical rates for the normalized SC and the Bayesian generalization error from our theorems.

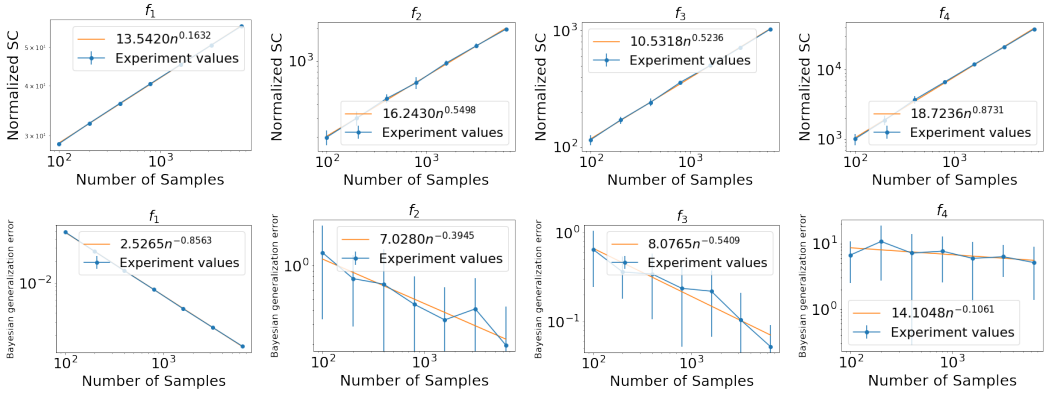

Figure 5: Normalized SC (top) and Bayesian generalization error (bottom) for GPR with kernel $k^{(2)}_{\text{w/ bias}}$ and the target functions in Table 5.

	function value	β	μ_0	$\mathbb{E}_{\boldsymbol{\epsilon}} F^0(D_n)$	$\mathbb{E}_{\boldsymbol{\epsilon}} G(D_n)$
f_1	$\cos 2\theta$	$+\infty$	> 0	$\Theta(n)$	$\Theta(1)$
f_2	$\mathrm{sign}(\theta)$	1	0	$\Theta(n^{1/2})$	$\Theta(n^{-1/2})$
f_3	$\pi/2 - \lvert\theta\rvert$	2	0	$\Theta(n^{1/2})$	$\Theta(n^{-1/2})$
f_4	$\begin{cases} \pi/2 - \theta, & \theta \in [0, \pi) \\ -\pi/2 - \theta, & \theta \in [-\pi, 0) \end{cases}$	1	> 0	$\Theta(n)$	$\Theta(1)$

Table 6: Target functions used in the experiments for the zero order arc-cosine kernel without bias, $k^{(0)}_{\mathrm{w/o\ bias}}$, their values of β and μ_0, and theoretical rates for the normalized SC and the Bayesian generalization error from our theorems.

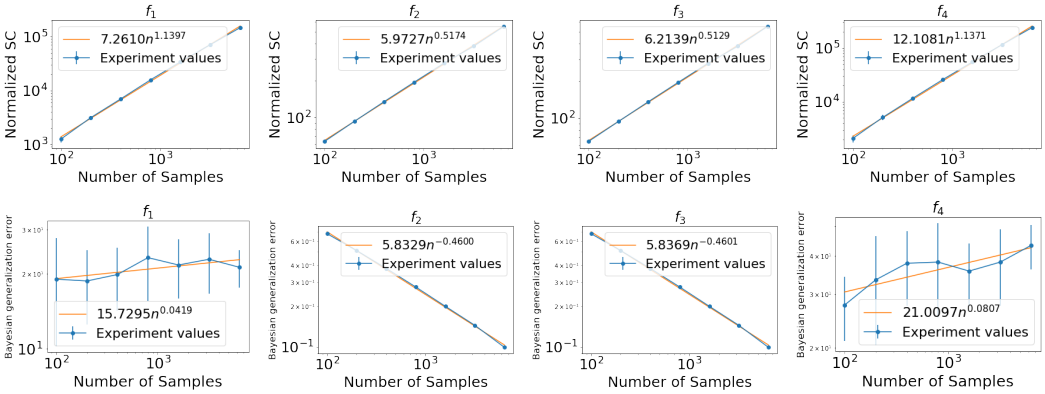

Figure 6: Normalized SC (top) and Bayesian generalization error (bottom) for GPR with kernel $k^{(0)}_{\mathrm{w/o\ bias}}$ and the target functions in Table 6.

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

Table 7: Target functions used in the experiments for the zero order arc-cosine kernel with bias, $k^{(0)}_{\text{w/ bias}}$, their values of β and μ_0, and theoretical rates for the normalized SC and the Bayesian generalization error from our theorems.

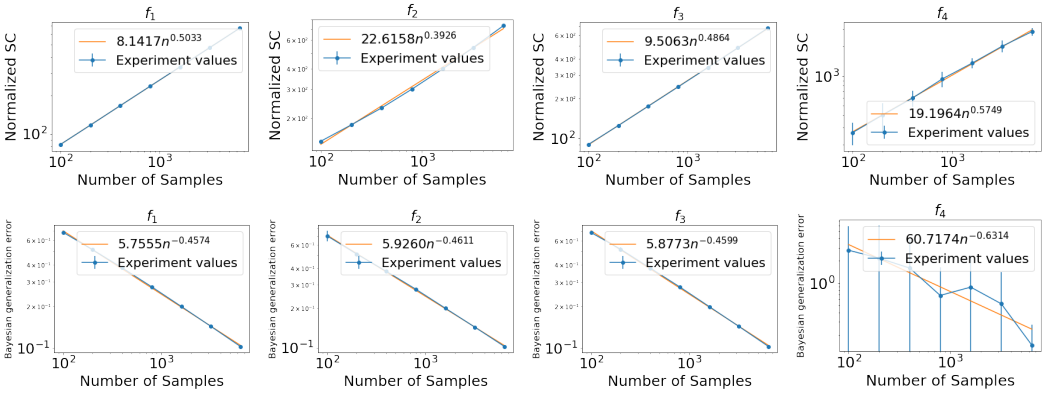

Figure 7: Normalized SC (top) and Bayesian generalization error (bottom) for GPR with kernel $k^{(0)}_{\text{w/ bias}}$ and the target functions in Table 7.