
Learning Expressive Random Feature Models via Parametrized Activations

Zailin Ma^{*1}

Jiansheng Yang¹

Yaodong Yang^{†2}

¹School of Mathematical Sciences, Peking University

²Institute for Artificial Intelligence, Peking University

Abstract

The random feature (RF) method is a powerful kernel approximation technique, but it typically uses fixed activation functions, limiting its adaptability across diverse tasks. To overcome this limitation, we introduce the Random Feature Model with Learnable Activation Functions (RFLAF), a novel statistical model that parameterizes activation functions as weighted sums of basis functions within the random feature framework. Examples of basis functions include radial basis functions (RBFs), spline functions, polynomials, and so forth. For theoretical results, we consider RBFs as representative basis functions. We start with a single RBF as the activation, and then extend the results to multiple RBFs, demonstrating that RF models with a learnable activation component substantially expand the represented function space. We provide estimates on the required number of samples and random features to achieve low excess risk. In our experiments, we test RFLAF with three types of bases: radial basis functions, spline functions and polynomials. Experimental results show that RFLAFs with RBFs and splines consistently outperform other RF models, where RBFs are three times more computationally efficient than splines. We then unfreeze the first-layer parameters and retrain the models, validating the expressivity advantage of learnable activation components on regular two-layer neural networks. Our work provides a deeper understanding of learnable activation components within modern neural network architectures.

1 INTRODUCTION

Kernel methods are powerful tools for solving nonlinear learning problems by leveraging kernel functions to implicitly map data into high-dimensional spaces. However, they can be computationally intensive and lack scalability with large datasets. The random feature (RF) model, introduced in [Rahimi and Recht, 2008b], offers a solution by approximating kernel functions with a finite number of random features, allowing the application of linear algorithms for large-scale computation [Li et al., 2021, Liu et al., 2021].

Despite their advantages, random feature models typically use a fixed activation function, limiting their adaptability during data fitting. This rigidity prevents the model from automatically searching for activation functions for optimal performance across various tasks. Meanwhile, recent work such as Kolmogorov-Arnold Networks (KANs) [Liu et al., 2025b] has demonstrated the effectiveness of learnable activation modules. Therefore, it is natural to study random feature models that incorporate learnable activation functions.

In this paper, we introduce the Random Feature models with Learnable Activation Functions (RFLAF), a novel statistical model that parametrizes learnable activation functions using weighted sums of basis functions within the random feature framework. Examples of basis functions that have universal approximation ability include radial basis functions (RBFs), B-spline functions [Fakhoury et al., 2022], polynomials [Goyal et al., 2019], and so forth. Among them, we consider RBFs as representative basis functions for theoretical analyses. We study the analytic properties of the kernel induced by a single RBF, and then provide approximation and generalization bounds for the multiple-RBF case. For experimental validation, we test RFLAF with RBF, spline functions and polynomials, and compare them with standard RF models across various tasks. Experimental results show that RFLAFs with RBFs or B-splines consistently exhibit greater expressivity and adaptability than standard RF models. More strikingly, RBFs are three times more

^{*}Email: mazailin@stu.pku.edu.cn

[†]Corresponding author: yaodong.yang@pku.edu.cn

computationally efficient than B-splines, achieving the best overall performance. We further unfreeze the first-layer parameters of RFLAFs and retrain the models, validating the expressivity advantage of learnable activation components on regular two-layer neural networks. This paper offers a comprehensive analysis of the model, covering both the theoretical foundations and the empirical validation. Our contributions are summarized as follows.

- We identify an unexplored kernel in the case of a single RBF as the activation. We provide the first result on the analytic form of this kernel, and investigate its representation and approximation properties (Section 3).
- For the general RFLAF, we provide rigorous analyses on the approximation and generalization properties. Our theories guarantee that our model improves its representational ability at a minimal cost of less than twice the number of parameters (Section 4), and the number of random features only needs to scale with the square root of the sample size (Section 5).
- We conduct extensive experiments to demonstrate the advantages of RFLAF (Section 6). We show that RFLAFs with RBFs or splines generally outperform other baseline RF models, with RBFs being three times more efficient than splines. Finally, we validate the effectiveness of the learnable activation function on regular two-layer networks.

The rest of the paper is organized as follows. Section 2 outlines the basics of random feature models and formally introduces our model. Theoretical analyses for models with the single and combined RBF activations are provided in Sections 3 and 4 respectively. Section 5 discusses guarantees on sample complexity, and Section 6 presents experimental results to evaluate our models.

1.1 RELATED WORK

Random Feature (RF) Models Random feature models [Rahimi and Recht, 2008b, 2007] are initially motivated by the fact that randomization is computationally cheaper than optimization [Amit and Geman, 1997, Moosmann et al., 2006]. More recently, by virtue of the relation between a kernel and its Fourier spectral density, random features act as a technique to scale up kernel methods [Lopez-Paz et al., 2014, Sun et al., 2018, Jacot et al., 2018, Arora et al., 2019b, Zandieh et al., 2021, Du et al., 2019, Zamboni et al., 2020, Fu et al., 2024, Shen et al., 2019]. For instance, researchers apply the random feature technique in Transformers to approximate the softmax or Gaussian kernel inside the attention block, successfully reducing the order of the time and space complexity with respect to token length L [Choromanski et al., 2021, Peng et al., 2021]. From a theoretical perspective, when viewed as a class of two-layer neural networks with fixed weights in the first layer, random feature models provide deep insights for partly understanding deep

neural networks [Cao and Gu, 2019, Arora et al., 2019a, Mei et al., 2022, Chizat and Bach, 2020]. Research effort has also been devoted to deriving approximation and generalization bounds with respect to the number of random features and sample size [Sutherland and Schneider, 2015, Rudi and Rosasco, 2017, Avron et al., 2017, Bach, 2017, Mei et al., 2022]. Li et al. [2021] contributes to a unified analysis of random Fourier features.

Learnable Activation Functions Previous work studies the learnable activation functions and attempts to incorporate them into neural network architectures. Activation functions are parametrized in a continuous way as splines [Liu et al., 2025b,a, Fakhoury et al., 2022, Bohra et al., 2020, Aziznejad and Unser, 2019], polynomials [Goyal et al., 2019], sigmoid linear units [Ramachandran et al., 2017] and neural networks [Zhang et al., 2022], or in a discrete way [Bingham and Miikkulainen, 2022]. The related notion of RBF networks introduced in [Lowe and Broomhead, 1988] is fundamentally distinct from our model. In our work, RBFs are used for universal approximation, whereas the RBF network applies them for functional interpolation. Optimal activation functions studied in [Wang and Bento, 2023, Demir and Doğan, 2024] directly relate to our work because RFLAF has the potential to learn optimal activations directly from data.

2 PRELIMINARIES

2.1 BASICS ON RANDOM FEATURE MODELS

In this section, we provide some foundations of random feature models [Rahimi and Recht, 2007, 2008a] related to our work.

Given a function $\sigma(x; w) : \mathcal{X} \times \mathcal{W} \rightarrow \mathbb{R}$. Let μ be a probability measure on $\mathcal{W}$. The class of infinite-width random feature model is defined as

$$\mathcal{F} = \left\{ f : f(x) = \int_{\mathcal{W}} \sigma(x; w) v(w) \mu(dw), v \in \mathcal{H}_{\mathcal{W}} \right\},$$

where $\mathcal{H}_{\mathcal{W}} = \{v(w) : \int_{\mathcal{W}} v(w)^2 \mu(dw) < \infty\}$ is a Hilbert space with norm $\|v\|_{\mathcal{H}_{\mathcal{W}}}^2 = \int_{\mathcal{W}} v(w)^2 \mu(dw)$ and inner product $\langle v, u \rangle_{\mathcal{H}_{\mathcal{W}}} = \int_{\mathcal{W}} v(w) u(w) \mu(dw)$. Furthermore, $\mathcal{F}$ is endowed with a norm $\|\cdot\|_{\mathcal{F}}$ and the inner product $\langle \cdot, \cdot \rangle_{\mathcal{F}} : \|f\|_{\mathcal{F}} = \inf_{f=\langle v, \sigma(\cdot) \rangle_{\mathcal{H}_{\mathcal{W}}}} \|v\|_{\mathcal{H}_{\mathcal{W}}}$, and $\langle f, g \rangle_{\mathcal{F}} = \frac{\|f+g\|_{\mathcal{F}}^2 - \|f-g\|_{\mathcal{F}}^2}{4}$.

Besides, we define the corresponding reproducing kernel $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ as

$$K(x, y) = \int_{\mathcal{W}} \sigma(x; w) \sigma(y; w) \mu(dw).$$

Define the RKHS induced by this kernel as $\mathcal{H}_K$ with corresponding norm $\|\cdot\|_{\mathcal{H}_K}$ and the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_K}$. Generally [Bai and Lee, 2020], for any feature map $\phi : \mathcal{X} \rightarrow \mathcal{H}$

(where $\mathcal{H}$ is a Hilbert space) that induces the kernel K , i.e., $K(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$, we have that for any function f ,

$$\|f\|_{\mathcal{H}_K} = \inf_{f=\langle \mathbf{u}, \phi(\cdot) \rangle_{\mathcal{H}}} \|\mathbf{u}\|_{\mathcal{H}},$$

which indicates the equivalence among different feature maps that generate the same kernel.

Finally, we have the following proposition according to [Minh et al., 2006].

Proposition 2.1 *Given the above definition of $\mathcal{F}$ and $\mathcal{H}_K$, we have that $(\mathcal{F}, \|\cdot\|_{\mathcal{F}}) = (\mathcal{H}_K, \|\cdot\|_{\mathcal{H}_K})$.*

2.2 PARAMETRIZATION OF ACTIVATION FUNCTIONS AND FINITE-WIDTH APPROXIMATION

Standard random feature models consider the case where the activation function σ is a fixed univariate function such as ReLU, and $\sigma(x; w) = \sigma(w^\top x)$. In this work, we broaden the target function class where σ can be any function in $C_c(\mathbb{R})$, namely the continuous functions with compact support.

Let $x \in \mathbb{R}^d$, and $w \sim \mathcal{N}(0, I_d)$. For technical convenience, we assume $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ and $v : \mathbb{R}^d \rightarrow \mathbb{R}$ to be Lipschitz continuous. Suppose that σ is supported on a bounded closed interval $\mathcal{K} \subseteq \mathbb{R}$. We define the target function class as

$$\mathcal{F}_{\mathcal{K}} := \{f : f(x) = \mathbb{E}_{w \sim \mathcal{N}(0, I_d)} [\sigma(w^\top x)v(w)], \quad \sigma \in C_c(\mathcal{K}), \|\sigma\|_{\text{Lip}} \leq L, \|v\|_{\text{Lip}} \leq L_v\}, \quad (1)$$

where $\|\cdot\|_{\text{Lip}}$ denotes the Lipschitz constant of a function.

Suppose the target function $f = \mathbb{E}[\sigma(w^\top x)v(w)] \in \mathcal{F}_{\mathcal{K}}$. The motivations of our model are twofold. In the first step, we consider using an array of basis functions $\{B_i(x)\}_{i \in [N]}$ to approximate any potentially targeted activation functions, i.e., $\tilde{\sigma}(x) := \sum_{i=1}^N a_i B_i(x)$, where a_i are learnable parameters. The corresponding function is then $\tilde{f}(x) := \mathbb{E}_{w \sim \mathcal{N}(0, I_d)} [\tilde{\sigma}(w^\top x)v(w)] \approx f(x)$. In the second step, we approximate $\tilde{f}(x)$ with the finite-width random feature model $\sum_{m=1}^M \tilde{\sigma}(w_m^\top x)v(w_m)/M \approx \mathbb{E}_{w \sim \mathcal{N}(0, I_d)} [\tilde{\sigma}(w^\top x)v(w)]$, where $\{w_m\}_{m=1}^M \stackrel{i.i.d.}{\sim} \mathcal{N}(0, I_d)$ are sampled independently and identically. Consequently, we formulate the Random Feature model with Learnable Activation Functions (RFLAF) as

$$\hat{f}(x; \mathbf{a}, \mathbf{v}) := \frac{1}{M} \sum_{m=1}^M \sum_{i=1}^N a_i B_i(w_m^\top x)v_m, \quad (2)$$

where $\mathbf{a} = (a_1, \dots, a_N) \in \mathbb{R}^N$, $\mathbf{v} = (v_1, \dots, v_M) \in \mathbb{R}^M$ are learnable parameters.

For the theoretical analysis hereafter, we consider the radial basis functions as representatives, namely, we choose the

array of basis functions to be

$$\left\{ B_i(x) = \exp\left(-\frac{(x - c_i)^2}{2h_i^2}\right) \right\}_{i \in [N]}$$

with centers c_i and widths h_i set in prior. The basis $\{B_i(x)\}$ in RFLAF can be replaced by any function class that has universal approximation properties [Nestoridis and Stefanopoulos, 2007] (e.g., B-splines in [Fakhoury et al., 2022], polynomials in [Goyal et al., 2019]).

3 RANDOM FEATURE MODELS WITH A SINGLE RADIAL BASIS FUNCTION

We first study the random feature model with a single radial basis function, which is a special case of RFLAF when $N = 1$. Theorem 3.1 provides the explicit expression of the corresponding kernel, which extends the result of the centered Gaussian activations in [Han et al., 2022] to general non-centered Gaussian activations. Our result is compatible with the former results but provides a more case-specific analysis.

The target function of interest admits representation

$$\varphi(x) = \mathbb{E}_{w \sim \mathcal{N}(0, I_d)} [B(w^\top x)v(w)], \quad (3)$$

where the activation function $B(x) = \exp(-(x - c)^2/(2h^2))$ is a radial basis function with center c and width h . The corresponding reproducing kernel is

$$K(x, x') := \mathbb{E}_{w \sim \mathcal{N}(0, I_d)} [B(w^\top x)B(w^\top x')]. \quad (4)$$

The first result presents the explicit expression of the kernel.

Theorem 3.1 *For any $x, x' \in \mathbb{R}^d$, we have that*

$$K(x, x') = \frac{h^2}{\sqrt{(h^2 + \|x\|^2)(h^2 + \|x'\|^2) - \langle x, x' \rangle^2}} \cdot \exp\left(-\frac{c^2}{2} \cdot \frac{(h^2 + \|x\|^2) + (h^2 + \|x'\|^2) - 2\langle x, x' \rangle}{(h^2 + \|x\|^2)(h^2 + \|x'\|^2) - \langle x, x' \rangle^2}\right). \quad (5)$$

Consider normalized inputs $\|x\|_2 = \|x'\|_2 = 1$, then $r = \langle x, x' \rangle \in [-1, 1]$. The kernel degenerates to a rotation-invariant kernel [Liu et al., 2021]. We slightly abuse the notation and define the univariate function $K(r)$ to be the rotation-invariant form of the kernel (5).

$$K(r) := \frac{h^2}{\sqrt{(1 + h^2)^2 - r^2}} \exp\left(-\frac{c^2}{1 + h^2 + r}\right). \quad (6)$$

We present the explicit expression of kernel (6).

Theorem 3.2 *The rotation-invariant kernel $K(r)$ has Taylor expansion as*

$$K(r) = e^{-p} \frac{h^2}{1+h^2} \sum_{n=0}^{\infty} \frac{R_n(p)}{n!(1+h^2)^n} r^n, \quad (7)$$

where $p = \frac{c^2}{1+h^2} \in [0, +\infty)$, and the polynomials are

$$R_n(x) = \begin{cases} P_k^2(x), & n = 2k, \\ xQ_k^2(x), & n = 2k+1, \end{cases}$$

$$P_k(x) = \sum_{i=0}^k (-1)^{k-i} \frac{(2k-1)!!}{(2i-1)!!} \cdot \binom{k}{i} x^i,$$

$$Q_k(x) = \sum_{i=0}^k (-1)^{k-i} \frac{(2k+1)!!}{(2i+1)!!} \cdot \binom{k}{i} x^i.$$

Therefore, the feature mapping with respect to the kernel (6) is

$$\phi(x) = \left(\frac{he^{-\frac{p}{2}} R_n^{\frac{1}{2}}(p)}{\sqrt{n!(1+h^2)^{n+1}}} x^{\otimes n} \right)_{n=0}^{\infty}.$$

Define the represented function class as

$$\mathcal{H}_{c,h} = \{\varphi : \varphi(x) = \sum_{n=0}^{\infty} \langle F_n, x^{\otimes n} \rangle, D_{c,h}(\varphi) < \infty\},$$

where $F_n \in \mathbb{R}^{d^n}$ and

$$D_{c,h}(\varphi) := \frac{e^p}{h^2} \sum_{n=0}^{\infty} \frac{n!(1+h^2)^{n+1}}{R_n(c^2/(1+h^2))} \|F_n\|_{\text{Fr}}^2.$$

Then we have the following representation theorem.

Corollary 3.3 *For any $f \in \mathcal{H}_{c,h}$, there exists $v : \mathbb{R}^d \rightarrow \mathbb{R}$ such that $f(x) = \mathbb{E}_w [B(w^\top x)v(w)]$ and $\mathbb{E}_w [v(w)^2] \leq D_{c,h}(f)$, where $w \sim \mathcal{N}(0, I_d)$ and $B(x) = \exp\left(-\frac{(x-c)^2}{2h^2}\right)$.*

Approximating φ with finite-width model $\hat{\varphi}(x) = \frac{1}{M} \sum_{m=1}^M B(w_m^\top x)v_m$ where $\{v_m\}_{m \in [M]}$ are learnable parameters, the approximation error can be estimated below.

Theorem 3.4 *Let $v(w)$ be L_v -Lipschitz and $R = \sqrt{2L_v^2 d + 2|v(\mathbf{0})|^2}$. Suppose that $\{w_m\}_{m=1}^M \stackrel{i.i.d.}{\sim} \mathcal{N}(0, I_d)$, then with probability of at least $1 - \delta$, there exists $\{v_m\}_{m=1}^M$ such that*

$$\mathbb{E}_x |\hat{\varphi}(x) - \varphi(x)| \leq 18R \sqrt{\log(4/\delta)/M},$$

and $\sum_{m=1}^M v_m^2 \leq 49MR^2 \log(2/\delta)$, where we assume $\delta < 1/2$. Note that the inequalities hold for whatever distribution x are sampled from.

Proofs of all the above statements are provided in Appendix B. The proof of Theorem 3.4 is not trivial, because the concentration property of $|\hat{\varphi}(x) - \varphi(x)|$ may not be uniform over x . We use some techniques to circumvent this problem.

Implied by Theorem 3.2 and Corollary 3.3, the represented function f corresponds to a fixed feature mapping with fast decaying coefficients F_n , indicating a narrow function class similar to other standard RF models. Hence, using a single RBF as the activation function does not necessarily lead to a leap in the expressivity of the RF model. The key step is to combine the RBFs with learnable weights. The mechanism of learnable activation functions results in evidently enhanced expressivity of RF models, as we will demonstrate in the next section.

4 RANDOM FEATURE MODELS WITH LEARNABLE ACTIVATION FUNCTIONS

This section provides the result on the approximation error between the RFLAF

$$\hat{f}(x) = \frac{1}{M} \sum_{m=1}^M \sum_{i=1}^N a_i B_i(w_m^\top x) v_m$$

of multiple RBFs and the target function $f^* \in \mathcal{F}_{\mathcal{K}}$ defined in Section 2.2. In the following, we denote $B_i(x) = \exp(-(x - c_i)^2/(2h_i^2))$. We recall the Gaussian universal approximation theorem in [Bacharoglou, 2010, Nestoridis and Stefanopoulos, 2007].

Gaussian Universal Approximation Theorem (Gaussian UAT) Suppose the target function $\sigma(x)$ is a continuous function with compact support $\mathcal{K}$. For any $\epsilon > 0$, there exists $N > 0$ and $\{h_i, c_i, a_i\}_{i=1}^N$ such that $\|\sigma(x) - \sum_{i=1}^N a_i B_i(x)\|_{\infty} < \epsilon$.

Inspired by the theorem, to bridge the gap between $\hat{f}$ and f^* , we consider an intermediate function

$$\tilde{f}(x) := \mathbb{E}_w \left[\sum_{i=1}^N a_i B_i(w^\top x) v(w) \right],$$

where $\{a_i\}_{i \in [N]}$ are learnable and $\{c_i, h_i\}_{i \in [N]}$ are set in prior. We expect that $\sum_{i=1}^N a_i B_i(x)$ in $\tilde{f}$ can approximate the ground truth $\sigma(x)$ to an arbitrarily low error. To describe c_i and h_i precisely, we partition the support set $\mathcal{K}$ of σ .

Let the grid number be N . We define the grid points as $y_0 = \min_{x \in \mathcal{K}} x$, $y_N = \max_{x \in \mathcal{K}} x$ and $y_i = y_0 + \frac{i}{N}(y_N - y_0)$ for $1 \leq i \leq N-1$. The diameter of the support is $|\mathcal{K}| := y_N - y_0$. The grid size then is $|\mathcal{K}|/N$. Because σ is continuous over the compact set $\mathcal{K}$, it is also bounded. Hence, $\|\sigma\|_{\infty} < \infty$. Our first result measures the approximation error between f^* and $\tilde{f}$ with respect to the choice of h_i and grid size.

Proposition 4.1 Suppose $f^* \in \mathcal{F}_\mathcal{K}$ with activation function σ . For any $\epsilon > 0$, by setting

$$h \leq \frac{\epsilon}{4\sqrt{2}LR\sqrt{\log \frac{16\|\sigma\|_\infty R}{\epsilon}}},$$

$$\frac{|\mathcal{K}|}{N} \leq \frac{\epsilon h \sqrt{\pi e}}{16\sqrt{2}\|\sigma\|_\infty R \log \left(\frac{8\|\sigma\|_\infty |\mathcal{K}| R}{\sqrt{2\pi e} h^2} \right)} \wedge \frac{\epsilon}{4LR},$$

and $h_i = h$, $c_i \in [y_{i-1}, y_i]$, there exists $\{a_i\}_{i=1}^N$ such that

$$\|\tilde{f}(x) - f^*(x)\|_\infty \leq \epsilon, \quad \sum_{i=1}^N |a_i|^2 \leq \frac{\|\sigma\|_\infty^2 |\mathcal{K}|^2}{2\pi h^2 N}.$$

We use the notation $a \wedge b := \min\{a, b\}$ for brevity. The proof of Proposition 4.1 is contained in Appendix C.1. Now we are ready to measure the approximation error between f^* and $\hat{f}$ in the sense of L_1 norm.

Theorem 4.2 Suppose $f^* \in \mathcal{F}_\mathcal{K}$ with activation function σ . For all $\epsilon > 0$, under the parameter settings of Proposition 4.1, let $\{w_m\}_{m=1}^M \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1)$, then with probability of at least $1 - \delta$, there exists $\{a_i\}_{i=1}^N$ and $\{v_m\}_{m=1}^M$ such that

$$\mathbb{E}_x |\hat{f}(x) - f^*(x)| \leq \frac{18(\|\sigma\|_\infty R + \epsilon) \sqrt{\log(4/\delta)}}{\sqrt{M}} + \epsilon,$$

and

$$\sum_{i=1}^N a_i^2 \leq \frac{\|\sigma\|_\infty^2 |\mathcal{K}|^2}{2\pi h^2 N}, \quad \sum_{m=1}^M v_m^2 \leq 49MR^2 \log \left(\frac{2}{\delta} \right),$$

where we assume $\delta < 1/2$.

The proof of Theorem 4.2 is contained in Appendix C.2. Theorem 4.2 indicates that to obtain $O(\epsilon)$ approximation error, the model requires $M = \Theta(1/\epsilon^2)$. Moreover, Proposition 4.1 indicates that $1/h = \tilde{\Theta}(1/\epsilon)^1$ and $N = \tilde{\Theta}(1/\epsilon h)$. Hence, $N = \tilde{\Theta}(1/\epsilon^2) = \tilde{\Theta}(M)$. The number of grid points N should scale with approximately the same order of M .

In practice, however, we find that a very humble number of basis functions N are sufficient to improve the expressivity of the model to a large extent. For instance, RFLAF with $N = 16$ already outperforms standard RF models across various tasks (see section 6). To summarize, RFLAF gains enhanced expressivity with very minor increase in parameter number.

5 GENERALIZATION BOUNDS AND SAMPLE COMPLEXITY OF LEARNING

To complete the theoretical analysis of the model, we provide the worst-case analysis regarding the generalization bounds of learning in this section.

¹ $\tilde{\Theta}(\cdot)$ stands for $\Theta(\cdot)$ but hides the logarithmic terms

Suppose the data distribution is $\mathcal{P}$ and the samples are $S = \{(x_i, y_i)\}_{i=1}^n \stackrel{i.i.d.}{\sim} \mathcal{P}$. Suppose the loss function $\ell(\hat{y}, y)$ is ρ -Lipschitz in $\hat{y}$ and $|\ell(0, y)| \leq \rho$ for any y (a common setting as in [Li et al., 2021]). The population risk and the empirical risk are defined respectively as

$$L_D(f) := \mathbb{E}_{x, y \sim \mathcal{P}}[\ell(f(x), y)], \quad L_S(f) := \frac{1}{n} \sum_{i=1}^n \ell(f(x_i), y_i).$$

The minimizer of the population risk is

$$f^* := \operatorname{argmin}_{f \in \mathcal{F}_\mathcal{K}} L_D(f). \quad (8)$$

Under the setting of Theorem 4.2, suppose $\{w_m\}_{m=1}^M \stackrel{i.i.d.}{\sim} \mathcal{N}(0, I_d)$ are sampled, and h , N and $\{c_i\}_{i \in [N]}$ are fixed. We aim at learning the parameters $V = (\mathbf{a}, \mathbf{v}) = (a_1, \dots, a_N, v_1, \dots, v_M)$ in $f_V(x) := \frac{1}{M} \sum_{m=1}^M \sum_{i=1}^N a_i B_i(w_m^\top x) v_m$. Guided by Theorem 4.2, the constrained set is set to be

$$\mathcal{V} := \left\{ V = (\mathbf{a}, \mathbf{v}) \in \mathbb{R}^N \times \mathbb{R}^M : \|\mathbf{a}\|_2 \leq \frac{\|\sigma\|_\infty |\mathcal{K}|}{h\sqrt{2\pi N}}, \right. \\ \left. \|\mathbf{v}\|_2 \leq 7R \sqrt{M \log \left(\frac{2}{\delta} \right)} \right\}. \quad (9)$$

Denote $f_\mathcal{V} = \{f_V\}_{V \in \mathcal{V}}$. The minimizer of the empirical risk is

$$f_S := \operatorname{argmin}_{f_V \in f_\mathcal{V}} L_S(f_V). \quad (10)$$

Theorem 5.1 Under the conditions and parameter settings of h , N , $\{c_i\}_{i=1}^N$ in Theorem 4.2, let f^* and f_S be the minimizers of the population risk and the empirical risk in Eq. (8) and (10) respectively. For all $\epsilon > 0$, with probability of at least $1 - \delta$ over $\{w_m\}_{m=1}^M \stackrel{i.i.d.}{\sim} \mathcal{N}(0, I_d)$ and $\{(x_i, y_i)\}_{i=1}^n \stackrel{i.i.d.}{\sim} \mathcal{P}$, the excess risk is bounded by

$$L_D(f_S) - L_D(f^*) \leq \frac{\rho C \log(16/\delta)}{h\sqrt{n}} + \frac{\rho C \sqrt{\log(8/\delta)}}{\sqrt{M}} + \rho\epsilon,$$

where $C = \max\{14(1 + 7\|\sigma\|_\infty |\mathcal{K}| R), 18(\|\sigma\|_\infty R + \epsilon)\}$, and we assume $\delta \leq 1/2$, $h \leq 1$.

The proof is contained in Appendix D, which is mainly reduced to estimating the Rademacher complexity of the function class induced by the constrained set (9).

Theorem 5.1 implies that to achieve $O(\epsilon)$ excess risk, it suffices to have the sample size n , the random feature number M and the grid number N to scale as

$$n = \tilde{\Theta}(1/\epsilon^2 h^2), \quad M = \Theta(1/\epsilon^2), \quad N = \tilde{\Theta}(1/\epsilon h).$$

Indicated by Proposition 4.1, we set h such that $1/h = \tilde{\Theta}(1/\epsilon)$. Hence, even in the worst case, only $M = \tilde{\Theta}(\sqrt{n})$

Table 1: Test Losses of Random Feature Models. $N = 16$ for RFLAFs. Results are reported as mean $\pm$ std. Best in bold. Second best in italics.

DATA SET	RFLAF			RFMLP			
	RBF	BS	PL	RELU	COS	TANH	SIGMOID
MNIST	<i>0.126$\pm$0.011</i>	0.165 $\pm$ 0.021	0.124$\pm$0.003	0.159 $\pm$ 0.002	1.390 $\pm$ 0.055	0.277 $\pm$ 0.003	0.498 $\pm$ 0.017
CIFAR-10	1.450$\pm$0.011	1.609 $\pm$ 0.025	1.482 $\pm$ 0.016	<i>1.466$\pm$0.004</i>	2.641 $\pm$ 0.125	1.769 $\pm$ 0.003	1.930 $\pm$ 0.028
ADULT	0.309$\pm$0.002	0.324 $\pm$ 0.005	-	<i>0.311$\pm$0.002</i>	0.363 $\pm$ 0.020	0.324 $\pm$ 0.003	0.327 $\pm$ 0.004
PROTEIN	<i>0.204$\pm$0.003</i>	0.194$\pm$0.002	-	0.241 $\pm$ 0.001	0.371 $\pm$ 0.022	0.650 $\pm$ 0.009	0.280 $\pm$ 0.025
CT	0.212$\pm$0.016	<i>0.302$\pm$0.038</i>	-	0.356 $\pm$ 0.096	0.589 $\pm$ 0.060	1.241 $\pm$ 0.133	0.692 $\pm$ 0.020
WORKLOADS	0.465$\pm$0.035	<i>0.546$\pm$0.014</i>	-	2.771 $\pm$ 0.039	2.634 $\pm$ 0.015	24.997 $\pm$ 0.265	1.707 $\pm$ 0.015
MILLIONSONGS	0.102$\pm$0.001	0.120 $\pm$ 0.002	-	0.951 $\pm$ 0.009	0.280 $\pm$ 0.094	8.434 $\pm$ 0.091	<i>0.118$\pm$0.007</i>

number of random features are required, matching the sharpest results on the number of features presented in [Li et al., 2021, Rudi and Rosasco, 2017] for standard random feature models.

6 NUMERICAL EXPERIMENTS

We test the models on seven real-world datasets, including three classification tasks (MNIST, CIFAR-10 and adult) and four large-scale UCI regression datasets (protein, ct, workloads and millionsongs), evaluated with squared error loss and cross-entropy loss respectively. For all random feature models, experiments are repeated for 10 different seeds and quantities including losses and time are averaged to provide statistical confidence.

For RFLAF, we set the grid range $\mathcal{K} = [-2, 2]$, consider a list of grid number $N = 8, 16, 32, 64, 128$ and set $h = 4/N$. Proposition 4.1 indicates that c_i can be arbitrarily chosen within each interval $[y_{i-1}, y_i]$, so we choose c_i to be on the grid points. The model width is $M = 1000$ for MNIST and $M = 3000$ for the other datasets.

6.1 OPTIMIZATION SETUP

We formulate the learning problem (10) in the case of MSE loss for regression tasks as an unconstrained optimization problem:

$$\min_{\mathbf{a}, \mathbf{v}} \frac{1}{n} \sum_{i=1}^n (\mathbf{a}^\top \mathbf{B}(x_i) \mathbf{v} - y_i)^2 + \lambda_1 (\|\mathbf{a}\|_2^2 - \|\mathbf{v}\|_2^2)^2 + \lambda_2 \|\mathbf{a}\|_1, \quad (11)$$

where $(\mathbf{B}(x_i))_{k,m} = B_k(w_m x_i)$.

The problem without the L_1 term is recognized as the matrix sensing problem, a canonical optimization problem in low-rank matrix factorization [Chi et al., 2019, Li et al., 2018, Tu et al., 2016]. The first regularizer $\mathcal{R}_1 := (\|\mathbf{a}\|_2^2 - \|\mathbf{v}\|_2^2)^2$ is

necessary to guarantee convergence by balancing the length of the two vectors, otherwise a longer $\mathbf{a}$ leads to a shorter $\mathbf{v}$ and the number of solutions can be infinite. The second regularizer is the common L_1 regularizer and is not strictly necessary. It aims to obtain sparse components for $\mathbf{a}$. For classification tasks, we substitute the MSE loss function with cross-entropy loss.

6.2 BASELINE COMPARISONS

We consider two types of RF models. The first type is RFLAFs with RBFs, B-splines of degree two (BS) and Taylor polynomials (PL). In both cases, N represents the number of the basis functions. The second type is the plain random feature models (RFMLP) with fixed activation functions (RELU, COS, TANH, SIGMOID). All models are compared within *the same width* M .

Table 2: Time Comparison between RBF and BS. $N = 16$. The values of train time and test time of RFMLP with ReLU are set to be 1 respectively.

DATA SET	TRAIN TIME		TEST TIME	
	RBF	BS	RBF	BS
MNIST	1.1	1.5	1.1	1.4
CIFAR-10	1.1	1.4	1.0	1.4
ADULT	1.7	4.2	2.1	6.2
PROTEIN	1.8	4.7	2.0	6.3
CT	1.6	3.9	1.6	4.4
WORKLOADS	1.3	3.4	1.4	4.4
MILLIONSONGS	2.0	5.1	2.4	7.3

Table 1 showcases the test losses among all RF models. For brevity, we will refer to various RF models as their

activation functions (e.g., RFLAF with RBFs are referred to as RBF). Several observations can be made:

(1) RFLAFs consistently outperform RFMLPs in all tasks. The results of accuracies for classification tasks in Appendix E.2 are similar. Because RFMLP represents fixed feature mappings, it inevitably performs well in tasks that match its feature mapping but poorly in other tasks. In contrast, the learnable activation module in RFLAF allows the model to adaptively fits the data.

(2) Specifically, RBF and BS achieves the best performances among all RF models. Furthermore, RBF shows 2 ~ 4 times faster computational efficiency than RBF (Table 2). As a result, RFLAF with RBF presents the best performance comprehensively. For PL, the models suffer from training instability due to the exploding magnitude of x^n when n is large, so they fail to converge in the last five tasks. To further substantiate the results, we also provide comparative results on low-degree polynomials that converge successfully, and results on RFLAF of $N = 8, 32, 64, 128$. In all cases, RFLAF presents consistent superiority, especially with RBF (see Appendix E.2).

(3) Finally, we highlight that the performance improvement of RBF is significant, but the cost of extra time consumption and extra parameter number compared to those of RFMLPs are actually minimal. For instance, in *workloads*, RBF improves the test losses compared to the best of RFMLP (SIGMOID) by around 70%, with training and testing time increased by no more than 30%. RFLAF with RBFs enhances standard RF models with minimal cost, showcasing the potential of RBF-based learnable activation module in enhancing modern neural network structures.

6.3 MODEL PERFORMANCE WITH RESPECT TO GRID NUMBER

In this part, we study how the performance of RFLAF (RBF) evolves as the grid number N increases.

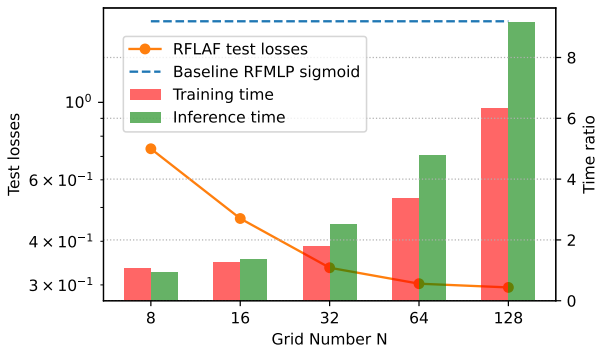


Figure 1: Test Losses and Time Consumption with respect to N on Dataset *workloads*.

Figure 1 shows that as N grows, the test error declines until limited by the representative function class. This fact substantiates Proposition 4.1 that finer grid size results in a lower approximation error. Figure 1 also shows that the running time grows linearly with N as the time complexity of the model is $O(MN)$. However, for $N = 8, 16, 32$, the model performance significantly improves as N increases, whilst the time consumption is acceptable (almost the same as RFMLP when $N = 8, 16$, around 2 times when $N = 32$). Results on other datasets are similar (see Appendix E.2).

Table 3: Epoch Number of Convergence.

DATA SET	RFLAF (RBF) WITH $N =$				
	8	16	32	64	128
MNIST	10	9	4	3	3
CIFAR-10	8	3	2	2	2
ADULT	14	4	2	1	1

Moreover, larger N equips the model with faster kernel learning ability. In Table 3, we consider three tasks where overfitting occurs, and record the epoch number when the model reaches the lowest test error. Table 3 shows that the epoch number of convergence declines as N increases. This indicates that the model with larger N represents a broader class of function, and hence the model is able to converge to the local minimizer very fast in the training phase.

To summarize, a small grid number such as $N = 16, 32, 64$ is sufficient for the model to fit the real data well, and is probably a good trade-off between model performance and time efficiency.

6.4 ABILITY TO APPROXIMATE THE OPTIMAL ACTIVATION FUNCTION

To further examine the kernel learning ability of RFLAF with RBF, we design synthetic functions to check if RFLAF successfully recovers the ground-truth activation function from data. We choose target functions to be of the form $f(x) = \mathbb{E}_{w \sim \mathcal{N}(0, I_d)} [\sigma(w^\top x) v(w)]$, where $x \in \mathbb{R}^d$ and $d = 2$. For RFLAF with RBF, we set $M = 1000$ and $N = 400$, $h = 0.005$ for the precise learning of the optimal activation. We show the result of

$$\sigma(x) = -\sin(\pi(x + 0.5))\mathbf{1}_{[-1.5, -0.5]} + \sin(\pi(x - 0.5))\mathbf{1}_{[0.5, 1.5]}$$

in Figure 2 as an example.

The blue dotted line represents the learned activation functions and the orange solid line represents the ground-truth function. RFLAF successfully recovers the ground truth activation functions σ through learning f , demonstrating its ability to directly learn from data the optimal activation function and skip the step for grid searching the optimal

Table 4: Test Losses of Regular Two-layer Networks. $N = 16$ for LAN and KAN.

DATA SET	LOSS				
	LAN			MLP	KAN
	RBF	BS	PL	RELU	
MNIST	0.124 ± 0.040	0.105 ± 0.003	0.181 ± 0.003	0.166 ± 0.003	0.161 ± 0.003
CIFAR-10	1.349 ± 0.010	1.375 ± 0.007	1.484 ± 0.115	1.449 ± 0.005	1.475 ± 0.003
ADULT	0.302 ± 0.001	0.308 ± 0.002	-	0.305 ± 0.000	0.305 ± 0.000
PROTEIN	0.220 ± 0.002	0.195 ± 0.002	-	0.234 ± 0.001	0.372 ± 0.000
CT	0.026 ± 0.000	0.023 ± 0.002	-	0.031 ± 0.001	4.938 ± 0.000
WORKLOADS	0.285 ± 0.012	0.266 ± 0.021	0.956 ± 0.276	2.090 ± 0.025	7.207 ± 0.000
MILLIONSONGS	0.077 ± 0.002	0.082 ± 0.001	-	0.550 ± 0.007	0.367 ± 0.000

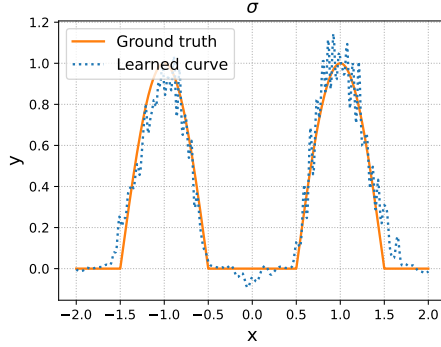


Figure 2: The Activation Function Learned in RFLAF.

parameters of the activation function as proposed in [Demir and Doğan, 2024]. We repeat the experiment with another two synthetic functions. The results are similar and are provided in Appendix E.1.

6.5 COMPARISONS AMONG THE REGULAR TWO-LAYER NETWORKS

In the last part, we unfreeze the first-layer parameters of all RF models, where RFLAF becomes Learnable Activation Networks (LAN), and test them over all datasets. We include the novel neural network KAN for comparison. The width of KAN is set M/N to ensure the same parameter number among models.

Table 4 shows that LAN with RBF or BS generally achieves the best performance, and Table 5 shows that RBF runs about 3 times faster than BS, consistent with the results of the random feature cases. We highlight that LAN, as an intermediate form between MLP and KAN, showcases its unique characteristics. KAN has superior interpretability in small-scale and science-related tasks, however, it suffers from scaling problem that LAN handles more easily. On

the other side, LAN enhances the capability of MLP with economic increase in time and parameter number. This result further substantiates the potential of learnable activation in modern neural network structure. More discussions are provided in Appendix E.3.

Table 5: Test Time Comparison among Regular Networks. $N = 16$ for LAN and KAN.

DATA SET	TEST TIME			
	LAN		MLP	KAN
	RBF	BS	RELU	
MNIST	1.1	1.4	1.0	1.1
CIFAR-10	1.1	1.3	1.0	1.1
ADULT	2.2	6.0	1.0	1.9
PROTEIN	1.7	5.1	1.0	1.8
CT	1.8	4.6	1.0	1.7
WORKLOADS	2.0	5.5	1.0	1.7
MILLIONSONGS	1.9	5.2	1.0	1.7

7 CONCLUSION

In this work, we propose the Random Feature Model with Learnable Activation Functions (RFLAF), which parameterizes activation functions as weighted sums of basis functions. Theoretically, we provide the first analytic form of the kernel induced by a single RBF as the activation and establish approximation and generalization bounds for RFLAF. Experimentally, RFLAF with RBFs or B-splines consistently outperforms standard RF models, with RBFs being three times more efficient. Our work deepens the understanding of learnable activation functions in neural architectures and opens future directions for random feature models.

Acknowledgements

This work is sponsored by the National Natural Science Foundation of China (62376013, 623B2003, 624B100026). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding agencies.

References

- Yali Amit and Donald Geman. Shape quantization and recognition with randomized trees. *Neural computation*, 9(7):1545–1588, 1997.
- Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pages 322–332. PMLR, 2019a.
- Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. *Advances in neural information processing systems*, 32, 2019b.
- Haim Avron, Michael Kapralov, Cameron Musco, Christopher Musco, Ameya Velingker, and Amir Zandieh. Random fourier features for kernel ridge regression: Approximation bounds and statistical guarantees. In *International conference on machine learning*, pages 253–262. PMLR, 2017.
- Shayan Aziznejad and Michael Unser. Deep spline networks with control of lipschitz regularity. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3242–3246. IEEE, 2019.
- Francis Bach. On the equivalence between kernel quadrature rules and random feature expansions. *Journal of machine learning research*, 18(21):1–38, 2017.
- Athanassia Bacharoglou. Approximation of probability distributions by convex mixtures of gaussian measures. *Proceedings of the American Mathematical Society*, 138(7):2619–2628, 2010.
- Yu Bai and Jason D. Lee. Beyond linearization: On quadratic and higher-order approximation of wide neural networks. 2020.
- Garrett Bingham and Risto Miikkulainen. Discovering parametric activation functions. *Neural Networks*, 148:48–65, 2022.
- Pakshal Bohra, Joaquim Campos, Harshit Gupta, Shayan Aziznejad, and Michael Unser. Learning activation functions in deep (spline) neural networks. *IEEE Open Journal of Signal Processing*, 1:295–309, 2020.
- Yuan Cao and Quanquan Gu. Generalization bounds of stochastic gradient descent for wide and deep neural networks. *Advances in neural information processing systems*, 32, 2019.
- Yuejie Chi, Yue M Lu, and Yuxin Chen. Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269, 2019.
- Lenaic Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Conference on learning theory*, pages 1305–1338. PMLR, 2020.
- Krzysztof Marcin Choromanski, Valerii Likhoshesterov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Quincy Davis, Afroz Mohiuddin, Lukasz Kaiser, David Benjamin Belanger, Lucy J Colwell, and Adrian Weller. Rethinking attention with performers. 2021.
- Samet Demir and Zafer Doğan. Optimal nonlinearities improve generalization performance of random features. In *Asian Conference on Machine Learning*, pages 247–262. PMLR, 2024.
- Simon S Du, Kangcheng Hou, Russ R Salakhutdinov, Barnabas Poczos, Ruosong Wang, and Keyulu Xu. Graph neural tangent kernel: Fusing graph neural networks with graph kernels. *Advances in neural information processing systems*, 32, 2019.
- Daniele Fakhoury, Emanuele Fakhoury, and Hendrik Speleers. Exspline: An interpretable and expressive spline-based neural network. *Neural Networks*, 152:332–346, 2022.
- Hengyu Fu, Tianyu Guo, Yu Bai, and Song Mei. What can a single attention layer learn? a study through the random features lens. *Advances in Neural Information Processing Systems*, 36, 2024.
- Mohit Goyal, Rajan Goyal, and Brejesh Lall. Learning activation functions: A new paradigm for understanding neural networks. *arXiv preprint arXiv:1906.09529*, 2019.
- Insu Han, Amir Zandieh, Jaehoon Lee, Roman Novak, Lechao Xiao, and Amin Karbasi. Fast neural kernel embeddings for general activations. *Advances in neural information processing systems*, 35:35657–35671, 2022.
- Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.
- Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Conference On Learning Theory*, pages 2–47. PMLR, 2018.

- Zhu Li, Jean-Francois Ton, Dino Oglic, and Dino Sejdinovic. Towards a unified analysis of random fourier features. *Journal of Machine Learning Research*, 22(108):1–51, 2021.
- Fanghui Liu, Xiaolin Huang, Yudong Chen, and Johan AK Suykens. Random features for kernel approximation: A survey on algorithms, theory, and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7128–7148, 2021.
- Ziming Liu, Max Tegmark, Pingchuan Ma, Wojciech Matysik, and Yixuan Wang. Kolmogorov-arnold networks meet science. 2025a.
- Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljacic, Thomas Hou, and Max Tegmark. Kan: Kolmogorov–arnold networks. *International conference on learning representations*, 2025: 70367–70413, 2025b.
- David Lopez-Paz, Suvrit Sra, Alex Smola, Zoubin Ghahramani, and Bernhard Schölkopf. Randomized nonlinear component analysis. In *International conference on machine learning*, pages 1359–1367. PMLR, 2014.
- David Lowe and D Broomhead. Multivariable functional interpolation and adaptive networks. *Complex systems*, 2(3):321–355, 1988.
- Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Generalization error of random feature and kernel methods: hypercontractivity and kernel matrix concentration. *Applied and Computational Harmonic Analysis*, 59:3–84, 2022.
- Ha Quang Minh, Partha Niyogi, and Yuan Yao. Mercer’s theorem, feature maps, and smoothing. In *International Conference on Computational Learning Theory*, pages 154–168. Springer, 2006.
- Frank Moosmann, Bill Triggs, and Frederic Jurie. Randomized clustering forests for building fast and discriminative visual vocabularies. NIPS, 2006.
- Vassili Nestoridis and Vangelis Stefanopoulos. Universal series and approximate identities. Technical report, Citeseer, 2007.
- Hao Peng, Nikolaos Pappas, Dani Yogatama, Roy Schwartz, Noah Smith, and Lingpeng Kong. Random feature attention. *International Conference on Learning Representations*, 2021.
- Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007.
- Ali Rahimi and Benjamin Recht. Uniform approximation of functions with random bases. In *2008 46th annual allerton conference on communication, control, and computing*, pages 555–561. IEEE, 2008a.
- Ali Rahimi and Benjamin Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. *Advances in neural information processing systems*, 21, 2008b.
- Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017.
- Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features. *Advances in neural information processing systems*, 30, 2017.
- Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.
- Yanning Shen, Tianyi Chen, and Georgios B Giannakis. Random feature-based online multi-kernel learning in environments with unknown dynamics. *Journal of Machine Learning Research*, 20(22):1–36, 2019.
- Yitong Sun, Anna Gilbert, and Ambuj Tewari. But how does it work in theory? linear svm with random features. *Advances in Neural Information Processing Systems*, 31, 2018.
- Danica J. Sutherland and Jeff Schneider. On the error of random fourier features. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, UAI’15, page 862–871, Arlington, Virginia, USA, 2015. AUAI Press. ISBN 9780996643108.
- Stephen Tu, Ross Boczar, Max Simchowitz, Mahdi Soltanolkotabi, and Ben Recht. Low-rank solutions of linear matrix equations via procrustes flow. In *International conference on machine learning*, pages 964–973. PMLR, 2016.
- Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.
- Jianxin Wang and José Bento. Optimal activation functions for the random features regression model. In *The Eleventh International Conference on Learning Representations*, 2023.
- Daniele Zambon, Cesare Alippi, and Lorenzo Livi. Graph random neural features for distance-preserving graph representations. In *International Conference on Machine Learning*, pages 10968–10977. PMLR, 2020.

Amir Zandieh, Insu Han, Haim Avron, Neta Shoham, Chaewon Kim, and Jinwoo Shin. Scaling neural tangent kernels via sketching and random features. *Advances in Neural Information Processing Systems*, 34:1062–1073, 2021.

Shijun Zhang, Zuowei Shen, and Haizhao Yang. Neural network architecture beyond width and depth. *Advances in Neural Information Processing Systems*, 35:5669–5681, 2022.

Learning Expressive Random Feature Models via Parametrized Activations (Supplementary Material)

Zailin Ma¹

Jiansheng Yang¹

Yaodong Yang²

¹School of Mathematical Sciences, Peking University

²Institute for Artificial Intelligence, Peking University

A TECHNICAL TOOLS

A.1 BASICS ON SUB-GAUSSIAN AND SUB-EXPONENTIAL RANDOM VARIABLES

Definition A.1 A random variable Y is a sub-gaussian random variable if there exists $K > 0$ such that $\mathbb{E} \exp(Y^2/K^2) \leq 2$. Define the sub-gaussian norm as $\|Y\|_{\psi_2} := \inf\{K > 0 : \mathbb{E} \exp(Y^2/K^2) \leq 2\}$.

Definition A.2 A random variable Y is a sub-exponential random variable if there exists $K > 0$ such that $\mathbb{E} \exp(|Y|/K) \leq 2$. Define the sub-exponential norm as $\|Y\|_{\psi_1} := \inf\{K > 0 : \mathbb{E} \exp(|Y|/K) \leq 2\}$.

Lemma A.3 If Y is a sub-gaussian random variable, then $\|Y^2\|_{\psi_1} = \|Y\|_{\psi_2}^2$.

The following properties of sub-gaussian random variable are stated in Proposition 2.5.2 in [Vershynin, 2018]. For this paper to be self-contained, we also state them here with explicit constants.

Lemma A.4 Suppose Y is a random variable.

1. If Y is a sub-gaussian random variable, then $P(|Y| \geq \epsilon) \leq 2 \exp(-\epsilon^2/\|Y\|_{\psi_2}^2)$.
2. If $P(|Y| \geq \epsilon) \leq 2 \exp(-\epsilon^2/K^2)$, then $\|Y\|_{\psi_2} \leq \sqrt{2}K$.

Lemma A.5 Suppose Y is a random variable.

1. If there exist $K_0 > 0$ such that $\mathbb{E} e^{\lambda^2 Y^2} \leq e^{K_0^2 \lambda^2}$ for all $\lambda^2 \leq 1/K_0^2$, then Y is a sub-gaussian random variable with sub-gaussian norm $\|Y\|_{\psi_2} \leq K_0/\sqrt{\log 2} \leq \sqrt{2}K_0$.
2. If Y is a sub-gaussian random variable, then $K_0 = 2\|Y\|_{\psi_2}$ such that $\mathbb{E} e^{\lambda^2 Y^2} \leq e^{K_0^2 \lambda^2}$ for all $\lambda^2 \leq 1/K_0^2$.

Lemma A.6 Suppose Y is a random variable and $\mathbb{E}Y = 0$.

1. If $\mathbb{E} e^{\lambda^2 Y^2} \leq e^{K_0^2 \lambda^2}$ for all $\lambda^2 \leq 1/K_0^2$, then $\mathbb{E} e^{\lambda Y} \leq e^{K_0^2 \lambda^2}$ for all $\lambda \in \mathbb{R}$.
2. If $\mathbb{E} e^{\lambda Y} \leq e^{K_0^2 \lambda^2}$ for all $\lambda \in \mathbb{R}$, then $\mathbb{E} e^{\lambda^2 Y^2} \leq e^{16K_0^2 \lambda^2}$ for all $\lambda^2 \leq 1/16K_0^2$.

For sums of independent sub-gaussian random variables, the Proposition 2.6.1 in [Vershynin, 2018] states that

Lemma A.7 Let $X_1, \dots, X_M$ be independent copies of a sub-gaussian random variable X and $\mathbb{E}X = 0$. Then

$$\left\| \sum_{m=1}^M X_m \right\|_{\psi_2} \leq 4\sqrt{M} \|X\|_{\psi_2}.$$

We also state a concentration inequality for sums of independent sub-exponential random variables.

Lemma A.8 (Bernstein's inequality (e.g., Theorem 2.8.1 in [Vershynin, 2018])) Let $X_1, \dots, X_M$ be independent copies of a sub-exponential random variable X and $\mathbb{E}X = 0$. Then, for every $t > 0$, we have

$$P\left(\frac{1}{M} \sum_{m=1}^M X_m > t\right) \leq \exp\left(-\min\left\{\frac{Mt^2}{16\|X\|_{\psi_1}^2}, \frac{Mt}{4\|X\|_{\psi_1}}\right\}\right).$$

B DEFERRED PROOFS IN SECTION 3

B.1 PROOF OF THEOREM 3.1

Proof. For the simplicity of calculation, we denote

$$s := \|x\|_2, \quad s' := \|x'\|_2, \quad \rho = \frac{\langle x, x' \rangle}{\|x\|_2 \|x'\|_2}.$$

The statistical properties of Gaussian distribution indicate that

$$\mathbb{E}_{w \sim \mathcal{N}(0, I_d)} [B(w^\top x) B(w^\top x')] = \mathbb{E}_{x \sim \mathcal{N}(0, 1), z \sim \mathcal{N}(0, 1-\rho^2)} [B(sx) B(s'(\rho x + z))],$$

where x, z are two independent Gaussian random variables. Then we do the calculation based on the latter expression. The calculation is quite complicated, so we illustrate it here in a detailed way.

$$\begin{aligned} & \mathbb{E}_{x \sim \mathcal{N}(0, 1), z \sim \mathcal{N}(0, 1-\rho^2)} [B(sx) B(s'(\rho x + z))] \\ &= \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{2\pi(1-\rho^2)}} \iint_{\mathbb{R}^2} \exp\left(-\frac{(sx-c)^2}{2h^2}\right) \exp\left(-\frac{(s'(\rho x+z)-c)^2}{2h^2}\right) \\ & \quad \exp\left(-\frac{x^2}{2}\right) \exp\left(-\frac{z^2}{2(1-\rho^2)}\right) dx dz \\ &= \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{2\pi(1-\rho^2)}} \iint_{\mathbb{R}^2} \exp\left(-\frac{1}{2h^2(1-\rho^2)} \left[(1-\rho^2)(sx-c)^2 + (1-\rho^2)(s'\rho x-c+s'z)^2 \right. \right. \\ & \quad \left. \left. + (1-\rho^2)h^2x^2 + h^2z^2 \right] \right) dx dz \\ &= \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{2\pi(1-\rho^2)}} \iint_{\mathbb{R}^2} \exp\left(-\frac{1}{2h^2(1-\rho^2)} \left[(1-\rho^2)(sx-c)^2 + (1-\rho^2)(s'\rho x-c)^2 + (1-\rho^2)(s'z)^2 \right. \right. \\ & \quad \left. \left. + 2(1-\rho^2)(s'\rho x-c)s'z + (1-\rho^2)h^2x^2 + h^2z^2 \right] \right) dx dz \\ &= \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{2\pi(1-\rho^2)}} \iint_{\mathbb{R}^2} \exp\left(-\frac{1}{2h^2(1-\rho^2)} \left[(1-\rho^2)(sx-c)^2 + (1-\rho^2)(s'\rho x-c)^2 + (1-\rho^2)h^2x^2 \right. \right. \\ & \quad \left. \left. + [(1-\rho^2)(s')^2 + h^2]z^2 + 2(1-\rho^2)(s'\rho x-c)s'z \right] \right) dx dz \\ &= \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{2\pi(1-\rho^2)}} \iint_{\mathbb{R}^2} \exp\left(-\frac{1}{2h^2(1-\rho^2)} \left[(1-\rho^2)(sx-c)^2 + (1-\rho^2)(s'\rho x-c)^2 + (1-\rho^2)h^2x^2 \right. \right. \\ & \quad \left. \left. + [(1-\rho^2)(s')^2 + h^2] \left[z + \frac{(1-\rho^2)(s'\rho x-c)s'}{(1-\rho^2)(s')^2 + h^2} \right]^2 - \frac{(1-\rho^2)^2(s'\rho x-c)^2(s')^2}{(1-\rho^2)(s')^2 + h^2} \right] \right) dx dz \\ &= \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{2\pi(1-\rho^2)}} \\ & \quad \int_{\mathbb{R}} \exp\left(-\frac{(1-\rho^2)(sx-c)^2 + (1-\rho^2)(s'\rho x-c)^2 + (1-\rho^2)h^2x^2 - \frac{(1-\rho^2)^2(s'\rho x-c)^2(s')^2}{(1-\rho^2)(s')^2 + h^2}}{2h^2(1-\rho^2)}\right) dx \\ & \quad \int_{\mathbb{R}} \exp\left(-\frac{\left[z + \frac{(1-\rho^2)(s'\rho x-c)s'}{(1-\rho^2)(s')^2 + h^2}\right]^2}{2\frac{h^2(1-\rho^2)}{(1-\rho^2)(s')^2 + h^2}}\right) dz \\ &= \frac{1}{\sqrt{2\pi}} \frac{\sqrt{\frac{h^2(1-\rho^2)}{(1-\rho^2)(s')^2 + h^2}}}{\sqrt{(1-\rho^2)}} \int_{\mathbb{R}} \exp\left(-\frac{(1-\rho^2)(sx-c)^2 + (1-\rho^2)(s'\rho x-c)^2 + (1-\rho^2)h^2x^2 - \frac{(1-\rho^2)^2(s'\rho x-c)^2(s')^2}{(1-\rho^2)(s')^2 + h^2}}{2h^2(1-\rho^2)}\right) dx. \end{aligned}$$

To continue, we have

$$\begin{aligned}
& \mathbb{E}_{x \sim \mathcal{N}(0,1), z \sim \mathcal{N}(0,1-\rho^2)} [B(sx)B(s'(\rho x + z))] \\
&= \frac{1}{\sqrt{2\pi}} \frac{h}{\sqrt{(1-\rho^2)(s')^2 + h^2}} \\
& \int_{\mathbb{R}} \exp \left(- \frac{(1-\rho^2)(sx-c)^2 + (1-\rho^2)(s'\rho x - c)^2 + (1-\rho^2)h^2x^2 - \frac{(1-\rho^2)^2(s'\rho x - c)^2(s')^2}{(1-\rho^2)(s')^2 + h^2}}{2h^2(1-\rho^2)} \right) dx \\
&= \frac{1}{\sqrt{2\pi}} \frac{h}{\sqrt{(1-\rho^2)(s')^2 + h^2}} \\
& \int_{\mathbb{R}} \exp \left(- \frac{(sx-c)^2 + (s'\rho x - c)^2 + h^2x^2 - \frac{(1-\rho^2)(s')^2(s'\rho x - c)^2}{(1-\rho^2)(s')^2 + h^2}}{2h^2} \right) dx \\
&= \frac{1}{\sqrt{2\pi}} \frac{h}{\sqrt{(1-\rho^2)(s')^2 + h^2}} \\
& \int_{\mathbb{R}} \exp \left(- \frac{[(1-\rho^2)(s')^2 + h^2][(sx-c)^2 + h^2x^2] + h^2(s'\rho x - c)^2}{2h^2[(1-\rho^2)(s')^2 + h^2]} \right) dx \\
&= \frac{1}{\sqrt{2\pi}} \frac{h}{\sqrt{(1-\rho^2)(s')^2 + h^2}} \\
& \int_{\mathbb{R}} \exp \left(- \frac{[(1-\rho^2)(s')^2 + h^2][(s^2 + h^2)x^2 - 2scx + c^2] + h^2[(s')^2\rho^2x^2 - 2s'\rho cx + c^2]}{2h^2[(1-\rho^2)(s')^2 + h^2]} \right) dx \\
&= \frac{1}{\sqrt{2\pi}} \frac{h}{\sqrt{(1-\rho^2)(s')^2 + h^2}} \\
& \int_{\mathbb{R}} \exp \left(- \frac{[h^4 + (1-\rho^2)s^2(s')^2 + h^2(s^2 + (s')^2)]x^2 - 2[(1-\rho^2)(s')^2s + h^2s + h^2s'\rho]cx + [(1-\rho^2)(s')^2 + 2h^2]c^2}{2h^2[(1-\rho^2)(s')^2 + h^2]} \right) dx \\
&= \frac{1}{\sqrt{2\pi}} \frac{h}{\sqrt{(1-\rho^2)(s')^2 + h^2}} \\
& \int_{\mathbb{R}} \exp \left(- \frac{[h^4 + (1-\rho^2)s^2(s')^2 + h^2(s^2 + (s')^2)] \left(x - \frac{[(1-\rho^2)(s')^2s + h^2s + h^2s'\rho]c}{h^4 + (1-\rho^2)s^2(s')^2 + h^2(s^2 + (s')^2)} \right)^2 - \frac{[(1-\rho^2)(s')^2s + h^2s + h^2s'\rho]^2c^2}{h^4 + (1-\rho^2)s^2(s')^2 + h^2(s^2 + (s')^2)} + [(1-\rho^2)(s')^2 + 2h^2]c^2}{2h^2[(1-\rho^2)(s')^2 + h^2]} \right) dx \\
&= \frac{h\sqrt{(1-\rho^2)(s')^2 + h^2}}{\sqrt{h^4 + (1-\rho^2)s^2(s')^2 + h^2(s^2 + (s')^2)}} \frac{h}{\sqrt{(1-\rho^2)(s')^2 + h^2}} \\
& \exp \left(\frac{[\frac{(1-\rho^2)(s')^2s + h^2s + h^2s'\rho]^2c^2}{h^4 + (1-\rho^2)s^2(s')^2 + h^2(s^2 + (s')^2)} - [(1-\rho^2)(s')^2 + 2h^2]c^2}{2h^2[(1-\rho^2)(s')^2 + h^2]} \right) \\
&= \frac{h^2}{\sqrt{(h^2 + s^2)(h^2 + (s')^2) - \rho^2s^2(s')^2}} \exp \left(\frac{[\frac{(1-\rho^2)(s')^2s + h^2s + h^2s'\rho]^2c^2}{h^4 + (1-\rho^2)s^2(s')^2 + h^2(s^2 + (s')^2)} - [(1-\rho^2)(s')^2 + 2h^2]c^2}{2h^2[(1-\rho^2)(s')^2 + h^2]} \right).
\end{aligned}$$

For the exponential term, we calculate as follows.

$$\begin{aligned}
& \frac{\frac{[(1-\rho^2)(s')^2 s + h^2 s + h^2 s' \rho]^2 c^2}{h^4 + (1-\rho^2)s^2(s')^2 + h^2(s^2 + (s')^2)} - [(1-\rho^2)(s')^2 + h^2 + h^2]c^2}{2h^2[(1-\rho^2)(s')^2 + h^2]} \\
&= -\frac{c^2}{2h^2} \left(1 - \frac{[(1-\rho^2)(s')^2 s + h^2 s + h^2 s' \rho]^2 - h^2[h^4 + (1-\rho^2)s^2(s')^2 + h^2(s^2 + (s')^2)]}{[(1-\rho^2)(s')^2 + h^2][h^4 + (1-\rho^2)s^2(s')^2 + h^2(s^2 + (s')^2)]} \right) \\
&= -\frac{c^2}{2h^2} \left(1 - \frac{[(1-\rho^2)(s')^2 + h^2]s + h^2 s' \rho - h^2[(1-\rho^2)(s')^2 + h^2]s^2 + h^2(h^2 + (s')^2)}{[(1-\rho^2)(s')^2 + h^2][(h^2 + s^2)(h^2 + (s')^2) - \rho^2 s^2(s')^2]} \right) \\
&= -\frac{c^2}{2h^2} \left(1 - \frac{[(1-\rho^2)(s')^2 + h^2]^2 s^2 + h^4(s')^2 \rho^2 + 2((1-\rho^2)(s')^2 + h^2)h^2 s s' \rho}{[(1-\rho^2)(s')^2 + h^2][(h^2 + s^2)(h^2 + (s')^2) - \rho^2 s^2(s')^2]} \right. \\
&\quad \left. - \frac{-h^2[(1-\rho^2)(s')^2 + h^2]s^2 - h^4(h^2 + (s')^2)}{[(1-\rho^2)(s')^2 + h^2][(h^2 + s^2)(h^2 + (s')^2) - \rho^2 s^2(s')^2]} \right) \\
&= -\frac{c^2}{2h^2} \left(1 - \frac{[(1-\rho^2)(s')^2 + h^2]\{[(1-\rho^2)(s')^2 + h^2]s^2 + 2h^2 s s' \rho - h^2 s^2\} - h^4[(1-\rho^2)(s')^2 + h^2]}{[(1-\rho^2)(s')^2 + h^2][(h^2 + s^2)(h^2 + (s')^2) - \rho^2 s^2(s')^2]} \right) \\
&= -\frac{c^2}{2h^2} \left(1 - \frac{[(1-\rho^2)(s')^2 + h^2]\{s^2(s')^2 - \rho^2 s^2(s')^2 + 2h^2 \rho s s' - h^4\}}{[(1-\rho^2)(s')^2 + h^2][(h^2 + s^2)(h^2 + (s')^2) - \rho^2 s^2(s')^2]} \right) \\
&= -\frac{c^2}{2h^2} \left(1 - \frac{s^2(s')^2 - (h^2 - \rho s s')^2}{(h^2 + s^2)(h^2 + (s')^2) - \rho^2 s^2(s')^2} \right).
\end{aligned}$$

Combining the former results, we obtain

$$\begin{aligned}
& \mathbb{E}_{x \sim \mathcal{N}(0,1), z \sim \mathcal{N}(0,1-\rho^2)} [B(sx)B(s'(\rho x + z))] \\
&= \frac{h^2}{\sqrt{(h^2 + s^2)(h^2 + (s')^2) - \rho^2 s^2(s')^2}} \exp \left(-\frac{c^2}{2h^2} \left(1 - \frac{s^2(s')^2 - (h^2 - \rho s s')^2}{(h^2 + s^2)(h^2 + (s')^2) - \rho^2 s^2(s')^2} \right) \right) \\
&= \frac{h^2}{\sqrt{(h^2 + s^2)(h^2 + (s')^2) - \rho^2 s^2(s')^2}} \exp \left(-\frac{c^2}{2} \cdot \frac{(h^2 + s^2) + (h^2 + (s')^2) - 2\rho s(s')}{(h^2 + s^2)(h^2 + (s')^2) - \rho^2 s^2(s')^2} \right).
\end{aligned}$$

Then using the relations

$$s = \|x\|_2, \quad s' = \|x'\|_2, \quad \rho s s' = \langle x, x' \rangle,$$

we obtain

$$\begin{aligned}
& \mathbb{E}_{x \sim \mathcal{N}(0,1), z \sim \mathcal{N}(0,1-\rho^2)} [B(sx)B(s'(\rho x + z))] \\
&= \frac{h^2}{\sqrt{(h^2 + \|x\|^2)(h^2 + \|x'\|^2) - \langle x, x' \rangle^2}} \exp \left(-\frac{c^2}{2} \cdot \frac{(h^2 + \|x\|^2) + (h^2 + \|x'\|^2) - 2\langle x, x' \rangle}{(h^2 + \|x\|^2)(h^2 + \|x'\|^2) - \langle x, x' \rangle^2} \right).
\end{aligned}$$

□

B.2 PROOF OF THEOREM 3.2

Proof.

Step 1. Transform $K(r)$.

To obtain a uniform expression regardless of h , we transform $K(r)$ in the following manner.

$$\begin{aligned}
K(r) &= \frac{h^2}{\sqrt{(1+h^2)^2 - r^2}} \exp \left(-\frac{c^2}{1+h^2+r} \right) \\
&= \frac{h^2}{1+h^2} \frac{1}{\sqrt{1 - \left(\frac{r}{1+h^2}\right)^2}} \exp \left(-\frac{\frac{c^2}{1+h^2}}{1 + \frac{r}{1+h^2}} \right).
\end{aligned}$$

Let $p = \frac{c^2}{1+h^2} \in [0, +\infty)$, $u = \frac{r}{1+h^2} \in [-\frac{1}{1+h^2}, \frac{1}{1+h^2}] \subsetneq (-1, 1)$, and

$$f(u) := \frac{1}{\sqrt{1-u^2}} \exp\left(-\frac{p}{1+u}\right).$$

Then

$$K(r) = \frac{h^2}{1+h^2} f(u).$$

Hence we only need to consider the Taylor expansion of $f(u)$ where $u \in (-1, 1)$.

$$f(u) = \sum_{n=0}^{\infty} \frac{f^{(n)}(0)}{n!} u^n.$$

Step 2. Deriving the recurrence relation of $f^{(n)}(0)$. Solving the Taylor coefficients of $f(u)$ at $u = 0$ is highly technical. For starters, we derive the recurrence formula. For notational convenience, let $y = f(u)$.

From the definition of y , we have the equality

$$y\sqrt{1-u^2} = \exp\left(-\frac{p}{1+u}\right).$$

Taking derivatives on both sides, we have

$$\begin{aligned} y'\sqrt{1-u^2} - \frac{uy}{\sqrt{1-u^2}} &= \frac{p}{(1+u)^2} \exp\left(-\frac{p}{1+u}\right) \\ &= \frac{p\sqrt{1-u^2}}{(1+u)^2} y. \end{aligned}$$

Multiplying $\frac{(u-1)^2(u+1)^2}{\sqrt{1-u^2}}$ on both sides, we have

$$(u^2-1)^2 y' + u(u^2-1)y = p(u-1)^2 y.$$

Eliminating the factor $(u-1)$ and expanding the polynomials lead to

$$(u^3 + u^2 - u - 1)y' + (u^2 + u)y = p(u-1)y.$$

Taking n -th derivatives on both sides and applying the Leibniz rule, we have

$$\begin{aligned} &y^{(n+1)}(u^3 + u^2 - u - 1) + y^{(n)}(u^2 + u) \\ &+ ny^{(n)}(3u^2 + 2u - 1) + ny^{(n-1)}(2u + 1) = y^{(n)}p(u-1) \\ &+ \frac{n(n-1)}{2}y^{(n-1)}(6u + 2) + \frac{n(n-1)}{2}y^{(n-2)} \cdot 2 + ny^{(n-1)}p. \\ &+ \frac{n(n-1)(n-2)}{6}y^{(n-2)} \cdot 6 \end{aligned}$$

Let $u = 0$, and let $y^{(n)} = y^{(n)}(0)$ in the statements hereafter, we have that

$$-y^{(n+1)} - ny^{(n)} + n^2y^{(n-1)} + n(n-1)^2y^{(n-2)} = -py^{(n)} + npy^{(n-1)}.$$

Finally, we have the recurrence formula.

$$y^{(n+1)} = (p-n)y^{(n)} - n(p-n)y^{(n-1)} + n(n-1)^2y^{(n-2)}. \quad (12)$$

To solve $\{y^{(n)}(0)\}_{n=0}^{\infty}$ from the recurrence relation, we also need to obtain $y(0), y'(0), y''(0)$ by hand. A simple calculation

shows that

$$\begin{aligned}
f(u) &= \frac{1}{\sqrt{1-u^2}} \exp\left(-\frac{p}{1+u}\right) \\
f'(u) &= \left(\frac{u}{\sqrt{(1-u^2)^3}} + \frac{1}{\sqrt{1-u^2}} \cdot \frac{p}{(1+u)^2}\right) \exp\left(-\frac{p}{1+u}\right) \\
f''(u) &= \left(\frac{1}{\sqrt{(1-u^2)^3}} + \frac{3u^2}{\sqrt{(1-u^2)^5}}\right. \\
&\quad + \frac{u}{\sqrt{(1-u^2)^3}} \cdot \frac{p}{(1+u)^2} + \frac{1}{\sqrt{1-u^2}} \cdot \frac{-2p}{(1+u)^3} + \\
&\quad \left.\left(\frac{u}{\sqrt{(1-u^2)^3}} + \frac{1}{\sqrt{1-u^2}} \cdot \frac{p}{(1+u)^2}\right) \frac{p}{(1+u)^2}\right) \exp\left(-\frac{p}{1+u}\right).
\end{aligned}$$

Hence, we obtain

$$\begin{aligned}
y^{(0)} &= e^{-p}, \\
y^{(1)} &= pe^{-p}, \\
y^{(2)} &= (p-1)^2 e^{-p}.
\end{aligned}$$

Solving $\{y^{(n)}\}_{n=0}^{\infty}$ remains to be difficult. To simplify the problem, we try to make some observations on the properties of $y^{(n)}$. We supplement $y^{(n)}$ till the first 8 terms.

$$\begin{aligned}
y^{(3)} &= (p^3 - 6p^2 + 9p)e^{-p}, \\
y^{(4)} &= (p^4 - 12p^3 + 42p^2 - 36p + 9)e^{-p}, \\
y^{(5)} &= (p^5 - 20p^4 + 130p^3 - 300p^2 + 225p)e^{-p}, \\
y^{(6)} &= (p^6 - 30p^5 + 315p^4 - 1380p^3 + 2475p^2 - 1350p + 225)e^{-p}, \\
y^{(7)} &= (p^7 - 42p^6 + 651p^5 - 4620p^4 + 15435p^3 - 22050p^2 + 11025p)e^{-p}, \\
y^{(8)} &= (p^8 - 56p^7 + 1204p^6 - 12600p^5 + 67830p^4 - 182280p^3 + 220500p^2 - 88200p + 11025)e^{-p}.
\end{aligned}$$

A further observation shows that

$$\begin{aligned}
y^{(3)} &= p(p-3)^2 e^{-p}, \\
y^{(4)} &= (p^2 - 6p + 3)^2 e^{-p}, \\
y^{(5)} &= p(p^2 - 10p + 15)^2 e^{-p}, \\
y^{(6)} &= (p^3 - 15p^2 + 45p - 15)^2 e^{-p}, \\
y^{(7)} &= p(p^3 - 21p^2 + 105p - 105)^2 e^{-p}, \\
y^{(8)} &= (p^4 - 28p^3 + 210p^2 - 420p + 105)^2 e^{-p}.
\end{aligned}$$

To conclude, we have the following observations.

1. $y^{(n)} = e^{-p} R_n(p)$, where $R_n(p)$ is a polynomial of degree n .
2. For $n = 2k$, $R_n(p) = P_k^2(p)$, where $P_k(p)$ is a polynomial of degree k .
3. For $n = 2k + 1$, $R_n(p) = p \cdot Q_k^2(p)$, where $Q_k(p)$ is a polynomial of degree k .

The correctness of the first observation is easily proved by induction. In the next step, we give a formal proof of the correctness of the second and third observations.

Step 3. Formal proof of the general term formula of the Taylor coefficients. The intuition of the proof is to directly derive the general term formula of $\{P_k\}$ and $\{Q_k\}$ from observations. Note that the observations are non-trivial.

We claim that

$$P_k(x) = \sum_{i=0}^k (-1)^{k-i} \frac{(2k-1)!!}{(2i-1)!!} \cdot \binom{k}{i} x^i, \quad (13)$$

and

$$Q_k(x) = \sum_{i=0}^k (-1)^{k-i} \frac{(2k+1)!!}{(2i+1)!!} \cdot \binom{k}{i} x^i, \quad (14)$$

where $(-1)!! := 1$, and $\binom{0}{0} := 1$ in the above expressions, and

$$R_{2k}(x) = P_k^2(x), \quad (15)$$

$$R_{2k+1}(x) = xQ_k^2(x). \quad (16)$$

We aim to prove the above four equalities true for all $k \in \mathbb{N}$ by induction.

First of all, it is easy to verify that the first three terms conform with the above expressions, where

$$\begin{aligned} P_0(x) &= 1, \\ Q_0(x) &= 1, \\ P_1(x) &= x - 1, \end{aligned}$$

and

$$\begin{aligned} y^{(0)} &= P_0^2(p)e^{-p}, \\ y^{(1)} &= pQ_0^2(p)e^{-p}, \\ y^{(2)} &= P_1^2(p)e^{-p}. \end{aligned}$$

For $n = 2k + 1$, where $k \geq 1$, suppose that Eq. (13) and Eq. (15) hold for all $i \leq k$ and Eq. (14) and Eq. (16) hold for all $i \leq k - 1$. We need to prove that Eq. (14) and Eq. (16) also hold for $i = k$. By Eq. (12), we only need to prove

$$xQ_k^2 = (x - 2k)P_k^2 - 2k(x - 2k)xQ_{k-1}^2 + 2k(2k - 1)^2P_{k-1}^2. \quad (17)$$

For $n = 2k$, where $k \geq 2$, suppose that Eq. (13) and Eq. (15) hold for all $i \leq k - 1$ and Eq. (14) and Eq. (16) hold for all $i \leq k - 1$. We need to prove that Eq. (13) and Eq. (15) also hold for $i = k$. By Eq. (12), we only need to prove

$$P_k^2 = (x - (2k - 1))xQ_{k-1}^2 - (2k - 1)(x - (2k - 1))P_{k-1}^2 + (2k - 1)(2k - 2)^2xQ_{k-2}^2. \quad (18)$$

For notational simplicity, we set for $i \in [k]$,

$$a_i^k = (-1)^{k-i} \frac{(2k-1)!!}{(2i-1)!!} \cdot \binom{k}{i}, \quad b_i^k = (-1)^{k-i} \frac{(2k+1)!!}{(2i+1)!!} \cdot \binom{k}{i}.$$

For $i \in [2k]$,

$$A_i^k = \sum_{j=0 \vee i-k}^{i \wedge k} a_j^k a_{i-j}^k, \quad B_i^k = \sum_{j=0 \vee i-k}^{i \wedge k} b_j^k b_{i-j}^k.$$

The polynomials are written as

$$\begin{aligned} P_k(x) &= \sum_{i=0}^k a_i^k x^i, \quad Q_k(x) = \sum_{i=0}^k b_i^k x^i. \\ (P_k(x))^2 &= \sum_{i=0}^{2k} A_i^k x^i, \quad (Q_k(x))^2 = \sum_{i=0}^{2k} B_i^k x^i. \end{aligned}$$

Proof of Eq. (17). Now consider the right-hand side of Eq. (17).

$$\begin{aligned}
\text{RHS} &= (x-2k) \sum_{i=0}^{2k} A_i^k x^i - 2kx(x-2k) \sum_{i=0}^{2k-2} B_i^{k-1} x^i + 2k(2k-1)^2 \sum_{i=0}^{2k-2} A_i^{k-1} x^i \\
&= \sum_{i=0}^{2k} A_i^k x^{i+1} + \sum_{i=0}^{2k} (-2k) A_i^k x^i \\
&\quad + \sum_{i=0}^{2k-2} (-2k) B_i^{k-1} x^{i+2} + \sum_{i=0}^{2k-2} (2k)^2 B_i^{k-1} x^{i+1} \\
&\quad + \sum_{i=0}^{2k-2} 2k(2k-1)^2 A_i^{k-1} x^i \\
&= \sum_{i=1}^{2k+1} A_{i-1}^k x^i + \sum_{i=0}^{2k} (-2k) A_i^k x^i \\
&\quad + \sum_{i=2}^{2k} (-2k) B_{i-2}^{k-1} x^i + \sum_{i=1}^{2k-1} (2k)^2 B_{i-1}^{k-1} x^i \\
&\quad + \sum_{i=0}^{2k-2} 2k(2k-1)^2 A_i^{k-1} x^i \\
&= \sum_{i=2}^{2k-2} (A_{i-1}^k - 2kA_i^k - 2kB_{i-2}^{k-1} + (2k)^2 B_{i-1}^{k-1} + 2k(2k-1)^2 A_i^{k-1}) x^i \\
&\quad + (A_{2k-2}^k - 2kA_{2k-1}^k - 2kB_{2k-3}^{k-1} + (2k)^2 B_{2k-2}^{k-1}) x^{2k-1} \\
&\quad + (A_{2k-1}^k - 2kA_{2k}^k - 2kB_{2k-2}^{k-1}) x^{2k} + A_{2k}^k x^{2k+1} \\
&\quad + (A_0^k - 2kA_1^k + (2k)^2 B_0^{k-1} + 2k(2k-1)^2 A_1^{k-1}) x \\
&\quad + (2k(2k-1)^2 A_0^{k-1} - 2kA_0^k).
\end{aligned}$$

For the constant term, the general term formula is

$$A_0^k = (a_0^k)^2 = ((2k-1)!!)^2.$$

Hence,

$$2k(2k-1)^2 A_0^{k-1} - 2kA_0^k = 2k[(2k-1)^2 \cdot ((2k-3)!!)^2 - ((2k-1)!!)^2] = 0.$$

Plug the result into the right-hand side, we obtain

$$\begin{aligned}
\text{RHS} &= x \left\{ \sum_{i=2}^{2k-2} (A_{i-1}^k - 2kA_i^k - 2kB_{i-2}^{k-1} + (2k)^2 B_{i-1}^{k-1} + 2k(2k-1)^2 A_i^{k-1}) x^{i-1} \right. \\
&\quad + (A_{2k-2}^k - 2kA_{2k-1}^k - 2kB_{2k-3}^{k-1} + (2k)^2 B_{2k-2}^{k-1}) x^{2k-2} \\
&\quad + (A_{2k-1}^k - 2kA_{2k}^k - 2kB_{2k-2}^{k-1}) x^{2k-1} + A_{2k}^k x^{2k} \\
&\quad \left. + (A_0^k - 2kA_1^k + (2k)^2 B_0^{k-1} + 2k(2k-1)^2 A_1^{k-1}) \right\} \\
&= x \left\{ \sum_{i=1}^{2k-3} (A_i^k - 2kA_{i+1}^k - 2kB_{i-1}^{k-1} + (2k)^2 B_i^{k-1} + 2k(2k-1)^2 A_{i+1}^{k-1}) x^i \right. \\
&\quad + (A_{2k-2}^k - 2kA_{2k-1}^k - 2kB_{2k-3}^{k-1} + (2k)^2 B_{2k-2}^{k-1}) x^{2k-2} \\
&\quad + (A_{2k-1}^k - 2kA_{2k}^k - 2kB_{2k-2}^{k-1}) x^{2k-1} + A_{2k}^k x^{2k} \\
&\quad \left. + (A_0^k - 2kA_1^k + (2k)^2 B_0^{k-1} + 2k(2k-1)^2 A_1^{k-1}) \right\}.
\end{aligned}$$

We then verify the coefficients are equal to those of $xQ_k^2(x) = x \left(\sum_{i=0}^{2k} B_i^k x^i \right)$.

For $i = 2k$,

$$A_{2k}^k = (a_k^k)^2 = 1^2 = (b_k^k)^2 = B_{2k}^k.$$

For $i = 2k - 1$,

$$\begin{aligned} & A_{2k-1}^k - 2kA_{2k}^k - 2kB_{2k-2}^{k-1} \\ &= 2a_{k-1}^k a_k^k - 2k(a_k^k)^2 - 2k(b_{k-1}^{k-1})^2 \\ &= -2(2k-1)k - 2k - 2k \\ &= -2k(2k+1) = B_{2k-1}^k. \end{aligned}$$

For $i = 2k - 2$,

$$\begin{aligned} & A_{2k-2}^k - 2kA_{2k-1}^k - 2kB_{2k-3}^{k-1} + (2k)^2 B_{2k-2}^{k-1} \\ &= (a_{k-1}^k)^2 + 2a_{k-2}^k a_k^k - 2k \cdot 2a_{k-1}^k a_k^k - 2k \cdot 2b_{k-2}^{k-1} b_{k-1}^{k-1} + (2k)^2 (b_{k-1}^{k-1})^2 \\ &= ((2k-1) \cdot k)^2 + 2 \cdot (2k-1)(2k-3) \frac{k(k-1)}{2} + 2k \cdot 2(2k-1)k \\ &\quad + 2k \cdot 2(2k-1)(k-1) + (2k)^2 \\ &= ((2k+1)k)^2 + 2(2k+1)(2k-1) \frac{k(k-1)}{2} = B_{2k-2}^k. \end{aligned}$$

For $i = 0$,

$$\begin{aligned} & A_0^k - 2kA_1^k + (2k)^2 B_0^{k-1} + 2k(2k-1)^2 A_1^{k-1} \\ &= ((2k-1)!!)^2 - 2k(-2(2k-1)!! \cdot (2k-1)!! \cdot k) \\ &\quad + (2k)^2 ((2k-1)!!)^2 + 2k(2k-1)^2 (-2(2k-3)!! \cdot (2k-3)!! \cdot (k-1)) \\ &= ((2k-1)!!)^2 + 2(2k)^2 ((2k-1)!!)^2 - (2k)(2k-2)((2k-1)!!)^2 \\ &= (2k+1)^2 ((2k-1)!!)^2 = ((2k+1)!!)^2 = B_0^k. \end{aligned}$$

For $1 \leq i \leq 2k - 3$, we need to show that

$$A_i^k - 2kA_{i+1}^k - 2kB_{i-1}^{k-1} + (2k)^2 B_i^{k-1} + 2k(2k-1)^2 A_{i+1}^{k-1} = B_i^k.$$

For starters, we have for the right-hand side that

$$\begin{aligned} (-1)^i B_i^k &= \sum_{j=0 \vee i-k}^{i \wedge k} \frac{(2k+1)!!}{(2j+1)!!} \binom{k}{j} \frac{(2k+1)!!}{(2(i-j)+1)!!} \binom{k}{i-j} \\ &= (2k+1)^2 \sum_{j=0 \vee i-k}^{i \wedge k} \frac{(2k-1)!!}{(2j+1)!!} \binom{k}{j} \frac{(2k-1)!!}{(2(i-j)+1)!!} \binom{k}{i-j}. \end{aligned}$$

For the left-hand side, we have

$$\begin{aligned}
& (-1)^i (A_i^k - 2kA_{i+1}^k - 2kB_{i-1}^{k-1} + (2k)^2 B_i^{k-1} + 2k(2k-1)^2 A_{i+1}^{k-1}) \\
&= \sum_{j=0 \vee i-k}^{i \wedge k} \frac{(2k-1)!!}{(2j-1)!!} \binom{k}{j} \frac{(2k-1)!!}{(2(i-j)-1)!!} \binom{k}{i-j} \\
&\quad + 2k \sum_{j=0 \vee i+1-k}^{i+1 \wedge k} \frac{(2k-1)!!}{(2j-1)!!} \binom{k}{j} \frac{(2k-1)!!}{(2(i+1-j)-1)!!} \binom{k}{i+1-j} \\
&\quad + 2k \sum_{j=0 \vee i-k}^{i-1 \wedge k-1} \frac{(2k-1)!!}{(2j+1)!!} \binom{k-1}{j} \frac{(2k-1)!!}{(2(i-1-j)+1)!!} \binom{k-1}{i-1-j} \\
&\quad + (2k)^2 \sum_{j=0 \vee i-k+1}^{i \wedge k-1} \frac{(2k-1)!!}{(2j+1)!!} \binom{k-1}{j} \frac{(2k-1)!!}{(2(i-j)+1)!!} \binom{k-1}{i-j} \\
&\quad - 2k(2k-1)^2 \sum_{j=0 \vee i-k+2}^{i+1 \wedge k-1} \frac{(2k-3)!!}{(2j-1)!!} \binom{k-1}{j} \frac{(2k-3)!!}{(2(i+1-j)-1)!!} \binom{k-1}{i+1-j} \\
&= \sum_{j=0 \vee i-k}^{i \wedge k} \frac{(2k-1)!!}{(2j-1)!!} \binom{k}{j} \frac{(2k-1)!!}{(2(i-j)-1)!!} \binom{k}{i-j} \\
&\quad + 2k \sum_{j=0 \vee i+1-k}^{i+1 \wedge k} \frac{(2k-1)!!}{(2j-1)!!} \binom{k}{j} \frac{(2k-1)!!}{(2(i-j)+1)!!} \binom{k}{i+1-j} \\
&\quad + 2k \sum_{j=0 \vee i-k}^{i-1 \wedge k-1} \frac{(2k-1)!!}{(2j+1)!!} \binom{k-1}{j} \frac{(2k-1)!!}{(2(i-j)-1)!!} \binom{k-1}{i-1-j} \\
&\quad + (2k)^2 \sum_{j=0 \vee i-k+1}^{i \wedge k-1} \frac{(2k-1)!!}{(2j+1)!!} \binom{k-1}{j} \frac{(2k-1)!!}{(2(i-j)+1)!!} \binom{k-1}{i-j} \\
&\quad - 2k \sum_{j=0 \vee i-k+2}^{i+1 \wedge k-1} \frac{(2k-1)!!}{(2j-1)!!} \binom{k-1}{j} \frac{(2k-1)!!}{(2(i-j)+1)!!} \binom{k-1}{i+1-j} \\
&= \sum_{j=0 \vee i-k}^{i \wedge k} [(2j+1)(2(i-j)+1)] \frac{(2k-1)!!}{(2j+1)!!} \binom{k}{j} \frac{(2k-1)!!}{(2(i-j)+1)!!} \binom{k}{i-j} \\
&\quad + 2k \sum_{j=0 \vee i+1-k}^{i+1 \wedge k} \frac{(2k-1)!!}{(2j-1)!!} \binom{k}{j} \frac{(2k-1)!!}{(2(i-j)+1)!!} \binom{k}{i+1-j} \\
&\quad + 2k \sum_{j=0 \vee i-k}^{i-1 \wedge k-1} \frac{(2k-1)!!}{(2j+1)!!} \binom{k-1}{j} \frac{(2k-1)!!}{(2(i-j)-1)!!} \binom{k-1}{i-1-j} \\
&\quad + \sum_{j=0 \vee i-k+1}^{i \wedge k-1} [2(k-j)2(k-(i-j))] \frac{(2k-1)!!}{(2j+1)!!} \binom{k}{j} \frac{(2k-1)!!}{(2(i-j)+1)!!} \binom{k}{i-j} \\
&\quad - 2k \sum_{j=0 \vee i-k+2}^{i+1 \wedge k-1} \frac{(2k-1)!!}{(2j-1)!!} \binom{k-1}{j} \frac{(2k-1)!!}{(2(i-j)+1)!!} \binom{k-1}{i+1-j}.
\end{aligned}$$

For the second, third and fifth terms, we have

$$\begin{aligned}
& 2k \sum_{j=0 \vee i+1-k}^{i+1 \wedge k} \frac{(2k-1)!!}{(2j-1)!!} \binom{k}{j} \frac{(2k-1)!!}{(2(i-j)+1)!!} \binom{k}{i+1-j} \\
& + 2k \sum_{j=0 \vee i-k}^{i-1 \wedge k-1} \frac{(2k-1)!!}{(2j+1)!!} \binom{k-1}{j} \frac{(2k-1)!!}{(2(i-j)-1)!!} \binom{k-1}{i-1-j} \\
& - 2k \sum_{j=0 \vee i-k+2}^{i+1 \wedge k-1} \frac{(2k-1)!!}{(2j-1)!!} \binom{k-1}{j} \frac{(2k-1)!!}{(2(i-j)+1)!!} \binom{k-1}{i+1-j} \\
& = 2k \sum_{j=0 \vee i+1-k}^{i+1 \wedge k} \frac{(2k-1)!!}{(2j-1)!!} \binom{k}{j} \frac{(2k-1)!!}{(2(i-j)+1)!!} \binom{k}{i+1-j} \\
& + 2k \sum_{j=0 \vee i+1-k}^{i \wedge k-1} \frac{(2k-1)!!}{(2j+1)!!} \binom{k-1}{j} \frac{(2k-1)!!}{(2(i-j)-1)!!} \binom{k}{i-j} \\
& - 2k \sum_{j=0 \vee i+1-k}^{i \wedge k-1} \frac{(2k-1)!!}{(2j+1)!!} \binom{k-1}{j} \frac{(2k-1)!!}{(2(i-j)-1)!!} \binom{k-1}{i-j} \\
& - 2k \sum_{j=0 \vee i-k+2}^{i+1 \wedge k-1} \frac{(2k-1)!!}{(2j-1)!!} \binom{k-1}{j} \frac{(2k-1)!!}{(2(i-j)+1)!!} \binom{k-1}{i+1-j} \\
& = 2k \sum_{j=0 \vee i+1-k}^{i+1 \wedge k} \frac{(2k-1)!!}{(2j-1)!!} \frac{k!}{j!(k-j)!} \frac{(2k-1)!!}{(2(i-j)+1)!!} \frac{k!}{(i-j+1)!(k-i+j-1)!} \\
& + 2k \sum_{j=0 \vee i+1-k}^{i \wedge k-1} \frac{(2k-1)!!}{(2j+1)!!} \frac{(k-1)!}{j!(k-j-1)!} \frac{(2k-1)!!}{(2(i-j)-1)!!} \frac{k!}{(i-j)!(k-i+j)!} \\
& - 2k \sum_{j=0 \vee i+1-k}^{i \wedge k-1} \frac{(2k-1)!!}{(2j+1)!!} \frac{(k-1)!}{j!(k-j-1)!} \frac{(2k-1)!!}{(2(i-j)-1)!!} \frac{(k-1)!}{(i-j)!(k-i+j-1)!} \\
& - 2k \sum_{j=0 \vee i-k+2}^{i+1 \wedge k-1} \frac{(2k-1)!!}{(2j-1)!!} \frac{(k-1)!}{j!(k-j-1)!} \frac{(2k-1)!!}{(2(i-j)+1)!!} \frac{(k-1)!}{(i-j+1)!(k-i+j-2)!} \\
& = 2k \sum_{j=0 \vee i+1-k}^{i+1 \wedge k} \frac{(2k-1)!!}{(2j-1)!!} \frac{k!}{j!(k-j)!} \frac{(2k-1)!!}{(2(i-j)+1)!!} \frac{k!}{(i-j+1)!(k-i+j-1)!} \\
& + \sum_{j=0 \vee i+1-k}^{i \wedge k-1} [2(k-j)(2(i-j)+1)] \frac{(2k-1)!!}{(2j+1)!!} \frac{k!}{j!(k-j)!} \frac{(2k-1)!!}{(2(i-j)+1)!!} \frac{k!}{(i-j)!(k-i+j)!} \\
& - 2k \sum_{j=0 \vee i+1-k}^{i \wedge k-1} \frac{(2k-1)!!}{(2j+1)!!} \frac{(k-1)!}{j!(k-j-1)!} \frac{(2k-1)!!}{(2(i-j)-1)!!} \frac{(k-1)!}{(i-j)!(k-i+j-1)!} \\
& - 2k \sum_{j=0 \vee i-k+2}^{i+1 \wedge k-1} \frac{(2k-1)!!}{(2j-1)!!} \frac{(k-1)!}{j!(k-j-1)!} \frac{(2k-1)!!}{(2(i-j)+1)!!} \frac{(k-1)!}{(i-j+1)!(k-i+j-2)!},
\end{aligned}$$

where in the first equality, we use the relation

$$\binom{k}{i-j} - \binom{k-1}{i-j} = \binom{k-1}{i-j-1}.$$

For the first, third and fourth terms of the former expression, we have

$$\begin{aligned}
& 2k \sum_{j=0 \vee i+1-k}^{i+1 \wedge k} \frac{(2k-1)!!}{(2j-1)!!} \frac{k!}{j!(k-j)!} \frac{(2k-1)!!}{(2(i-j)+1)!!} \frac{k!}{(i-j+1)!(k-i+j-1)!} \\
& - 2k \sum_{j=0 \vee i+1-k}^{i \wedge k-1} \frac{(2k-1)!!}{(2j+1)!!} \frac{(k-1)!}{j!(k-j-1)!} \frac{(2k-1)!!}{(2(i-j)-1)!!} \frac{(k-1)!}{(i-j)!(k-i+j-1)!} \\
& - 2k \sum_{j=0 \vee i-k+2}^{i+1 \wedge k-1} \frac{(2k-1)!!}{(2j-1)!!} \frac{(k-1)!}{j!(k-j-1)!} \frac{(2k-1)!!}{(2(i-j)+1)!!} \frac{(k-1)!}{(i-j+1)!(k-i+j-2)!} \\
& = 2k \sum_{j=0 \vee i+1-k}^{i+1 \wedge k} \frac{(2k-1)!!}{(2j-1)!!} \frac{k!}{j!(k-j)!} \frac{(2k-1)!!}{(2(i-j)+1)!!} \frac{k!}{(i-j+1)!(k-i+j-1)!} \\
& - 2k \sum_{j=0 \vee i+1-k}^{i \wedge k-1} \frac{(2k-1)!!}{(2j+1)!!} \frac{(k-1)!}{j!(k-j-1)!} \frac{(2k-1)!!}{(2(i-j)-1)!!} \frac{(k-1)!}{(i-j)!(k-i+j-1)!} \\
& - 2 \sum_{j=0 \vee i-k+2}^{i+1 \wedge k-1} (k-j) \frac{(2k-1)!!}{(2j-1)!!} \frac{k!}{j!(k-j)!} \frac{(2k-1)!!}{(2(i-j)+1)!!} \frac{(k-1)!}{(i-j+1)!(k-i+j-2)!} \\
& = 2k \sum_{j=0 \vee i+1-k}^{i+1 \wedge k} \frac{(2k-1)!!}{(2j-1)!!} \frac{k!}{j!(k-j)!} \frac{(2k-1)!!}{(2(i-j)+1)!!} \frac{k!}{(i-j+1)!(k-i+j-1)!} \\
& - 2k \sum_{j=0 \vee i+1-k}^{i \wedge k-1} \frac{(2k-1)!!}{(2j+1)!!} \frac{(k-1)!}{j!(k-j-1)!} \frac{(2k-1)!!}{(2(i-j)-1)!!} \frac{(k-1)!}{(i-j)!(k-i+j-1)!} \\
& - 2k \sum_{j=0 \vee i-k+2}^{i+1 \wedge k-1} \frac{(2k-1)!!}{(2j-1)!!} \frac{k!}{j!(k-j)!} \frac{(2k-1)!!}{(2(i-j)+1)!!} \frac{(k-1)!}{(i-j+1)!(k-i+j-2)!} \\
& + \sum_{j=0 \vee i-k+2}^{i+1 \wedge k-1} (2j) \frac{(2k-1)!!}{(2j-1)!!} \frac{k!}{j!(k-j)!} \frac{(2k-1)!!}{(2(i-j)+1)!!} \frac{(k-1)!}{(i-j+1)!(k-i+j-2)!} \\
& = 2k \sum_{j=0 \vee i-k}^{i \wedge k} \frac{(2k-1)!!}{(2j-1)!!} \frac{k!}{j!(k-j)!} \frac{(2k-1)!!}{(2(i-j)+1)!!} \frac{(k-1)!}{(i-j)!(k-i+j-1)!} \\
& - 2k \sum_{j=0 \vee i+1-k}^{i \wedge k-1} \frac{(2k-1)!!}{(2j+1)!!} \frac{(k-1)!}{j!(k-j-1)!} \frac{(2k-1)!!}{(2(i-j)-1)!!} \frac{(k-1)!}{(i-j)!(k-i+j-1)!} \\
& + \sum_{j=0 \vee i-k+2}^{i+1 \wedge k-1} (2j) \frac{(2k-1)!!}{(2j-1)!!} \frac{k!}{j!(k-j)!} \frac{(2k-1)!!}{(2(i-j)+1)!!} \frac{(k-1)!}{(i-j+1)!(k-i+j-2)!} \\
& = \sum_{j=0 \vee i-k}^{i \wedge k} [(2j+1)2(k-i+j)] \frac{(2k-1)!!}{(2j+1)!!} \frac{k!}{j!(k-j)!} \frac{(2k-1)!!}{(2(i-j)+1)!!} \frac{(k-1)!}{(i-j)!(k-i+j-1)!} \\
& - 2k \sum_{j=0 \vee i+1-k}^{i \wedge k-1} \frac{(2k-1)!!}{(2j+1)!!} \frac{(k-1)!}{j!(k-j-1)!} \frac{(2k-1)!!}{(2(i-j)-1)!!} \frac{(k-1)!}{(i-j)!(k-i+j-1)!} \\
& + \sum_{j=0 \vee i-k+2}^{i+1 \wedge k-1} (2j) \frac{(2k-1)!!}{(2j-1)!!} \frac{k!}{j!(k-j)!} \frac{(2k-1)!!}{(2(i-j)+1)!!} \frac{(k-1)!}{(i-j+1)!(k-i+j-2)!},
\end{aligned}$$

where in the third equality, we combine the first and third terms using the relation

$$\binom{k}{i-j+1} - \binom{k-1}{i-j+1} = \binom{k-1}{i-j}.$$

For the last two terms, we have

$$\begin{aligned}
& -2k \sum_{j=0 \vee i+1-k}^{i \wedge k-1} \frac{(2k-1)!!}{(2j+1)!!} \frac{(k-1)!}{j!(k-j-1)!} \frac{(2k-1)!!}{(2(i-j)-1)!!} \frac{(k-1)!}{(i-j)!(k-i+j-1)!} \\
& + \sum_{j=0 \vee i-k+2}^{i+1 \wedge k-1} (2j) \frac{(2k-1)!!}{(2j-1)!!} \frac{k!}{j!(k-j)!} \frac{(2k-1)!!}{(2(i-j)+1)!!} \frac{(k-1)!}{(i-j+1)!(k-i+j-2)!} \\
& = -2k \sum_{j=0 \vee i+1-k}^{i \wedge k-1} \frac{(2k-1)!!}{(2j+1)!!} \frac{(k-1)!}{j!(k-j-1)!} \frac{(2k-1)!!}{(2(i-j)-1)!!} \frac{(k-1)!}{(i-j)!(k-i+j-1)!} \\
& + \sum_{j=0 \vee i-k+1}^{i \wedge k-1} (2j+2) \frac{(2k-1)!!}{(2j+1)!!} \frac{k!}{(j+1)!(k-j-1)!} \frac{(2k-1)!!}{(2(i-j)-1)!!} \frac{(k-1)!}{(i-j)!(k-i+j-1)!} \\
& = -2 \sum_{j=0 \vee i+1-k}^{i \wedge k-1} \frac{(2k-1)!!}{(2j+1)!!} \frac{k!}{j!(k-j-1)!} \frac{(2k-1)!!}{(2(i-j)-1)!!} \frac{(k-1)!}{(i-j)!(k-i+j-1)!} \\
& + 2 \sum_{j=0 \vee i-k+1}^{i \wedge k-1} \frac{(2k-1)!!}{(2j+1)!!} \frac{k!}{j!(k-j-1)!} \frac{(2k-1)!!}{(2(i-j)-1)!!} \frac{(k-1)!}{(i-j)!(k-i+j-1)!} \\
& = 0.
\end{aligned}$$

Combine the four parts illustrated above, we have that

$$\begin{aligned}
& (-1)^i (A_i^k - 2kA_{i+1}^k - 2kB_{i-1}^{k-1} + (2k)^2 B_i^{k-1} + 2k(2k-1)^2 A_{i+1}^{k-1}) \\
& = \sum_{j=0 \vee i-k}^{i \wedge k} [(2j+1)(2(i-j)+1) + 2(k-j)2(k-i+j) \\
& \quad + 2(k-j)(2(i-j)+1) + (2j+1)2(k-i+j) + 0] \\
& \quad \cdot \frac{(2k-1)!!}{(2j+1)!!} \binom{k}{j} \frac{(2k-1)!!}{(2(i-j)+1)!!} \binom{k}{i-j} \\
& = \sum_{j=0 \vee i-k}^{i \wedge k} (2k+1)^2 \frac{(2k-1)!!}{(2j+1)!!} \binom{k}{j} \frac{(2k-1)!!}{(2(i-j)+1)!!} \binom{k}{i-j} \\
& = \sum_{j=0 \vee i-k}^{i \wedge k} \frac{(2k+1)!!}{(2j+1)!!} \binom{k}{j} \frac{(2k+1)!!}{(2(i-j)+1)!!} \binom{k}{i-j} \\
& = (-1)^i B_i^k.
\end{aligned}$$

Finally, we complete the proof of Eq. (17).

Proof of Eq. (18). Consider the right-hand side of Eq. (18).

$$\begin{aligned}
\text{RHS} &= (x - (2k - 1))x \sum_{i=0}^{2k-2} B_i^{k-1} x^i - (2k - 1)(x - (2k - 1)) \sum_{i=0}^{2k-2} A_i^{k-1} x^i + (2k - 1)(2k - 2)^2 x \sum_{i=0}^{2k-4} B_i^{k-2} x^i \\
&= \sum_{i=0}^{2k-2} B_i^{k-1} x^{i+2} + \sum_{i=0}^{2k-2} (-(2k - 1)) B_i^{k-1} x^{i+1} \\
&\quad + \sum_{i=0}^{2k-2} (-(2k - 1)) A_i^{k-1} x^{i+1} + \sum_{i=0}^{2k-2} (2k - 1)^2 A_i^{k-1} x^i \\
&\quad + \sum_{i=0}^{2k-4} (2k - 1)(2k - 2)^2 B_i^{k-2} x^{i+1} \\
&= \sum_{i=2}^{2k} B_{i-2}^{k-1} x^i + \sum_{i=1}^{2k-1} (-(2k - 1)) B_{i-1}^{k-1} x^i \\
&\quad + \sum_{i=1}^{2k-1} (-(2k - 1)) A_{i-1}^{k-1} x^i + \sum_{i=0}^{2k-2} (2k - 1)^2 A_i^{k-1} x^i \\
&\quad + \sum_{i=1}^{2k-3} (2k - 1)(2k - 2)^2 B_{i-1}^{k-2} x^i \\
&= \sum_{i=2}^{2k-3} (B_{i-2}^{k-1} - (2k - 1) B_{i-1}^{k-1} - (2k - 1) A_{i-1}^{k-1} + (2k - 1)^2 A_i^{k-1} + (2k - 1)(2k - 2)^2 B_{i-1}^{k-2}) x^i \\
&\quad + (B_{2k-4}^{k-1} - (2k - 1) B_{2k-3}^{k-1} - (2k - 1) A_{2k-3}^{k-1} + (2k - 1)^2 A_{2k-2}^{k-1}) x^{2k-2} \\
&\quad + (B_{2k-3}^{k-1} - (2k - 1) B_{2k-2}^{k-1} - (2k - 1) A_{2k-2}^{k-1}) x^{2k-1} + B_{2k-2}^{k-1} x^{2k} \\
&\quad + (-(2k - 1) B_0^{k-1} - (2k - 1) A_0^{k-1} + (2k - 1)^2 A_1^{k-1} + (2k - 1)(2k - 2)^2 B_0^{k-2}) x \\
&\quad + (2k - 1)^2 A_0^{k-1}.
\end{aligned}$$

It suffices to prove that the coefficients of the above expression are equal to those of $P_k^2(x) = \sum_{i=0}^{2k} A_i^k x^i$.

For $i = 0, 1, 2k - 2, 2k - 1, 2k$, the verifications are trivial. We only need to show that for $2 \leq i \leq 2k - 3$, it holds that

$$B_{i-2}^{k-1} - (2k - 1) B_{i-1}^{k-1} - (2k - 1) A_{i-1}^{k-1} + (2k - 1)^2 A_i^{k-1} + (2k - 1)(2k - 2)^2 B_{i-1}^{k-2} = A_i^k.$$

Consider the right-hand side, we have

$$(-1)^i A_i^k = \sum_{j=0 \vee i-k}^{i \wedge k} \frac{(2k-1)!!}{(2j-1)!!} \binom{k}{j} \frac{(2k-1)!!}{(2(i-j)-1)!!} \binom{k}{i-j}.$$

For the left-hand side, we first have

$$\begin{aligned}
& (-1)^i (B_{i-2}^{k-1} - (2k-1)B_{i-1}^{k-1} - (2k-1)A_{i-1}^{k-1} + (2k-1)^2 A_i^{k-1} + (2k-1)(2k-2)^2 B_{i-1}^{k-2}) \\
&= \sum_{j=0 \vee i-1-k}^{i-2 \wedge k-1} \frac{(2k-1)!!}{(2j+1)!!} \binom{k-1}{j} \frac{(2k-1)!!}{(2(i-2-j)+1)!!} \binom{k-1}{i-2-j} \\
&\quad + (2k-1) \sum_{j=0 \vee i-k}^{i-1 \wedge k-1} \frac{(2k-1)!!}{(2j+1)!!} \binom{k-1}{j} \frac{(2k-1)!!}{(2(i-1-j)+1)!!} \binom{k-1}{i-1-j} \\
&\quad + (2k-1) \sum_{j=0 \vee i-k}^{i-1 \wedge k-1} \frac{(2k-3)!!}{(2j-1)!!} \binom{k-1}{j} \frac{(2k-3)!!}{(2(i-1-j)-1)!!} \binom{k-1}{i-1-j} \\
&\quad + (2k-1)^2 \sum_{j=0 \vee i-k+1}^{i \wedge k-1} \frac{(2k-3)!!}{(2j-1)!!} \binom{k-1}{j} \frac{(2k-3)!!}{(2(i-j)-1)!!} \binom{k-1}{i-j} \\
&\quad - (2k-1)(2k-2)^2 \sum_{j=0 \vee i-k+1}^{i-1 \wedge k-2} \frac{(2k-3)!!}{(2j+1)!!} \binom{k-2}{j} \frac{(2k-3)!!}{(2(i-1-j)+1)!!} \binom{k-2}{i-1-j} \\
&= \sum_{j=0 \vee i-1-k}^{i-2 \wedge k-1} \frac{(2k-1)!!}{(2j+1)!!} \binom{k-1}{j} \frac{(2k-1)!!}{(2(i-j)-3)!!} \binom{k-1}{i-2-j} \\
&\quad + (2k-1) \sum_{j=0 \vee i-k}^{i-1 \wedge k-1} \frac{(2k-1)!!}{(2j+1)!!} \binom{k-1}{j} \frac{(2k-1)!!}{(2(i-j)-1)!!} \binom{k-1}{i-1-j} \\
&\quad + (2k-1) \sum_{j=0 \vee i-k}^{i-1 \wedge k-1} \frac{(2k-3)!!}{(2j-1)!!} \binom{k-1}{j} \frac{(2k-3)!!}{(2(i-j)-3)!!} \binom{k-1}{i-1-j} \\
&\quad + (2k-1)^2 \sum_{j=0 \vee i-k+1}^{i \wedge k-1} \frac{(2k-3)!!}{(2j-1)!!} \binom{k-1}{j} \frac{(2k-3)!!}{(2(i-j)-1)!!} \binom{k-1}{i-j} \\
&\quad - (2k-1)(2k-2)^2 \sum_{j=0 \vee i-k+1}^{i-1 \wedge k-2} \frac{(2k-3)!!}{(2j+1)!!} \binom{k-2}{j} \frac{(2k-3)!!}{(2(i-j)-1)!!} \binom{k-2}{i-1-j} \\
&= \sum_{j=0 \vee i-k}^{i-1 \wedge k-1} \frac{(2k-1)!!}{(2j+1)!!} \binom{k-1}{j} \frac{(2k-1)!!}{(2(i-j)-3)!!} \binom{k}{i-1-j} \\
&\quad - \sum_{j=0 \vee i-k}^{i-1 \wedge k-1} \frac{(2k-1)!!}{(2j+1)!!} \binom{k-1}{j} \frac{(2k-1)!!}{(2(i-j)-3)!!} \binom{k-1}{i-1-j} \\
&\quad + (2k-1) \sum_{j=0 \vee i-k}^{i-1 \wedge k-1} \frac{(2k-1)!!}{(2j+1)!!} \binom{k-1}{j} \frac{(2k-1)!!}{(2(i-j)-1)!!} \binom{k-1}{i-1-j} \\
&\quad + (2k-1) \sum_{j=0 \vee i-k}^{i-1 \wedge k-1} \frac{(2k-3)!!}{(2j-1)!!} \binom{k-1}{j} \frac{(2k-3)!!}{(2(i-j)-3)!!} \binom{k-1}{i-1-j} \\
&\quad + \sum_{j=0 \vee i-k+1}^{i \wedge k-1} \frac{(2k-1)!!}{(2j-1)!!} \binom{k-1}{j} \frac{(2k-1)!!}{(2(i-j)-1)!!} \binom{k-1}{i-j} \\
&\quad - (2k-1) \sum_{j=0 \vee i-k+1}^{i-1 \wedge k-2} 2(k-1-j)2(k-i+j) \frac{(2k-3)!!}{(2j+1)!!} \binom{k-1}{j} \frac{(2k-3)!!}{(2(i-j)-1)!!} \binom{k-1}{i-1-j}.
\end{aligned}$$

To continue, we have

$$\begin{aligned}
\text{LHS} &= \sum_{j=(0 \vee i-k)+1}^{i \wedge k} \frac{(2k-1)!!}{(2j-1)!!} \binom{k-1}{j-1} \frac{(2k-1)!!}{(2(i-j)-1)!!} \binom{k}{i-j} \\
&+ \sum_{j=0 \vee i-k+1}^{i \wedge k-1} \frac{(2k-1)!!}{(2j-1)!!} \binom{k-1}{j} \frac{(2k-1)!!}{(2(i-j)-1)!!} \binom{k-1}{i-j} \\
&- \sum_{j=0 \vee i-k}^{i-1 \wedge k-1} \frac{(2k-1)!!}{(2j+1)!!} \binom{k-1}{j} \frac{(2k-1)!!}{(2(i-j)-3)!!} \binom{k-1}{i-1-j} \\
&+ (2k-1) \sum_{j=0 \vee i-k}^{i-1 \wedge k-1} \frac{(2k-1)!!}{(2j+1)!!} \binom{k-1}{j} \frac{(2k-1)!!}{(2(i-j)-1)!!} \binom{k-1}{i-1-j} \\
&+ (2k-1) \sum_{j=0 \vee i-k}^{i-1 \wedge k-1} \frac{(2k-3)!!}{(2j-1)!!} \binom{k-1}{j} \frac{(2k-3)!!}{(2(i-j)-3)!!} \binom{k-1}{i-1-j} \\
&- (2k-1) \sum_{j=0 \vee i-k+1}^{i-1 \wedge k-2} 2(k-1-j)2(k-i+j) \frac{(2k-3)!!}{(2j+1)!!} \binom{k-1}{j} \frac{(2k-3)!!}{(2(i-j)-1)!!} \binom{k-1}{i-1-j} \\
&= \sum_{j=(0 \vee i-k)+1}^{i \wedge k} \frac{(2k-1)!!}{(2j-1)!!} \binom{k-1}{j-1} \frac{(2k-1)!!}{(2(i-j)-1)!!} \binom{k}{i-j} \\
&+ \sum_{j=0 \vee i-k+1}^{i \wedge k-1} \frac{(2k-1)!!}{(2j-1)!!} \binom{k-1}{j} \frac{(2k-1)!!}{(2(i-j)-1)!!} \binom{k-1}{i-j} \\
&- \sum_{j=0 \vee i-k}^{i-1 \wedge k-1} (2k-1)(2(i-j)-1) \frac{(2k-3)!!}{(2j+1)!!} \binom{k-1}{j} \frac{(2k-1)!!}{(2(i-j)-1)!!} \binom{k-1}{i-1-j} \\
&+ \sum_{j=0 \vee i-k}^{i-1 \wedge k-1} (2k-1)^2 \frac{(2k-3)!!}{(2j+1)!!} \binom{k-1}{j} \frac{(2k-1)!!}{(2(i-j)-1)!!} \binom{k-1}{i-1-j} \\
&+ \sum_{j=0 \vee i-k}^{i-1 \wedge k-1} (2j+1)(2(i-j)-1) \frac{(2k-3)!!}{(2j+1)!!} \binom{k-1}{j} \frac{(2k-1)!!}{(2(i-j)-1)!!} \binom{k-1}{i-1-j} \\
&- \sum_{j=0 \vee i-k+1}^{i-1 \wedge k-2} (2k-2(j+1))(2k-2(i-j)) \frac{(2k-3)!!}{(2j+1)!!} \binom{k-1}{j} \frac{(2k-1)!!}{(2(i-j)-1)!!} \binom{k-1}{i-1-j}.
\end{aligned}$$

We consider combining the last four terms. Because

$$\begin{aligned}
&- (2k-1)(2(i-j)-1) + (2k-1)^2 \\
&+ (2j+1)(2(i-j)-1) - (2k-2(j+1))(2k-2(i-j)) \\
&= - (2k-2j-2)(2(i-j)-1) + (2k-1)^2 \\
&- (2k-2j-2)(2k-2(i-j)) \\
&= (2k-1)(2k-1-2k+2j+2) \\
&= (2k-1)(2j+1),
\end{aligned}$$

we have

$$\begin{aligned}
\text{LHS} &= \sum_{j=(0 \vee i-k)+1}^{i \wedge k} \frac{(2k-1)!!}{(2j-1)!!} \binom{k-1}{j-1} \frac{(2k-1)!!}{(2(i-j)-1)!!} \binom{k}{i-j} \\
&+ \sum_{j=0 \vee i-k+1}^{i \wedge k-1} \frac{(2k-1)!!}{(2j-1)!!} \binom{k-1}{j} \frac{(2k-1)!!}{(2(i-j)-1)!!} \binom{k-1}{i-j} \\
&+ \sum_{j=0 \vee i-k}^{i-1 \wedge k-1} (2k-1)(2j+1) \frac{(2k-3)!!}{(2j+1)!!} \binom{k-1}{j} \frac{(2k-1)!!}{(2(i-j)-1)!!} \binom{k-1}{i-1-j} \\
&= \sum_{j=(0 \vee i-k)+1}^{i \wedge k} \frac{(2k-1)!!}{(2j-1)!!} \binom{k-1}{j-1} \frac{(2k-1)!!}{(2(i-j)-1)!!} \binom{k}{i-j} \\
&+ \sum_{j=0 \vee i-k+1}^{i \wedge k-1} \frac{(2k-1)!!}{(2j-1)!!} \binom{k-1}{j} \frac{(2k-1)!!}{(2(i-j)-1)!!} \binom{k-1}{i-j} \\
&+ \sum_{j=0 \vee i-k}^{i-1 \wedge k-1} \frac{(2k-1)!!}{(2j-1)!!} \binom{k-1}{j} \frac{(2k-1)!!}{(2(i-j)-1)!!} \binom{k-1}{i-1-j} \\
&= \sum_{j=0 \vee i-k+1}^{i \wedge k-1} \frac{(2k-1)!!}{(2j-1)!!} \binom{k}{j} \frac{(2k-1)!!}{(2(i-j)-1)!!} \binom{k}{i-j} \\
&= (-1)^i A_i^k.
\end{aligned}$$

Finally, we complete the proof of Eq. (18).

□

B.3 PROOF OF COROLLARY 3.3

Proof. The Taylor expansion of $K(r)$ is

$$\begin{aligned}
K(r) &= e^{-p} \frac{h^2}{1+h^2} \sum_{n=0}^{\infty} \frac{R_n(p)}{n!(1+h^2)^n} r^n \\
&= \sum_{n=0}^{\infty} \frac{e^{-p} R_n(p)}{n!} \cdot \frac{h^2}{(1+h^2)^{n+1}} \langle x, x' \rangle^n \\
&= \sum_{n=0}^{\infty} \frac{e^{-p} R_n(p)}{n!} \cdot \frac{h^2}{(1+h^2)^{n+1}} \langle x^{\otimes n}, x'^{\otimes n} \rangle \\
&= \sum_{n=0}^{\infty} \left\langle \frac{he^{-\frac{p}{2}} R_n^{\frac{1}{2}}(p)}{\sqrt{n!(1+h^2)^{n+1}}} x^{\otimes n}, \frac{he^{-\frac{p}{2}} R_n^{\frac{1}{2}}(p)}{\sqrt{n!(1+h^2)^{n+1}}} x'^{\otimes n} \right\rangle.
\end{aligned} \tag{19}$$

Hence the feature mapping with respect to kernel (6) is

$$\phi(x) = \left(\frac{he^{-\frac{p}{2}} R_n^{\frac{1}{2}}(p)}{\sqrt{n!(1+h^2)^{n+1}}} x^{\otimes n} \right)_{n=0}^{\infty}.$$

For any target function

$$f(x) = \sum_{n=0}^{\infty} \langle F_n, x^{\otimes n} \rangle,$$

where $F_n \in \mathbb{R}^{d^n}$, we have

$$\begin{aligned} f(x) &= \sum_{n=0}^{\infty} \langle F_n, x^{\otimes n} \rangle \\ &= \left\langle \frac{\sqrt{n!(1+h^2)^{n+1}}}{he^{-\frac{p}{2}} R_n^{\frac{1}{2}}(p)} F_n, (\phi(x))_n \right\rangle. \end{aligned}$$

Hence, we have

$$\begin{aligned} \|f\|_{\mathcal{H}_K}^2 &\leq \left\| \left(\frac{\sqrt{n!(1+h^2)^{n+1}}}{he^{-\frac{p}{2}} R_n^{\frac{1}{2}}(p)} F_n \right)_{n=0}^{\infty} \right\|_{\mathcal{H}}^2 \\ &= \sum_{n=0}^{\infty} \frac{n!(1+h^2)^{n+1}}{h^2 e^{-p} R_n(p)} \|F_n\|_{\text{Fr}}^2 \\ &= \frac{e^p}{h^2} \sum_{n=0}^{\infty} \frac{n!(1+h^2)^{n+1}}{R_n(p)} \|F_n\|_{\text{Fr}}^2, \end{aligned}$$

where $\|\cdot\|_{\text{Fr}}$ the the Frobenius norm.

Let

$$D(f) := \frac{e^p}{h^2} \sum_{n=0}^{\infty} \frac{n!(1+h^2)^{n+1}}{R_n(p)} \|F_n\|_{\text{Fr}}^2.$$

By Proposition (2.1), we conclude that

$$\|f\|_{\mathcal{F}} = \|f\|_{\mathcal{H}_K} \leq \sqrt{D(f)}.$$

Furthermore, there exist $v : \mathcal{W} \rightarrow \mathbb{R}$ such that

$$f(x) = \mathbb{E}_{w \sim \mathcal{N}(0, I_d)} [B(w^\top x) v(w)],$$

and

$$\|v\|_{\mathcal{H}_W} \leq \sqrt{D(f)}.$$

□

B.4 PROOF OF THEOREM 3.4

Recall that in (1), we assume a mild condition that v is L_v -Lipschitz continuous. Because

$$|v(w)|^2 \leq (|v(\mathbf{0})| + L_v \|w\|)^2 \leq 2v(\mathbf{0})^2 + 2L_v^2 \|w\|^2.$$

By setting $R = \sqrt{2L_v^2 d + 2|v(\mathbf{0})|^2}$, we have

$$\mathbb{E}_{w \sim \mathcal{N}(0, I_d)} [v(w)^2] \leq R^2.$$

Proof. Let $W = (w_1, w_2, \dots, w_M)$ and $v_m = v(w_m)$. We already have $\varphi(x) := \mathbb{E}_{w \sim \mathcal{N}(0, I_d)} [B(w^\top x) v(w)]$.

To obtain the desired result, we consider the concentration property of the random variable

$$\mathbb{E}_x |\hat{\varphi}(x) - \varphi(x)| = \mathbb{E}_x \left| \frac{1}{M} \sum_{m=1}^M B(w_m^\top x) v(w_m) - \varphi(x) \right|,$$

in which the randomness comes from W .

Naturally, we consider

$$\begin{aligned}
& \mathbb{E}_W \exp \left(\lambda^2 \left(\mathbb{E}_x \left| \frac{1}{M} \sum_{m=1}^M B(w_m^\top x) v(w_m) - \varphi(x) \right| \right)^2 \right) \\
& \leq \mathbb{E}_W \exp \left(\lambda^2 \mathbb{E}_x \left(\frac{1}{M} \sum_{m=1}^M B(w_m^\top x) v(w_m) - \varphi(x) \right)^2 \right) \\
& \leq \mathbb{E}_W \mathbb{E}_x \exp \left(\lambda^2 \left(\frac{1}{M} \sum_{m=1}^M B(w_m^\top x) v(w_m) - \varphi(x) \right)^2 \right) \\
& = \mathbb{E}_x \mathbb{E}_W \exp \left(\lambda^2 \left(\frac{1}{M} \sum_{m=1}^M B(w_m^\top x) v(w_m) - \varphi(x) \right)^2 \right),
\end{aligned} \tag{20}$$

where we used Jensen's inequality twice.

Next, we prove that $B(w_m^\top x) v(w_m) - \varphi(x)$ are sub-gaussian random variables for every $w_m \sim \mathcal{N}(0, I_d)$ and every $x \in \mathbb{R}$. In addition, they have a uniform sub-gaussian norm.

To start with, for every $x \in \mathbb{R}$, we have the following estimation.

$$\begin{aligned}
& (B(w_m^\top x) v(w_m) - \varphi(x))^2 \\
& \leq 2B(w_m^\top x)^2 v(w_m)^2 + 2\varphi(x)^2 \\
& \leq 2B(w_m^\top x)^2 (L_v \|w_m - \mathbf{0}\|_2 + |v(\mathbf{0})|)^2 + 2(\mathbb{E}_{w \sim \mathcal{N}(0, I_d)} [B(w^\top x) v(w)])^2 \\
& \leq 2B(w_m^\top x)^2 (2L_v^2 \|w_m - \mathbf{0}\|_2^2 + 2|v(\mathbf{0})|^2) + 2\mathbb{E}_{w \sim \mathcal{N}(0, I_d)} [B(w^\top x)^2] \mathbb{E}_{w \sim \mathcal{N}(0, I_d)} [v(w)^2] \\
& \leq 4L_v^2 \|w_m\|_2^2 + 4|v(\mathbf{0})|^2 + 2R^2,
\end{aligned}$$

where we used the fact that v is L_v -Lipschitz and $0 \leq B(w^\top x) \leq 1$.

Therefore, we have

$$\begin{aligned}
& \mathbb{E}_W \exp (\lambda^2 (B(w_m^\top x) v(w_m) - \varphi(x))^2) \\
& \leq \mathbb{E}_W \exp (\lambda^2 (4L_v^2 \|w_m\|_2^2 + 4|v(\mathbf{0})|^2 + 2R^2)) \\
& = \exp (\lambda^2 (4|v(\mathbf{0})|^2 + 2R^2)) \cdot \mathbb{E}_W \exp (4L_v^2 \lambda^2 \|w_m\|_2^2) \\
& = \exp (\lambda^2 (4|v(\mathbf{0})|^2 + 2R^2)) \cdot \prod_{i=1}^d \mathbb{E}_{w_{m,d} \sim \mathcal{N}(0,1)} \exp (4L_v^2 \lambda^2 w_{m,d}^2) \\
& = \exp (\lambda^2 (4|v(\mathbf{0})|^2 + 2R^2)) \cdot \prod_{i=1}^d \frac{1}{\sqrt{1 - 8L_v^2 \lambda^2}}
\end{aligned} \tag{21}$$

By applying $\frac{1}{1-x} \leq e^{2x}$ over $x \in [0, 1/2]$, we have that for $\lambda^2 \leq \frac{1}{16L_v^2}$,

$$\begin{aligned}
& \exp (\lambda^2 (4|v(\mathbf{0})|^2 + 2R^2)) \cdot \prod_{i=1}^d \frac{1}{\sqrt{1 - 8L_v^2 \lambda^2}} \\
& \leq \exp (\lambda^2 (4|v(\mathbf{0})|^2 + 2R^2)) \exp (8dL_v^2 \lambda^2) \\
& = \exp (\lambda^2 (8dL_v^2 + 4|v(\mathbf{0})|^2 + 2R^2)) \\
& \leq \exp (\lambda^2 (16dL_v^2 + 4|v(\mathbf{0})|^2 + 2R^2)) \\
& \leq \exp (\lambda^2 \cdot 10R^2).
\end{aligned} \tag{22}$$

To summarize, let $Y_m = B(w_m^\top x) v(w_m) - \varphi(x)$, then for $\lambda^2 \leq 1/(10R^2)$, it holds that

$$\mathbb{E}_W \exp (\lambda^2 Y_m^2) \leq \exp (\lambda^2 \cdot 10R^2).$$

By Lemma A.6, we have that for all $\lambda \in \mathbb{R}$,

$$\mathbb{E}_W \exp(\lambda Y_m) \leq \exp(\lambda^2 \cdot 10R^2). \quad (23)$$

Note that $Y_1, Y_2, \dots, Y_M$ are independent. Therefore, we have

$$\begin{aligned} & \mathbb{E}_W \exp \left(\lambda \left(\frac{1}{M} \sum_{m=1}^M B(w_m^\top x) v(w_m) - \varphi(x) \right) \right) \\ &= \mathbb{E}_W \exp \left(\frac{\lambda}{M} \sum_{m=1}^M Y_m \right) = \prod_{m=1}^M \mathbb{E}_{w_m} \exp \left(\frac{\lambda}{M} Y_m \right) \\ &\leq \exp(\lambda^2 \cdot 10R^2/M). \end{aligned} \quad (24)$$

By Lemma A.6 again, we have that for $\lambda^2 \leq M/(160R^2)$,

$$\begin{aligned} & \mathbb{E}_W \exp \left(\lambda^2 \left(\frac{1}{M} \sum_{m=1}^M B(w_m^\top x) v(w_m) - \varphi(x) \right)^2 \right) \\ &\leq \exp(160R^2 \lambda^2/M). \end{aligned} \quad (25)$$

Taking expectation over x on both sides and plugging it back to (20), we have that

$$\mathbb{E}_W \exp \left(\lambda^2 \left(\mathbb{E}_x \left| \frac{1}{M} \sum_{m=1}^M B(w_m^\top x) v(w_m) - \varphi(x) \right| \right)^2 \right) \leq \exp(160R^2 \lambda^2/M).$$

Because $\sqrt{2}\sqrt{160R^2} \leq 18R$, by Lemma A.5, we conclude that

$$\left\| \mathbb{E}_x \left| \frac{1}{M} \sum_{m=1}^M B(w_m^\top x) v(w_m) - \varphi(x) \right| \right\|_{\psi_2} \leq \frac{18R}{\sqrt{M}}.$$

Consequently, applying Lemma A.4, for $\delta > 0$, by taking some $\epsilon = \frac{18R\sqrt{\log(4/\delta)}}{\sqrt{M}}$, we have that

$$\begin{aligned} & P(\mathbb{E}_x |\hat{\varphi}(x) - \varphi(x)| \geq \epsilon) \\ &= P \left(\mathbb{E}_x \left| \frac{1}{M} \sum_{m=1}^M B(w_m^\top x) v(w_m) - \varphi(x) \right| \geq \epsilon \right) \\ &\leq 2 \exp \left(-\frac{M\epsilon^2}{(18R)^2} \right) \leq \delta/2. \end{aligned}$$

Hence, with probability of at least $1 - \delta/2$, it holds that

$$\mathbb{E}_x |\hat{\varphi}(x) - \varphi(x)| \leq \frac{18R\sqrt{\log(4/\delta)}}{\sqrt{M}}.$$

In the remaining part of the proof, we consider the high probability bound of $\sum_{m=1}^M v_m^2$. To start with, we show that $v(w)$ is a sub-gaussian random variable in which $w \sim \mathcal{N}(0, I_d)$.

$$\begin{aligned} \mathbb{E} \exp(\lambda^2 v(w)^2) &\leq \mathbb{E} \exp(\lambda^2 (2L_v^2 \|w_m\|_2^2 + 2|v(\mathbf{0})|^2)) \\ &\leq \exp((4L_v^2 d + 2|v(\mathbf{0})|^2) \lambda^2), \end{aligned}$$

for λ such that $(4L_v^2 d + 2|v(\mathbf{0})|^2) \lambda^2 \leq 1$. By Lemma A.5, we have $\|v(w)\|_{\psi_2}^2 \leq (4L_v^2 d + 2|v(\mathbf{0})|^2)/\log 2 \leq 4R^2$. Hence, by Lemma A.3, we have $\|v(w)^2\|_{\psi_1} = \|v(w)\|_{\psi_2}^2 \leq 4R^2$. By triangle inequality, we have $\|v(w)^2 - \mathbb{E}[v(w)^2]\|_{\psi_1} \leq$

$\|v(w)^2\|_{\psi_1} + \|\mathbb{E}[v(w)^2]\|_{\psi_1}$. Given that $\mathbb{E}[v(w)^2]$ is a constant with an upper bound R^2 , by the definition of the sub-exponential norm, we have $\|\mathbb{E}[v(w)^2]\|_{\psi_1} \leq \mathbb{E}[v(w)^2]/\log 2 \leq 2R^2$. To conclude, we have that $\|v(w)^2 - \mathbb{E}[v(w)^2]\|_{\psi_1} \leq 6R^2$.

We apply Lemma A.8 for random variables $X_m = v(w_m)^2 - \mathbb{E}[v(w)^2]$ by setting $t = 24R^2 \left(\sqrt{\frac{\log(2/\delta)}{M}} + \frac{\log(2/\delta)}{M} \right)$. We obtain

$$P \left(\frac{1}{M} \sum_{m=1}^M v(w_m)^2 - \mathbb{E}[v(w)^2] > t \right) \leq \exp \left(- \min \left\{ \frac{Mt^2}{16\|X\|_{\psi_1}^2}, \frac{Mt}{4\|X\|_{\psi_1}} \right\} \right) \leq \frac{\delta}{2}.$$

Because $\mathbb{E}[v(w)^2] \leq R^2$, we obtain that

$$P \left(\frac{1}{M} \sum_{m=1}^M v(w_m)^2 - R^2 > t \right) \leq P \left(\frac{1}{M} \sum_{m=1}^M v(w_m)^2 - \mathbb{E}[v(w)^2] > t \right) \leq \frac{\delta}{2}.$$

Therefore, with probability of at least $1 - \delta/2$, we have

$$\frac{1}{M} \sum_{m=1}^M v(w_m)^2 \leq R^2 + 24R^2 \left(\sqrt{\frac{\log(2/\delta)}{M}} + \frac{\log(2/\delta)}{M} \right) \leq R^2 + 24R^2 \left(\sqrt{\log(2/\delta)} + \log(2/\delta) \right).$$

Without loss of generality, we assume $\delta < 1/2$, then $1 < \sqrt{\log(2/\delta)} < \log(2/\delta)$ and hence

$$\frac{1}{M} \sum_{m=1}^M v(w_m)^2 \leq 49R^2 \log(2/\delta).$$

Combining the two inequalities and taking the union bound of the probabilities, we have that with probability of at least $1 - \delta$, it holds that

$$\mathbb{E}_x |\hat{\varphi}(x) - \varphi(x)| \leq \frac{18R\sqrt{\log(4/\delta)}}{\sqrt{M}},$$

and

$$\frac{1}{M} \sum_{m=1}^M v(w_m)^2 \leq 49R^2 \log(2/\delta).$$

□

C DEFERRED PROOFS IN SECTION 4

C.1 PROOF OF PROPOSITION 4.1

Proof. To start with, we define the Gaussian function with parameter h as

$$\phi_h(x) := \frac{1}{\sqrt{2\pi}h} \exp \left(-\frac{x^2}{2h^2} \right).$$

First, we approximate σ by $\sigma * \phi_h = \int_{\mathbb{R}} \sigma(x-y)\phi_h(y)dy$. Because σ is L -Lipschitz continuous, we have that $|\sigma(x) -$

$|\sigma(x - y)| \leq L|y|$. Together with the fact $|\sigma| \leq \|\sigma\|_\infty$, we have that

$$\begin{aligned}
& |\sigma(x) - (\sigma * \phi_h)(x)| \\
&= \left| \sigma(x) - \int_{\mathbb{R}} \sigma(x - y) \phi_h(y) dy \right| \\
&\leq \int_{\mathbb{R}} |\sigma(x) - \sigma(x - y)| \phi_h(y) dy \\
&= \int_{[-\delta, \delta]} |\sigma(x) - \sigma(x - y)| \phi_h(y) dy + \int_{\mathbb{R} - [-\delta, \delta]} |\sigma(x) - \sigma(x - y)| \phi_h(y) dy \\
&\leq \int_{[-\delta, \delta]} L|y| \phi_h(y) dy + \int_{\mathbb{R} - [-\delta, \delta]} 2\|\sigma\|_\infty \phi_h(y) dy \\
&\leq L\delta + 2\|\sigma\|_\infty \cdot P\left(|Z| \geq \frac{\delta}{h}\right),
\end{aligned}$$

where $Z \sim \mathcal{N}(0, 1)$. The tail probability of Gaussian random variable is estimated as

$$\begin{aligned}
P\left(|Z| \geq \frac{\delta}{h}\right) &= 2P\left(Z \geq \frac{\delta}{h}\right) \\
&\stackrel{\lambda \geq 0}{\leq} 2P\left(e^{\lambda Z} \geq e^{\frac{\lambda \delta}{h}}\right) \\
&\leq 2 \inf_{\lambda > 0} \frac{\mathbb{E} e^{\lambda Z}}{e^{\frac{\lambda \delta}{h}}} \\
&= 2 \exp\left(-\frac{\delta^2}{2h^2}\right).
\end{aligned}$$

By taking

$$\delta = \frac{\epsilon}{4L}, \quad h \leq \frac{\epsilon}{4\sqrt{2}L\sqrt{\log \frac{16\|\sigma\|_\infty}{\epsilon}}},$$

we have

$$\begin{aligned}
& \left| \sigma(x) - \int_{\mathbb{R}} \sigma(x - y) \phi_h(y) dy \right| \\
&\leq L\delta + 2\|\sigma\|_\infty \cdot P\left(|Z| \geq \frac{\delta}{h}\right) \\
&\leq \frac{\epsilon}{4} + \frac{\epsilon}{4} = \frac{\epsilon}{2}.
\end{aligned}$$

In the second step, we approximate $\sigma * \phi_h$ by the Riemann sum $\sum_{i=1}^N f(y_i) \cdot (y_i - y_{i-1}) \cdot \phi_h(x - y_i)$.

For the convolution part, we have

$$\begin{aligned}
(\sigma * \phi_h)(x) &= \int_{\mathbb{R}} \sigma(x - y) \phi_h(y) dy \\
&= \int_{\mathbb{R}} \sigma(y) \phi_h(x - y) dy \\
&= \int_{\mathcal{K}} \sigma(y) \phi_h(x - y) dy \\
&= \sum_{i=1}^N \int_{y_{i-1}}^{y_i} \sigma(y) \phi_h(x - y) dy.
\end{aligned}$$

Then we have

$$\begin{aligned}
& \left| (\sigma * \phi_h)(x) - \sum_{i=1}^N \sigma(y_i) \cdot (y_i - y_{i-1}) \cdot \phi_h(x - y_i) \right| \\
&= \left| \sum_{i=1}^N \int_{y_{i-1}}^{y_i} \sigma(y) \phi_h(x - y) dy - \sum_{i=1}^N \int_{y_{i-1}}^{y_i} \sigma(y_i) \phi_h(x - y_i) dy \right| \\
&\leq \left| \sum_{i=1}^N \int_{y_{i-1}}^{y_i} \sigma(y) \phi_h(x - y) dy - \sum_{i=1}^N \int_{y_{i-1}}^{y_i} \sigma(y_i) \phi_h(x - y) dy \right| \\
&\quad + \left| \sum_{i=1}^N \int_{y_{i-1}}^{y_i} \sigma(y_i) \phi_h(x - y) dy - \sum_{i=1}^N \int_{y_{i-1}}^{y_i} \sigma(y_i) \phi_h(x - y_i) dy \right| \\
&\leq \sum_{i=1}^N \int_{y_{i-1}}^{y_i} |\sigma(y) - \sigma(y_i)| \phi_h(x - y) dy + \sum_{i=1}^N \int_{y_{i-1}}^{y_i} |\sigma(y_i)| \cdot |\phi_h(x - y) - \phi_h(x - y_i)| dy \\
&\leq \sum_{i=1}^N L(y_i - y_{i-1}) \int_{y_{i-1}}^{y_i} \phi_h(x - y) dy + \|\sigma\|_\infty \sum_{i=1}^N \int_{y_{i-1}}^{y_i} |\phi_h(x - y) - \phi_h(x - y_i)| dy.
\end{aligned} \tag{26}$$

For the first term, if $|\mathcal{K}|/N \leq \epsilon/4L$, then we have

$$\sum_{i=1}^N L(y_i - y_{i-1}) \int_{y_{i-1}}^{y_i} \phi_h(x - y) dy \leq \frac{L|\mathcal{K}|}{N} \sum_{i=1}^N \int_{y_{i-1}}^{y_i} \phi_h(x - y) dy \leq \frac{L|\mathcal{K}|}{N} \leq \frac{\epsilon}{4}. \tag{27}$$

For the second term, we first consider the derivative of $\phi_h(x)$.

$$\begin{aligned}
|\phi'_h(x)| &= \left| \frac{1}{\sqrt{2\pi}h^2} \cdot \frac{x}{h} \cdot \exp\left(-\frac{1}{2}\left(\frac{x}{h}\right)^2\right) \right| \\
&\leq \frac{1}{\sqrt{2\pi}h^2} \exp\left(-\frac{1}{4}\left(\frac{x}{h}\right)^2\right)
\end{aligned}$$

where we use the inequality $x \leq \exp(x^2/4)$.

Taking $t = \sqrt{4 \log\left(\frac{8\|\sigma\|_\infty |\mathcal{K}|^2}{\sqrt{2\pi}} \cdot \frac{1}{\epsilon N h^2}\right)}$, if $|x| > th$, then

$$|\phi'_h(x)| \leq \frac{\epsilon N}{8\|\sigma\|_\infty |\mathcal{K}|^2}.$$

If $|x| \leq th$, then

$$\begin{aligned}
|\phi'_h(x)| &= \left| \frac{1}{\sqrt{2\pi}h^2} \cdot \frac{x}{h} \cdot \exp\left(-\frac{1}{2}\left(\frac{x}{h}\right)^2\right) \right| \\
&\leq \left| \frac{1}{\sqrt{2\pi}h^2} \sup_{t \in \mathbb{R}} \left\{ t \exp\left(-\frac{t^2}{2}\right) \right\} \right| \\
&= \frac{1}{\sqrt{2\pi}eh^2}.
\end{aligned}$$

Consequently, for the second term, it holds that

$$\begin{aligned}
& \|\sigma\|_\infty \sum_{i=1}^N \int_{y_{i-1}}^{y_i} |\phi_h(x-y) - \phi_h(x-y_i)| dy \\
&= \|\sigma\|_\infty \sum_{i=1}^N \int_{y_{i-1}}^{y_i} \left| \int_y^{y_i} \phi'_h(x-z) dz \right| dy \\
&\leq \|\sigma\|_\infty \sum_{i=1}^N \int_{y_{i-1}}^{y_i} \left| \sup_{z \in [x-y_i, x-y_{i-1}]} |\phi'_h(z)| |y-y_i| \right| dy \\
&\leq \|\sigma\|_\infty \sum_{i=1}^N \sup_{z \in [x-y_i, x-y_{i-1}]} |\phi'_h(z)| \left(\frac{|\mathcal{K}|}{N} \right)^2 \\
&\leq \|\sigma\|_\infty \frac{2th}{\frac{|\mathcal{K}|}{N}} \cdot \frac{1}{\sqrt{2\pi}eh^2} \cdot \left(\frac{|\mathcal{K}|}{N} \right)^2 + \|\sigma\|_\infty N \cdot \frac{\epsilon N}{8\|\sigma\|_\infty |\mathcal{K}|^2} \cdot \left(\frac{|\mathcal{K}|}{N} \right)^2 \\
&= \|\sigma\|_\infty \sqrt{\frac{2}{\pi e}} \cdot \frac{t|\mathcal{K}|}{hN} + \frac{\epsilon}{8} \\
&= \|\sigma\|_\infty \sqrt{\frac{2}{\pi e}} \sqrt{4 \log \left(\frac{8\|\sigma\|_\infty |\mathcal{K}|}{\sqrt{2\pi}} \cdot \frac{|\mathcal{K}|}{\epsilon h^2 N} \right)} \cdot \frac{|\mathcal{K}|}{hN} + \frac{\epsilon}{8}.
\end{aligned} \tag{28}$$

The fifth line holds because there are at most $2thN/|\mathcal{K}|$ intervals in which $|\phi'_h| > \frac{\epsilon N}{8\|\sigma\|_\infty |\mathcal{K}|^2}$.

Let

$$\frac{|\mathcal{K}|}{N} \leq \frac{\epsilon h \sqrt{\pi e}}{16\sqrt{2}\|\sigma\|_\infty \log \left(\frac{8\|\sigma\|_\infty |\mathcal{K}|}{\sqrt{2\pi}eh^2} \right)} \wedge \frac{\epsilon}{4L} \ll 1.$$

Then

$$\begin{aligned}
& \|\sigma\|_\infty \sqrt{\frac{2}{\pi e}} \sqrt{4 \log \left(\frac{8\|\sigma\|_\infty |\mathcal{K}|}{\sqrt{2\pi}} \cdot \frac{|\mathcal{K}|}{\epsilon h^2 N} \right)} \cdot \frac{|\mathcal{K}|}{hN} \\
&= \|\sigma\|_\infty \sqrt{\frac{2}{\pi e}} \sqrt{4 \log \left(\frac{8\|\sigma\|_\infty |\mathcal{K}|}{\sqrt{2\pi}eh^2} \right)} + 4 \log \left(\frac{|\mathcal{K}|}{N} \right) \cdot \frac{|\mathcal{K}|}{hN} \\
&\leq \|\sigma\|_\infty \sqrt{\frac{2}{\pi e}} \sqrt{4 \log \left(\frac{8\|\sigma\|_\infty |\mathcal{K}|}{\sqrt{2\pi}eh^2} \right)} \cdot \frac{|\mathcal{K}|}{hN} \leq \frac{\epsilon}{8}.
\end{aligned} \tag{29}$$

Putting (27), (28) and (29) into (26), we conclude that

$$\left| (\sigma * \phi_h)(x) - \sum_{i=1}^N \sigma(y_i) \cdot (y_i - y_{i-1}) \cdot \phi_h(x - y_i) \right| \leq \frac{\epsilon}{2}.$$

Hence,

$$\begin{aligned}
& \left| \sigma(x) - \sum_{i=1}^N \sigma(y_i) \cdot (y_i - y_{i-1}) \cdot \phi_h(x - y_i) \right| \\
&\leq |\sigma(x) - (\sigma * \phi_h)(x)| + \left| (\sigma * \phi_h)(x) - \sum_{i=1}^N \sigma(y_i) \cdot (y_i - y_{i-1}) \cdot \phi_h(x - y_i) \right| \\
&\leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.
\end{aligned}$$

Let

$$B_i(x) = \exp \left(-\frac{(x - y_i)^2}{2h^2} \right), \quad a_i = \frac{|\mathcal{K}|}{\sqrt{2\pi}hN} \cdot \sigma(y_i),$$

then

$$\sum_{i=1}^N \sigma(y_i) \cdot (y_i - y_{i-1}) \cdot \phi_h(x - y_i) = \sum_{i=1}^N a_i B_i(x).$$

Hence

$$\left\| \sigma - \sum_{i=1}^N a_i B_i(x) \right\|_{\infty} \leq \epsilon,$$

and

$$\sum_{i=1}^N |a_i| \leq \sum_{i=1}^N \frac{|\mathcal{K}|}{\sqrt{2\pi}hN} \cdot |\sigma(y_i)| \leq \frac{\|\sigma\|_{\infty} |\mathcal{K}|}{\sqrt{2\pi}h}.$$

In addition,

$$\sum_{i=1}^N |a_i|^2 \leq \sum_{i=1}^N \frac{|\mathcal{K}|^2}{2\pi h^2 N^2} \cdot |\sigma(y_i)|^2 \leq \frac{\|\sigma\|_{\infty}^2 |\mathcal{K}|}{2\pi} \cdot \frac{|\mathcal{K}|}{Nh^2}.$$

To conclude, if one sets

$$h_i \equiv h \leq \frac{\epsilon}{4\sqrt{2}L\sqrt{\log \frac{16\|\sigma\|_{\infty}}{\epsilon}}}, \quad \frac{|\mathcal{K}|}{N} \leq \frac{\epsilon h \sqrt{\pi e}}{16\sqrt{2}\|\sigma\|_{\infty} \log \left(\frac{8\|\sigma\|_{\infty} |\mathcal{K}|}{\sqrt{2\pi} \epsilon h^2} \right)} \wedge \frac{\epsilon}{4L}, \quad (30)$$

and c_i be the grid points of $\mathcal{K}$, then there exists $\{a_i\}_{i=1}^N$ such that

$$\left\| \sigma - \sum_{i=1}^N a_i B_i(x) \right\|_{\infty} \leq \epsilon,$$

and

$$\sum_{i=1}^N |a_i| \leq \frac{\|\sigma\|_{\infty} |\mathcal{K}|}{\sqrt{2\pi}h}, \quad \sum_{i=1}^N |a_i|^2 \leq \frac{\|\sigma\|_{\infty}^2 |\mathcal{K}|}{2\pi} \cdot \frac{|\mathcal{K}|}{Nh^2}.$$

We remark that the choice of c_i could be arbitrary as long as $c_i \in [y_{i-1}, y_i]$. And the L_2 bound actually implies the L_1 bound because $\sum_{i=1}^N |a_i| \leq \sqrt{\sum_{i=1}^N |a_i|^2} \sqrt{N}$.

Now, replacing ϵ with ϵ/R in (30), there exists $N > 0$ and $\{h_i, c_i, a_i\}_{i=1}^N$ such that

$$\left\| \sigma(x) - \sum_{i=1}^N a_i B_i(x) \right\|_{\infty} < \frac{\epsilon}{R}.$$

Thus

$$\begin{aligned} \|f^*(x) - \tilde{f}(x)\|_{\infty} &= \left\| \mathbb{E}_{w \sim \mathcal{N}(0,1)} \left[\left(\sigma(w^{\top} x) - \sum_{i=1}^N a_i B_i(w^{\top} x) \right) v(w) \right] \right\|_{\infty} \\ &\leq \mathbb{E}_{w \sim \mathcal{N}(0,1)} \left[\left\| \sigma(w^{\top} x) - \sum_{i=1}^N a_i B_i(w^{\top} x) \right\|_{\infty} |v(w)| \right] \\ &\leq \left\| \sigma(w^{\top} x) - \sum_{i=1}^N a_i B_i(w^{\top} x) \right\|_{\infty} \left(\mathbb{E}_{w \sim \mathcal{N}(0,1)} [v(w)^2] \right)^{\frac{1}{2}} \\ &\leq \frac{\epsilon}{R} \cdot R \leq \epsilon. \end{aligned} \quad (31)$$

□

C.2 PROOF OF THEOREM 4.2

In the proof, we attempt to approximate $\tilde{f}$ with finite-width random feature model. We clarify the notations here and denote

$$\varphi_i(x) = \mathbb{E}_{w \sim \mathcal{N}(0, I_d)} [B_i(w^\top x) v(w)], \quad \hat{\varphi}_i(x) = \frac{1}{M} \sum_{m=1}^M B_i(w_m^\top x) v_m.$$

Then

$$\tilde{f}(x) = \sum_{i=1}^N a_i \varphi_i(x), \quad \hat{f}(x) = \sum_{i=1}^N a_i \hat{\varphi}_i(x).$$

Proof.

For all $\epsilon > 0$, under the parameter settings of Proposition 4.1, there exists $\{a_i\}_{i=1}^N$ such that

$$\left\| \tilde{f}(x) - f^*(x) \right\|_\infty \leq \epsilon, \quad \left\| \sigma(x) - \sum_{i=1}^N a_i B_i(x) \right\|_\infty < \epsilon/R, \quad \sum_{i=1}^N a_i^2 \leq \frac{\|\sigma\|_\infty^2 |\mathcal{K}|^2}{2\pi h^2 N}.$$

So we first have

$$\begin{aligned} \mathbb{E}_x \left| \hat{f}(x) - f^*(x) \right| &= \mathbb{E}_x \left| \hat{f}(x) - \tilde{f}(x) + \tilde{f}(x) - f^*(x) \right| \\ &\leq \mathbb{E}_x \left| \hat{f}(x) - \tilde{f}(x) \right| + \mathbb{E}_x \left| \tilde{f}(x) - f^*(x) \right| \\ &\leq \mathbb{E}_x \left| \sum_{i=1}^N a_i (\hat{\varphi}_i(x) - \varphi_i(x)) \right| + \epsilon. \end{aligned} \tag{32}$$

Next, we aim to derive a high probability bound on $\mathbb{E}_x \left| \sum_{i=1}^N a_i (\hat{\varphi}_i(x) - \varphi_i(x)) \right|$. The proof techniques are similar to those of Theorem 3.4. First, we have

$$\begin{aligned} \sum_{i=1}^N a_i (\hat{\varphi}_i(x) - \varphi_i(x)) &= \sum_{i=1}^N a_i \left(\frac{1}{M} \sum_{m=1}^M B_i(w_m^\top x) v(w_m) - \varphi_i(x) \right) \\ &= \frac{1}{M} \sum_{m=1}^M \left(\sum_{i=1}^N a_i B_i(w_m^\top x) v(w_m) - \sum_{i=1}^N a_i \varphi_i(x) \right) \end{aligned}$$

It boils down to estimating the sub-gaussian norms of the random variables $Z_m = \sum_{i=1}^N a_i B_i(w_m^\top x) v(w_m) - \sum_{i=1}^N a_i \varphi_i(x)$ where $\{w_m\}_{m \in [M]} \stackrel{i.i.d.}{\sim} \mathcal{N}(0, I_d)$.

Consider

$$\begin{aligned} Z_m^2 &= \left(\sum_{i=1}^N a_i B_i(w_m^\top x) v(w_m) - \sum_{i=1}^N a_i \varphi_i(x) \right)^2 \\ &\leq 2 \left(\sum_{i=1}^N a_i B_i(w_m^\top x) v(w_m) \right)^2 + 2 \left(\sum_{i=1}^N a_i \varphi_i(x) \right)^2 \\ &= 2v(w_m)^2 \left(\sum_{i=1}^N a_i B_i(w_m^\top x) \right)^2 + 2 \left(\mathbb{E}_w \sum_{i=1}^N a_i B_i(w^\top x) v(w) \right)^2 \\ &\leq 2v(w_m)^2 \left(\sum_{i=1}^N a_i B_i(w_m^\top x) \right)^2 + 2\mathbb{E}_w \left(\sum_{i=1}^N a_i B_i(w^\top x) \right)^2 \mathbb{E}_w (v(w)^2). \end{aligned}$$

Because $\left\| \sigma(x) - \sum_{i=1}^N a_i B_i(x) \right\|_\infty < \epsilon/R$, we have $\left| \sum_{i=1}^N a_i B_i(x) \right| \leq \|\sigma\|_\infty + \epsilon/R$ for all x . Hence,

$$\begin{aligned} Z_m^2 &\leq 2v(w_m)^2 (\|\sigma\|_\infty + \epsilon/R)^2 + 2R^2 (\|\sigma\|_\infty + \epsilon/R)^2 \\ &\leq 2(L_v \|w_m\| + |v(\mathbf{0})|)^2 (\|\sigma\|_\infty + \epsilon/R)^2 + 2R^2 (\|\sigma\|_\infty + \epsilon/R)^2 \\ &\leq (4L_v^2 \|w_m\|^2 + 4|v(\mathbf{0})|^2 + 2R^2) (\|\sigma\|_\infty + \epsilon/R)^2. \end{aligned}$$

Similar to the estimation in Eq. (21) and Eq. (22), we have that for λ such that $10(\|\sigma\|_\infty R + \epsilon)^2 \lambda^2 \leq 1$, it holds that

$$\mathbb{E}_W e^{\lambda^2 Z_m^2} \leq e^{10(\|\sigma\|_\infty R + \epsilon)^2 \lambda^2}.$$

Similar to the estimation in Eq. (23), Eq. (24) and Eq. (25), we have that for λ such that $160(\|\sigma\|_\infty R + \epsilon)^2 \lambda^2 \leq M$, it holds that

$$\mathbb{E}_W e^{\lambda^2 (\sum_{m=1}^M Z_m/M)^2} \leq e^{160(\|\sigma\|_\infty R + \epsilon)^2 \lambda^2 / M}.$$

Hence, similar to Eq. (20), we have

$$\begin{aligned} \mathbb{E}_W e^{\lambda^2 (\mathbb{E}_x |\sum_{i=1}^N a_i (\hat{\varphi}_i(x) - \varphi_i(x))|)^2} &\leq \mathbb{E}_x \mathbb{E}_W e^{\lambda^2 (\sum_{i=1}^N a_i (\hat{\varphi}_i(x) - \varphi_i(x)))^2} \\ &= \mathbb{E}_x \mathbb{E}_W e^{\lambda^2 (\sum_{m=1}^M Z_m/M)^2} \\ &\leq e^{160(\|\sigma\|_\infty R + \epsilon)^2 \lambda^2 / M}. \end{aligned}$$

By Lemma A.5, we obtain $\left\| \mathbb{E}_x \left| \sum_{i=1}^N a_i (\hat{\varphi}_i(x) - \varphi_i(x)) \right| \right\|_{\psi_2} \leq 18(\|\sigma\|_\infty R + \epsilon)/\sqrt{M}$. By Lemma A.4, we have that

$$P \left(\mathbb{E}_x \left| \sum_{i=1}^N a_i (\hat{\varphi}_i(x) - \varphi_i(x)) \right| \geq \frac{18(\|\sigma\|_\infty R + \epsilon)\sqrt{\log(4/\delta)}}{\sqrt{M}} \right) \leq \frac{\delta}{2}.$$

Namely, with probability of at least $1 - \delta/2$, we have

$$\mathbb{E}_x \left| \sum_{i=1}^N a_i (\hat{\varphi}_i(x) - \varphi_i(x)) \right| \leq \frac{18(\|\sigma\|_\infty R + \epsilon)\sqrt{\log(4/\delta)}}{\sqrt{M}}.$$

Therefore, putting it back to (32), with probability of at least $1 - \delta/2$, we have

$$\mathbb{E}_x \left| \hat{f}(x) - f^*(x) \right| \leq \frac{18(\|\sigma\|_\infty R + \epsilon)\sqrt{\log(4/\delta)}}{\sqrt{M}} + \epsilon.$$

Further, by the proof of Theorem 3.4, the event

$$\frac{1}{M} \sum_{m=1}^M v_m^2 \leq 49R^2 \log(2/\delta)$$

happens with probability of at least $1 - \delta/2$.

Taking the union bounds of the probability, we conclude that with probability of at least $1 - \delta$, the inequalities hold:

$$\mathbb{E}_x \left| \hat{f}(x) - f^*(x) \right| \leq \frac{18(\|\sigma\|_\infty R + \epsilon)\sqrt{\log(4/\delta)}}{\sqrt{M}} + \epsilon,$$

and

$$\frac{1}{M} \sum_{m=1}^M v_m^2 \leq 49R^2 \log(2/\delta).$$

□

D DEFERRED PROOFS IN SECTION 5

We use Rademacher complexity to obtain the result in Theorem 5.1. We first recall the definition of Rademacher complexity. Suppose we are given samples $S = \{z_i = (x_i, y_i)\}_{i=1}^n$. Let

$$\ell \circ f_V := \{(x, y) \mapsto \ell(f(x), y) : f \in f_V\}$$

be the function class. Let

$$f_V \circ S := \{(f(x_1), \dots, f(x_n)) : f \in f_V\},$$

$$\ell \circ f_{\mathcal{V}} \circ S := \{(\ell(f(x_1), y_1), \dots, \ell(f(x_n), y_n)) : f \in f_{\mathcal{V}}\}$$

be vector sets. The Rademacher complexity of a function class $\mathcal{H}$ with respect to S is defined as

$$\mathcal{R}(\mathcal{H} \circ S) := \frac{1}{n} \mathbb{E}_{\boldsymbol{\xi}} \sup_{h \in \mathcal{H}} \sum_{i=1}^n \xi_i h(z_i),$$

where $\boldsymbol{\xi} = (\xi_1, \dots, \xi_n)$ and $\{\xi_i\}_{i \in [n]}$ are independent symmetric Bernoulli random variables.

Next, we introduce three lemmas for proving Theorem 5.1. The first one is a technical tool.

Lemma D.1 (Talagrand's contraction principle (e.g., Exercise 6.7.7 in [Vershynin, 2018])) *Consider a bounded subset $T \subset \mathbb{R}^n$, and let $\{\xi_i\}_{i \in [n]}$ be independent symmetric Bernoulli random variables. If $\phi_i : \mathbb{R} \rightarrow \mathbb{R}$ are ρ -Lipschitz functions, then*

$$\mathbb{E}_{\boldsymbol{\xi}} \sup_{t \in T} \sum_{i=1}^n \xi_i \phi_i(t_i) \leq \rho \mathbb{E}_{\boldsymbol{\xi}} \sup_{t \in T} \sum_{i=1}^n \xi_i t_i.$$

Then, through Lemma D.1, we can obtain the following result describing the Rademacher complexity of the function class of interests.

Lemma D.2 *All $f \in f_{\mathcal{V}}$ are bounded:*

$$\|f\|_{\infty} \leq \frac{7\|\sigma\|_{\infty} |\mathcal{K}| R \sqrt{\log(2/\delta)}}{h \sqrt{2\pi}}.$$

Furthermore, the Rademacher complexity of $\ell \circ f_{\mathcal{V}}$ with respect to samples S is bounded as

$$\mathcal{R}(\ell \circ f_{\mathcal{V}} \circ S) \leq \frac{7\rho\|\sigma\|_{\infty} |\mathcal{K}| R \sqrt{\log(2/\delta)/2\pi}}{h \sqrt{n}}.$$

For the coherence of the statements, we give the proof of Lemma D.2 at the end of this section. Finally, we derive the excess risk from the Rademacher complexity using the well known result in supervised learning illustrated below.

Lemma D.3 (e.g., Theorem 26.5 in [Shalev-Shwartz and Ben-David, 2014]) *Assume that for all $z = (x, y) \sim \mathbb{P}$ and $f \in f_{\mathcal{V}}$ we have that $|\ell(f(x), y)| \leq c$. Then for any $\hat{f} \in f_{\mathcal{V}}$, with probability of at least $1 - \delta$ over $\{(x_i, y_i)\}_{i \in [n]} \stackrel{i.i.d.}{\sim} \mathbb{P}$, it holds that*

$$L_D(f_S) - L_D(\hat{f}) \leq 2\mathcal{R}(\ell \circ f_{\mathcal{V}} \circ S) + 5c \sqrt{\frac{2 \log(8/\delta)}{n}}.$$

Formal Proof of Theorem 5.1. Under the conditions and parameter settings of $h, N, \{c_i\}_{i=1}^N$ in Theorem 4.2, with probability of at least $1 - \delta$ over $W = (w_1, \dots, w_M)$, there exists $\hat{f} \in f_{\mathcal{V}}$ such that

$$\mathbb{E}_x \left| \hat{f}(x) - f^*(x) \right| \leq \frac{18(\|\sigma\|_{\infty} R + \epsilon) \sqrt{\log(4/\delta)}}{\sqrt{M}} + \epsilon.$$

On the other hand, for all $f \in f_{\mathcal{V}}$ and (x, y) , we have that

$$|\ell(f(x), y)| \leq |\ell(0, y)| + \rho|f(x) - 0| \leq \rho \left(1 + \frac{7\|\sigma\|_{\infty} |\mathcal{K}| R \sqrt{\log(2/\delta)}}{h \sqrt{2\pi}} \right) =: c,$$

where we use the first part of Lemma D.2, the Lipschitz property of ℓ and the relation $|\ell(0, y)| \leq \rho$.

Apply the second part of Lemma D.2 and D.3 for $\hat{f}$. Then with probability of at least $1 - \delta$ over $\{(x_i, y_i)\}_{i \in [n]} \stackrel{i.i.d.}{\sim} \mathbb{P}$, we have that

$$\begin{aligned} L_D(f_S) - L_D(\hat{f}) &\leq 2\mathcal{R}(\ell \circ f_{\mathcal{V}} \circ S) + 5c \sqrt{\frac{2 \log(8/\delta)}{n}} \\ &\leq 2\rho \frac{7\|\sigma\|_{\infty} |\mathcal{K}| R \sqrt{\log(2/\delta)/2\pi}}{h} \sqrt{\frac{1}{n}} + 5\rho \left(1 + \frac{7\|\sigma\|_{\infty} |\mathcal{K}| R \sqrt{\log(2/\delta)/2\pi}}{h} \right) \sqrt{\frac{2 \log(8/\delta)}{n}} \\ &\leq 7\rho \left(1 + \frac{7\|\sigma\|_{\infty} |\mathcal{K}| R \sqrt{\log(2/\delta)}}{h} \right) \sqrt{\frac{2 \log(8/\delta)}{n}}. \end{aligned}$$

Next, we notice that with probability of at least $1 - \delta$ over $W = (w_1, \dots, w_M)$,

$$\begin{aligned} L_D(\hat{f}) - L_D(f^*) &= \mathbb{E}_{x, y \sim \mathcal{P}}[\ell(\hat{f}(x), y) - \ell(f^*(x), y)] \\ &\leq \rho \mathbb{E}_x \left| \hat{f}(x) - f^*(x) \right| \\ &\leq \rho \frac{18(\|\sigma\|_\infty R + \epsilon) \sqrt{\log(4/\delta)}}{\sqrt{M}} + \rho\epsilon. \end{aligned}$$

Combining the two inequalities and taking the union bounds of the probabilities, we conclude that with probability of at least $1 - 2\delta$ over W and S , it holds that

$$L_D(f_S) - L_D(f^*) \leq 7\rho \left(1 + \frac{7\|\sigma\|_\infty |\mathcal{K}| R \sqrt{\log(2/\delta)}}{h} \right) \sqrt{\frac{2 \log(8/\delta)}{n}} + \rho \frac{18(\|\sigma\|_\infty R + \epsilon) \sqrt{\log(4/\delta)}}{\sqrt{M}} + \rho\epsilon.$$

Without loss of generality, assume $h \leq 1$ and $\delta \leq 1/2$, then $1 \leq \sqrt{\log(2/\delta)}/h$, $\sqrt{\log(2/\delta)} \leq \sqrt{2 \log(8/\delta)}$. Consequently,

$$L_D(f_S) - L_D(f^*) \leq \rho \frac{14(1 + 7\|\sigma\|_\infty |\mathcal{K}| R) \log(8/\delta)}{h\sqrt{n}} + \rho \frac{18(\|\sigma\|_\infty R + \epsilon) \sqrt{\log(4/\delta)}}{\sqrt{M}} + \rho\epsilon.$$

Replacing δ with $\delta/2$, with probability of at least $1 - \delta$, we have

$$L_D(f_S) - L_D(f^*) \leq \rho \frac{14(1 + 7\|\sigma\|_\infty |\mathcal{K}| R) \log(16/\delta)}{h\sqrt{n}} + \rho \frac{18(\|\sigma\|_\infty R + \epsilon) \sqrt{\log(8/\delta)}}{\sqrt{M}} + \rho\epsilon.$$

Let $C = \max\{14(1 + 7\|\sigma\|_\infty |\mathcal{K}| R), 18(\|\sigma\|_\infty R + \epsilon)\}$, we obtain that

$$L_D(f_S) - L_D(f^*) \leq \frac{\rho C \log(16/\delta)}{h\sqrt{n}} + \frac{\rho C \sqrt{\log(8/\delta)}}{\sqrt{M}} + \rho\epsilon.$$

□

At the end of the proof, we supplement the proof of the second lemma. The proof of Lemma D.1 and D.3 can be found readily in the literature and are hence omitted.

Proof of Lemma D.2. Let $\phi_i(t) = \ell(t, y_i)$ and $t_i = f(x_i)$. Then $\phi_i(t)$ is ρ -Lipschitz continuous with respect to t . For the boundedness of $T = \{(f(x_1), \dots, f(x_n)) : f \in \mathcal{F}_\mathcal{V}\}$, we can see that for all $f \in \mathcal{F}_\mathcal{V}$, it holds that

$$\begin{aligned} |f| &= \left| \frac{1}{M} \sum_{k=1}^N a_k \sum_{m=1}^M B_k(w_m^\top x) v_m \right| \\ &\leq \frac{1}{M} \sqrt{\sum_{k=1}^N a_k^2} \cdot \sqrt{\sum_{k=1}^N \left(\sum_{m=1}^M B_k(w_m^\top x) v_m \right)^2} \\ &\leq \frac{1}{M} \sqrt{\sum_{k=1}^N a_k^2} \cdot \sqrt{\sum_{k=1}^N \left(\sum_{m=1}^M B_k^2(w_m^\top x) \sum_{m=1}^M v_m^2 \right)} \\ &\leq \frac{1}{M} \|\mathbf{a}\|_2 \cdot \sqrt{NM} \|\mathbf{v}\|_2 \\ &\leq \sqrt{\frac{N}{M}} \frac{\|\sigma\|_\infty |\mathcal{K}|}{h\sqrt{2\pi N}} \cdot \sqrt{49MR^2 \log\left(\frac{2}{\delta}\right)} \\ &= \frac{7\|\sigma\|_\infty |\mathcal{K}| R \sqrt{\log(2/\delta)}}{h\sqrt{2\pi}}. \end{aligned} \tag{33}$$

Hence,

$$\|f\|_\infty \leq \frac{7\|\sigma\|_\infty |\mathcal{K}| R \sqrt{\log(2/\delta)}}{h\sqrt{2\pi}},$$

and for all $t \in T$, $t = (f(x_1), \dots, f(x_n))$ and $\|t\| \leq \sqrt{n}\|f\|_\infty \leq \frac{7\|\sigma\|_\infty|\mathcal{K}|R\sqrt{n\log(2/\delta)}}{h\sqrt{2\pi}}$.

By applying Lemma D.1, we have

$$\mathbb{E}_\xi \sup_{f \in f_V} \sum_{i=1}^n \xi_i \ell(f(x_i), y_i) \leq \rho \mathbb{E}_\xi \sup_{f \in f_V} \sum_{i=1}^n \xi_i f(x_i).$$

To continue, let $K_1 = \frac{\|\sigma\|_\infty|\mathcal{K}|}{h\sqrt{2\pi N}}$, $K_2 = \sqrt{49MR^2\log\left(\frac{2}{\delta}\right)}$, $\mathbf{B}_i \in \mathbb{R}^{N \times M}$ with $(\mathbf{B}_i)_{k,m} = B_k(w_m^\top x_i)$, then we have

$$\begin{aligned} & \mathbb{E}_\xi \sup_{f \in f_V} \sum_{i=1}^n \xi_i f(x_i) \\ &= \mathbb{E}_\xi \sup_{\substack{\|\mathbf{a}\|_2 \leq K_1 \\ \|\mathbf{v}\|_2 \leq K_2}} \sum_{i=1}^n \xi_i \frac{1}{M} \sum_{k=1}^N a_k \sum_{m=1}^M B_k(w_m^\top x_i) v_m \\ &= \frac{1}{M} \mathbb{E}_\xi \sup_{\substack{\|\mathbf{a}\|_2 \leq K_1 \\ \|\mathbf{v}\|_2 \leq K_2}} \sum_{k=1}^N a_k \sum_{m=1}^M \left(\sum_{i=1}^n \xi_i B_k(w_m^\top x_i) \right) v_m \\ &= \frac{1}{M} \mathbb{E}_\xi \sup_{\substack{\|\mathbf{a}\|_2 \leq K_1 \\ \|\mathbf{v}\|_2 \leq K_2}} \mathbf{a}^\top \left(\sum_{i=1}^n \xi_i \mathbf{B}_i \right) \mathbf{v}, \end{aligned}$$

Let $\|\cdot\|$ be the operator norm of a matrix, namely the largest singular value of a matrix. Then by the equivalent definition of the operator norm, we have that

$$\frac{1}{M} \mathbb{E}_\xi \sup_{\substack{\|\mathbf{a}\|_2 \leq K_1 \\ \|\mathbf{v}\|_2 \leq K_2}} \mathbf{a}^\top \left(\sum_{i=1}^n \xi_i \mathbf{B}_i \right) \mathbf{v} = \frac{K_1 K_2}{M} \mathbb{E}_\xi \left\| \sum_{i=1}^n \xi_i \mathbf{B}_i \right\|.$$

Furthermore, we have that for any matrix $\mathbf{A}$, it holds that

$$\|\mathbf{A}\| \leq \|\mathbf{A}\|_{\text{Fr}} = \sqrt{\text{Tr}(\mathbf{A}\mathbf{A}^\top)}.$$

Plugging it into the former expression, we have

$$\begin{aligned} \frac{K_1 K_2}{M} \mathbb{E}_\xi \left\| \sum_{i=1}^n \xi_i \mathbf{B}_i \right\| &\leq \frac{K_1 K_2}{M} \mathbb{E}_\xi \left\| \sum_{i=1}^n \xi_i \mathbf{B}_i \right\|_{\text{Fr}} \\ &\leq \frac{K_1 K_2}{M} \sqrt{\mathbb{E}_\xi \left\| \sum_{i=1}^n \xi_i \mathbf{B}_i \right\|_{\text{Fr}}^2} \\ &= \frac{K_1 K_2}{M} \sqrt{\mathbb{E}_\xi \text{Tr} \left(\sum_{i=1}^n \xi_i \mathbf{B}_i \right) \left(\sum_{i=1}^n \xi_i \mathbf{B}_i \right)^\top} \\ &= \frac{K_1 K_2}{M} \sqrt{\text{Tr} \mathbb{E}_\xi \left(\sum_{i=1}^n \xi_i^2 \mathbf{B}_i \mathbf{B}_i^\top + \sum_{\substack{i \neq j \\ i, j \in [n]}} \xi_i \xi_j \mathbf{B}_i \mathbf{B}_j^\top \right)} \\ &= \frac{K_1 K_2}{M} \sqrt{\text{Tr} \left(\sum_{i=1}^n \mathbf{B}_i \mathbf{B}_i^\top \right)} \\ &= \frac{K_1 K_2}{M} \sqrt{\sum_{i=1}^n \|\mathbf{B}_i\|_{\text{Fr}}^2} \\ &\leq \frac{K_1 K_2}{M} \sqrt{nNM} = \frac{7\|\sigma\|_\infty|\mathcal{K}|R\sqrt{n\log(2/\delta)}}{h\sqrt{2\pi}}. \end{aligned}$$

Finally, we conclude that

$$\begin{aligned}
\mathcal{R}(\ell \circ f_{\mathcal{V}} \circ S) &= \frac{1}{n} \mathbb{E}_{\boldsymbol{\xi}} \sup_{t \in T} \sum_{i=1}^n \xi_i \ell(f(x_i), y_i) \\
&\leq \frac{\rho}{n} \mathbb{E}_{\boldsymbol{\xi}} \sup_{t \in T} \sum_{i=1}^n \xi_i f(x_i) \\
&\leq \frac{7\rho \|\sigma\|_{\infty} |\mathcal{K}| R \sqrt{\log(2/\delta)/2\pi}}{h\sqrt{n}}.
\end{aligned}$$

□

E FURTHER DETAILS ON EXPERIMENTS

E.1 DATASETS

Benchmark Datasets. MNIST and CIFAR-10 are loaded using `torchvision.datasets` in Python. The UCI datasets are downloaded from the urls listed in Table 6.

Table 6: Urls for downloading the UCI datasets.

DATA SET	URLS
ADULT	HTTPS://ARCHIVE.ICS.UCI.EDU/ML/MACHINE-LEARNING-DATABASES/ADULT/ADULT.DATA
PROTEIN	HTTPS://ARCHIVE.ICS.UCI.EDU/ML/MACHINE-LEARNING-DATABASES/00265/CASP.CSV
CT	HTTPS://ARCHIVE.ICS.UCI.EDU/ML/MACHINE-LEARNING-DATABASES/00206/SLICE_LOCALIZATION_DATA.ZIP
WORKLOADS	HTTPS://ARCHIVE.ICS.UCI.EDU/ML/MACHINE-LEARNING-DATABASES/00493/DATASETS.ZIP
MILLIONSONGS	HTTPS://ARCHIVE.ICS.UCI.EDU/ML/MACHINE-LEARNING-DATABASES/00203/YEARPREDICTIONMSD.TXT.ZIP

Synthetic Datasets. We choose target functions to be of the form

$$f(x) = \mathbb{E}_{w \sim \mathcal{N}(0, I_d)} [\sigma(w^\top x) v(w)]$$

where $x \in \mathbb{R}^d$ and $d = 2$. We set f_1, f_2, f_3 with the corresponding $\sigma_1, \sigma_2, \sigma_3$ as

$$\begin{aligned} \sigma_1(x) &= \sin(\pi x) \mathbf{1}_{[-1, 1]}, & \sigma_2(x) &= \sin(\pi x) \mathbf{1}_{[0, 1]}, \\ \sigma_3(x) &= -\sin(\pi(x + 0.5)) \mathbf{1}_{[-1.5, -0.5]} + \sin(\pi(x - 0.5)) \mathbf{1}_{[0.5, 1.5]}, \end{aligned}$$

and $v_i(w) = c_i \max\{b_1^\top w, b_2^\top w\}$, $i \in [3]$, where b_1, b_2 are two fixed vectors, and c_i are set as to ensure that $\mathbb{E}_x |f_i(x)| \approx 1$. To create the synthetic datasets, we sampled 10^5 values of w and using the empirical average $\sum_{m=1}^{10^5} \sigma_i(w_m^\top x) v_i(w_m) / 10^5$ to approximate $f_i(x)$, so that the approximation error is around $C * 10^{-3}$. We sampled $\{x_i\}_{i \in [n]} \stackrel{i.i.d.}{\sim} \mathcal{N}(0, I_d)$ for sample size $n = 15000$ and $d = 2$.

We show all the learned activation functions in RFLAF in Figure 3. We can see that the learned activation function is close to the ground-truth function with a level of approximation error.

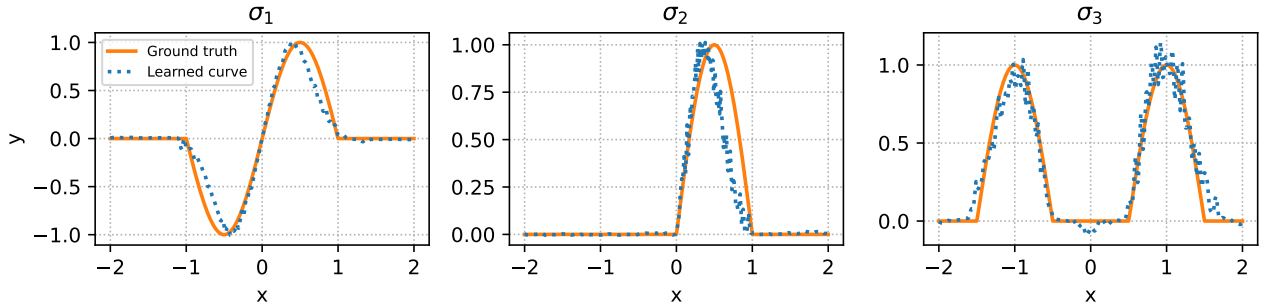


Figure 3: The Activation Functions Learned in RFLAF.

E.2 MORE EXPERIMENTAL RESULTS

In Table 7, we show the test accuracies on the three classification tasks as supplementary results for Table 1. RFLAFs show consistently advantages over RFMLPs.

Table 7: Test Accuracies (%) of Random Feature Models. $N = 16$ for RFLAFs.
Results are reported as mean $\pm$ std. Best in bold. Second best in italics.

DATA SET	RFLAF			RFMLP			
	RBF	BS	PL	RELU	COS	TANH	SIGMOID
MNIST	<i>96.10$\pm$0.30</i>	94.95 $\pm$ 0.63	96.26$\pm$0.13	95.40 $\pm$ 0.09	65.98 $\pm$ 4.43	92.30 $\pm$ 0.06	86.86 $\pm$ 0.56
CIFAR-10	48.95 $\pm$ 0.36	44.29 $\pm$ 0.58	49.20$\pm$0.42	<i>48.99$\pm$0.23</i>	18.44 $\pm$ 1.52	39.62 $\pm$ 0.23	34.12 $\pm$ 1.60
ADULT	85.66$\pm$0.15	85.05 $\pm$ 0.18	-	<i>85.63$\pm$0.11</i>	83.85 $\pm$ 0.74	85.05 $\pm$ 0.15	84.90 $\pm$ 0.31

In Table 8, we show the test accuracies on the three classification tasks as supplementary results for Table 4. LANs show better performances in the first two tasks and similar performance in the third task compared to MLP and KAN.

Table 8: Test Accuracies (%) of Regular Two-layer Networks. $N = 16$ for LAN and KAN.

DATA SET	LOSS				
	LAN			MLP	KAN
	RBF	BS	PL	RELU	
MNIST	<i>96.40$\pm$1.15</i>	96.85$\pm$0.15	95.10 $\pm$ 0.16	95.99 $\pm$ 0.07	95.19 $\pm$ 0.14
CIFAR-10	<i>52.74$\pm$0.36</i>	53.18$\pm$0.61	48.13 $\pm$ 3.91	49.61 $\pm$ 0.17	47.90 $\pm$ 0.18
ADULT	85.77 $\pm$ 0.08	85.76 $\pm$ 0.30	83.15 $\pm$ 0.50	85.85$\pm$0.08	<i>85.80$\pm$0.07</i>

In the following, we additionally summarize the results on RFLAFs of $N = 8, 16, 32, 64, 128$ and low-degree polynomials, including the test losses (Table 9), training time (Table 10) and testing time (Table 11).

Table 9: Test Losses for All N . (Best in Bold.)

MODEL-N	MNIST	CIFAR-10	ADULT	PROTEIN	CT	WORKLOADS	MILLIONSONGS
RELU	0.159	1.466	0.311	0.241	0.356	2.771	0.951
COS	1.390	2.641	0.363	0.371	0.589	2.634	0.280
TANH	0.277	1.769	0.324	0.650	1.241	24.997	8.434
SIGMOID	0.498	1.930	0.327	0.280	0.692	1.707	0.118
RBF-8	0.204	1.440	0.307	0.230	0.202	0.736	0.103
BS-8	0.146	1.490	0.312	0.198	0.259	0.476	0.129
PL-8	0.129	1.470	0.322	INF	INF	0.827	INF
RBF-16	0.126	1.450	0.309	0.204	0.212	0.465	0.102
BS-16	0.165	1.609	0.324	0.194	0.302	0.546	0.120
PL-16	0.124	1.482	INF	INF	INF	INF	INF
RBF-32	0.137	1.462	0.314	0.184	0.201	0.336	0.103
BS-32	0.152	1.493	0.331	0.196	0.283	0.548	0.135
PL-32	0.132	3.493	INF	INF	INF	INF	INF
RBF-64	0.142	1.480	0.316	0.180	0.200	0.302	0.102
BS-64	0.153	1.498	0.349	0.225	0.356	0.616	0.215
PL-64	10.599	INF	INF	INF	INF	INF	INF
RBF-128	0.145	1.471	0.315	0.183	0.185	0.295	0.100
BS-128	0.161	1.537	0.392	0.286	0.419	0.791	0.377
PL-128	INF	INF	INF	INF	INF	INF	INF
PL-2	0.189	1.484	0.313	0.266	4.007	1.292	0.119
PL-4	0.137	1.462	0.313	0.330	INF	0.857	0.405
PL-6	0.144	1.470	0.312	26.614	INF	0.810	INF

Table 10: Train Time for All N .

MODEL-N	MNIST	CIFAR-10	ADULT	PROTEIN	CT	WORKLOADS	MILLIONSONGS
RELU	1.000	1.000	1.000	1.000	1.000	1.000	1.000
COS	0.999	1.000	0.991	1.046	1.049	1.017	1.135
TANH	0.994	1.001	0.986	1.040	1.018	1.028	1.147
SIGMOID	0.993	0.996	0.943	0.974	0.993	0.995	1.166
RBF-8	1.065	1.041	1.525	1.499	1.399	1.065	1.691
BS-8	1.267	1.226	2.886	3.119	2.647	2.280	3.426
PL-8	1.059	1.043	1.388	1.415	1.259	1.007	1.589
RBF-16	1.099	1.063	1.683	1.756	1.618	1.269	1.974
BS-16	1.452	1.367	4.236	4.732	3.864	3.372	5.139
PL-16	1.078	1.057	1.431	1.410	1.319	1.005	1.669
RBF-32	1.153	1.126	2.380	2.496	2.197	1.782	2.746
BS-32	1.812	1.699	7.134	8.122	6.471	5.773	8.650
PL-32	1.090	1.082	1.819	1.796	1.578	1.263	1.993
RBF-64	1.261	1.341	4.303	4.709	3.948	3.366	5.122
BS-64	2.352	2.310	12.927	14.683	11.867	10.549	15.883
PL-64	1.046	1.173	2.797	2.989	2.578	2.148	3.312
RBF-128	1.520	1.738	7.878	8.716	7.152	6.325	9.563
BS-128	3.742	3.686	25.557	28.601	22.978	20.695	31.085
PL-128	1.151	1.412	5.008	5.467	4.538	3.955	5.997
PL-2	1.085	1.040	1.336	1.395	1.236	0.963	1.523
PL-4	1.077	1.049	1.355	1.355	1.276	0.977	1.530
PL-6	1.064	1.044	1.366	1.373	1.278	0.995	1.527

Table 11: Test Time for All N .

MODEL-N	MNIST	CIFAR-10	ADULT	PROTEIN	CT	WORKLOADS	MILLIONSONGS
RELU	1.000	1.000	1.000	1.000	1.000	1.000	1.000
COS	1.002	1.000	1.029	1.022	1.053	0.996	1.179
TANH	0.998	1.001	0.996	1.028	1.006	1.026	1.218
SIGMOID	0.992	0.997	0.931	0.903	0.965	0.988	1.241
RBF-8	1.022	0.997	1.474	1.337	1.214	0.924	1.630
BS-8	1.242	1.193	3.730	3.707	2.730	2.683	4.330
PL-8	1.018	1.010	1.312	1.191	1.020	0.814	1.396
RBF-16	1.060	1.042	2.103	1.971	1.613	1.381	2.387
BS-16	1.438	1.360	6.197	6.292	4.400	4.425	7.324
PL-16	1.043	1.037	1.677	1.486	1.277	1.041	1.874
RBF-32	1.137	1.154	3.664	3.523	2.683	2.503	4.182
BS-32	1.843	1.750	11.612	11.918	8.112	8.439	13.763
PL-32	1.080	1.108	2.722	2.508	1.888	1.732	2.924
RBF-64	1.290	1.375	6.785	6.736	4.845	4.802	7.969
BS-64	2.483	2.390	21.305	21.835	15.033	15.594	25.539
PL-64	1.091	1.206	4.415	4.285	3.173	3.075	5.124
RBF-128	1.584	1.776	12.730	12.679	8.916	9.170	15.144
BS-128	3.983	3.818	42.399	42.576	29.200	30.676	50.142
PL-128	1.200	1.444	8.028	7.884	5.604	5.691	9.398
PL-2	1.041	0.994	1.068	0.991	0.920	0.673	1.216
PL-4	1.035	1.007	1.161	1.035	0.994	0.713	1.212
PL-6	1.032	1.001	1.220	1.122	0.981	0.750	1.273

E.3 DISCUSSIONS OVER THE REGULAR TWO-LAYER NETWORKS

There are a few points that we would like to supplement.

(1) Implementation details of KAN.

We do not adopt the `pykan` package developed by the author of KAN due to its inefficiency. KAN was initially proposed to address problems typical of science-related tasks, which are generally smaller in scale than typical machine learning tasks. The author states that they will continue developing the repository primarily for scientific discovery and computing, but without significant updates for efficiency¹. Furthermore, we tested the model on the baseline MNIST dataset and found that the original KAN struggles to scale as the width M increases due to its inefficiency.

To address this issue, we adopt the `efficient-KAN` package², which enables the model to be sufficiently fast and scalable for comparisons. We use the default KAN configuration in this package, employing B-splines of degree three. To ensure consistency, all models are trained using the Adam optimizer with identical parameter settings.

(2) Why do LAN perform better than MLP and KAN?

The fact that LAN performs better than MLP is natural because the structure of learnable activation functions enables the two-layer networks to represent a broader class of functions. The improvement over MLP is not free though, but in the cost of additional computation for combining the basis functions. However, as shown in Table 4, LAN of $N = 16$ runs within two times of the running time of MLP, which is acceptable in practice. And we highlight that this is probably not the speed limit of LAN as more techniques over codes may be applied to speed up the training and computation process of LAN (e.g., parallel computing of the N basis functions), just as how KAN has evolves. We do not contribute to the code improvements of LAN, since our study does not focus on LAN but on RFLAF, but we believe that this is probably another interesting future direction.

For the other method, KAN fails to consistently outperform even MLP in the results. This is probably due to the convergence problem of KAN. To control the variables, we apply Adam for all models as the optimizer and train all models with the same epoch number. The author of KAN proposed to use LBFGS to boost convergence, but we do not know how sensitive the model is to the use of the optimizer. But for LAN, the common optimizer Adam works well. The difficulty of optimization in KAN probably comes from the high degree of freedom induced by the extensive number of learnable activation functions. While MLP and LAN (with RBF or BS) successfully achieve very low test errors, KAN seems to converge very slowly and be underfitting in some tasks. In contrast, the two-layer LAN only contains one learnable activation function of N extra learnable parameters compared to MLP. Hence, training LAN is almost as easy as training MLP (at least for the case of two layer), which indicates that LAN is probably more capable of scaling up in typical machine learning problems than KAN.

¹See <https://github.com/KindXiaoming/pykan?tab=readme-ov-file>

²See <https://github.com/Blealtan/efficient-kan>