On Risk of Minimum-Norm Interpolants and Restricted Lower Isometry of Kernels

Tengyuan Liang
University of Chicago

Alexander Rakhlin
MIT

Xiyu Zhai
MIT

August 26, 2019

Abstract

We study the risk of minimum-norm interpolants of data in a Reproducing Kernel Hilbert Space where kernel defined as a function of an inner product. Our upper bounds on the risk are of a multiple-descent shape for the various scalings of $d = n^{\alpha}$, $\alpha \in (0, 1)$. At the heart of our analysis is a study of spectral properties of the random kernel matrix restricted to a filtration of eigen-spaces of the population covariance operator.

1 Introduction

We investigate the generalization and consistency of minimum-norm interpolants

$$\begin{aligned} \widehat{f} \in \ & \underset{f\in\mathcal{H}}{\operatorname{argmin}} \quad \|f\|_{\mathcal{H}} \\ & \text{s.t.} \quad f(x_i) = y_i, \quad i = 1, \ldots, n \end{aligned} \tag{1}$$

of the data $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$ with respect to a norm in a Reproducing Kernel Hilbert Space $\mathcal{H}$. The interpolant, also termed "Kernel Ridgeless Regression," can be viewed as a limiting solution of

$$\underset{f\in\mathcal{H}}{\operatorname{argmin}} \ \frac{1}{n}\sum_{i=1}^{n}(f(x_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2 \tag{2}$$

as $\lambda \to 0$. Classical statistical analyses of Kernel Ridge Regression (see Caponnetto and De Vito (2007) and references therein) rely on a carefully chosen regularization parameter to control the bias-variance tradeoff, and the question of consistency of the non-regularized solution falls outside the scope of these classical results.

Recent literature has focused on understanding risk of estimators that interpolate data, including work on nonparametric local rules (Belkin et al., 2018b,d), high-dimensional linear regression (Bartlett et al., 2019; Hastie et al., 2019), and kernel (ridgeless) regression (Belkin et al., 2018c; Liang and Rakhlin, 2018; Rakhlin and Zhai, 2018).

This paper continues the line of work on kernel regression. More precisely, Rakhlin and Zhai (2018) showed that the minimum-norm interpolant with respect to the Laplace kernel is *not* consistent (that is, risk does not go to zero with $n \to \infty$) if dimensionality d of the data is constant with respect to n, even if the bandwidth of the kernel is chosen adaptively. On the other hand, Liang and Rakhlin (2018) investigated the regime $n \asymp d$ and showed that risk can be upper bounded by a quantity that can be small under favorable spectral properties of the data and the kernel matrix. The present paper aims to paint a more comprehensive picture, studying the performance of the minimum-norm interpolants in a general scaling regime $d \asymp n^{\alpha}$, $\alpha \in (0, 1)$.

It is not hard to see that spectral properties of the kernel are key for analyzing the variance and the bias of the minimum-norm interpolant. Note that the eigenvalues of the empirical kernel matrix have one-to-one correspondance to that of the empirical covariance operator. We prove that on a filtration of eigen-spaces of the covariance operator defined by the population distribution, the empirical covariance operator satisfies a certain *restricted lower isometry* property. This spectral analysis is the main technical part of this paper.

Figure 1 summarizes the behavior of our upper bound on the risk of the minimum-norm interpolant. We make two observations. First, for any integer $k \geq 1$, for $\alpha \in [\frac{1}{k+1}, \frac{1}{k})$, there exists a "valley" on the curve at each $d = n^{\frac{1}{k+1/2}}$ where the rate is fast (of the order $n^{-\beta}$ with $\beta = \frac{1}{2k+1}$). Second, towards the lower dimensional regime (α moving towards 0), the fastest possible rate even at the bottom of the valley is getting worse. This result complements the double-descent behavior on the risk of interpolated solutions investigated previously in the literature (Belkin et al., 2018a). Since the lower bound matching our upper bound is still not available (and this is a challenging open problem), we do not know whether the multiple-descent behavior is indeed a generalization of the double-descent curve for kernel regression.

The main results of the paper can be stated as follows, informally.

Theorem 1 (Informal)**.** For any integer $\iota \geq 1$, consider $d = n^{\alpha}$ where $\alpha \in [\frac{1}{\iota+1}, \frac{1}{\iota})$. Consider a general function $h \in C^{\infty}$ and define the inner product kernel $k(x, z) = h(x^{\top} z/d)$. Consider n *i.i.d.* data pairs (x_i, y_i) drawn from $\mathcal{P}_{X,Y}$, and denote the target function $f_*(x) = \mathbb{E}[Y|X = x]$. Suppose the conditions on f_*, h and $\mathcal{P}_{X,Y}$ specified by Theorems 2-3 are satisfied. With probability at least $1 - \delta - e^{-n/d^{\iota}}$ on the design $\mathbf{X} \in \mathbb{R}^{n \times d}$,

$$
\begin{aligned}
\mathbb{E}\left[\|\widehat{f} - f_*\|^2_{\mathcal{P}_X} | \mathbf{X}\right] &\leq C \cdot \left(\frac{d^{\iota}}{n} + \frac{n}{d^{\iota+1}}\right) \asymp n^{-\beta}, \\
\beta &:= \min\left\{(\iota+1)\alpha - 1, 1 - \iota\alpha\right\}.
\end{aligned}
$$

Here the constant $C(\delta, K, \iota, \mathcal{P})$ does not depend on d, n.

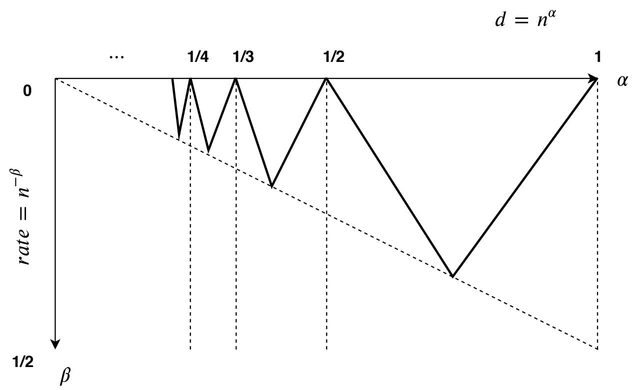

Figure 1: **Multiple-descent behavior** of the rates as the scaling $d = n^\alpha$ changes.

2 Formulation

Data Assumption. Consider i.i.d random vectors $x_1, \cdots, x_n \in \mathbb{R}^d$ drawn from $\mathcal{P}_X^{\otimes d}$, where the distribution for each coordinate $\mathcal{P}_X$ satisfies:

1. $\mathcal{P}_X$ has the following tail property:

$$\forall t > 0, \mathbb{P}(|X| \geq t) \leq C(1+t)^{-\nu} \tag{3}$$

   where ν is a positive constant.

2. for any set S of finitely many real numbers, $\mathbb{P}(X \in S) < 1$.

In addition, we assume that $\forall x \in \mathcal{X}$, the conditional variance is bounded with a constant,

$$\mathrm{Var}[Y|X = x] \leq M. \tag{4}$$

Definition of Kernel. Consider the function $f \in \mathcal{C}^\infty$ whose Taylor expansion converges for all $x \in \mathbb{R}$

$$f(x) = \sum_{i=0}^{\infty} \alpha_i x^i \tag{5}$$

with all coefficients $\alpha_i \geq 0$. We define a kernel function $k(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ induced by f

$$k(x, z) := f\left(\frac{x^\top z}{d}\right). \tag{6}$$

Similarly, we define the normalized finite dimensional kernel matrix $\mathbf{K} \in \mathbb{R}^{n \times n}$,

$$\mathbf{K}_{ij} := k(x_i, x_j)/n. \tag{7}$$

In other words, $\mathbf{K} = k(\mathbf{X}, \mathbf{X})/n$ with the $1/n$ normalization.

Denote the truncated version of the function as

$$f^{[\leq \iota]}(x) = \sum_{\iota=0}^{\iota} \alpha_\iota x^\iota, \tag{8}$$

and the corresponding truncated kernel matrix $\mathbf{K}^{[\iota]}$ (used only in the proof) as

$$\mathbf{K}_{ij}^{[\leq \iota]} := f^{[\leq \iota]}(x_i^\top x_j/d)/n. \tag{9}$$

Similarly, we define for convenience

$$f^{[\iota]}(x) = \alpha_\iota x^\iota \ , \tag{10}$$

and correspondingly

$$\mathbf{K}_{ij}^{[\iota]} := f^{[\iota]}(x_i^\top x_j/d)/n \ . \tag{11}$$

Regime. We are interested in the following high dimensional regime: there is a fixed positive integer ι and we assume that

$$\iota < \frac{\log n}{\log d(n)} < \iota + 1. \tag{12}$$

Our investigation focuses on dimension $d(n)$ growing with the sample size n and n being sufficiently large.

Notation. We use the notation "r_i';" to represent a sequence of indices $r_1 \cdots r_d$. For example, we can use this to abbreviate a monomial with order $r_1 \cdots r_d$: for monomial p (recall x is a d-dimensional vector),

$$p_{r_i;}(x) := p_{r_1 \cdots r_d}(x) = \prod_{i=1}^{d} (x[i])^{r_i} \tag{13}$$

where $x[i]$ denotes the i-th coordinate of x. This notation is also used in tensors, for example:

$$T_{r_i;} := T_{r_1 \cdots r_d}. \tag{14}$$

Restricted Lower Isometry for High Dimensional Kernel In Section 3, we prove the *Restricted Lower Isometry Property* for the high dimensional kernel matrix of interest. This property proves crucial in bounding the generalization error for the kernel ridgeless regression and the wide neural networks.

Proposition 1 (Restricted Lower Isometry Property for Kernel)**.** Assume that the first $\iota_0 + 1$ th Taylor coefficients $\alpha_0, \cdots, \alpha_{\iota_0}$ are positive and $d^{\iota_0} \log d = o(n)$.

Then there are positive constants C, C' depending only on ι_0 and $\mathcal{P}_X$ and $\alpha_0, \cdots, \alpha_{\iota_0}$ such that for n large enough, with probability at least $1 - e^{-\Omega(n/d^{\iota_0})}$ the following holds: for any $\iota \leq \iota_0$, $\mathbf{K}^{[\leq \iota]}$ has $\binom{\iota+d}{\iota}$ nonzero eigenvalues, all of them larger than $C' d^{-\iota}$, and the range of $\mathbf{K}^{[\leq \iota]}$ is

$$\{(p(x_1), \cdots, p(x_n)) : p \text{ is a multivariable polynomial of degree not larger than } \iota\}.$$

We note that concurrent work of Ghorbani et al. (2019) implies a similar control on the least eigenvalue under somewhat different assumptions on the underlying data.

3 Proof of Main Proposition

The proof aims to establish the restricted lower isometry behavior for the empirical kernel K when restricting to the population eigenspace of rank $\binom{i+d}{i}$ (sorted according to the eigenvalues). We show a lower bound for the restricted lower isometry, as multiplicatively equivalent to the population eigenvalues. The approach proceeds along the lines of (Koltchinskii and Mendelson, 2015; Mendelson, 2014). One technical contribution is establishing the "small-ball" property for the polynomial basis of the kernel.

3.1 Preparation

First we fix an index $\iota \leq \iota_0$. After we prove it for ι, the conclusion shall follow easily from a union bound over the case $\iota = 1, \cdots, \iota_0$.

Consider the Taylor expansion, with the multi-index $r_1, \cdots, r_d$

$$n\mathbf{K}_{ij}^{[\leq \iota]} = \sum_{r_1, \cdots, r_d \geq 0}^{\iota} c_{r_1 \cdots r_d} \alpha_{r_1 + \cdots + r_d} p_{r_1 \cdots r_d}(x_i) p_{r_1 \cdots r_d}(x_j) / d^{r_1 + \cdots + r_d} \tag{15}$$

where

$$c_{r_1 \cdots r_d} = \frac{(r_1 + \cdots + r_d)!}{r_1! \cdots r_d!}, p_{r_1 \cdots r_d}(x_i) = (x_i[1])^{r_1} \cdots (x_i[d])^{r_d} \quad . \tag{16}$$

Fix an ordering of all $(r_1, \cdots, r_d)$ such that $r_i \geq 0, \sum r_i \leq \iota$, and let

$$(r_1 \cdots r_d)_\iota \in \left\{1, 2, \cdots, \binom{\iota+d}{\iota}\right\}$$

denote the index of $(r_1, \cdots, r_d)$ in the ordering.

Define a $n \times \binom{\iota+d}{\iota}$ matrix Φ as

$$\Phi_{i,(r_1 \cdots r_d)_\iota} = \sqrt{c_{r_1 \cdots r_d} \alpha_{r_1 + \cdots + r_d}} p_{r_1 \cdots r_d}(x_i) / d^{(r_1 + \cdots + r_n)/2}, \tag{17}$$

then

$$\mathbf{K}^{[\leq \iota]} = \frac{1}{n}\Phi\Phi^\top \quad , \tag{18}$$

which has the same nonzero spectrum as the covariance

$$\Theta = \frac{1}{n}\Phi^\top\Phi. \tag{19}$$

We have

$$\Theta_{(r_1\cdots r_d)_\iota,(r_1'\cdots r_d')_\iota} = \frac{1}{n}\sum_{i=1}^{n}\sqrt{c_{r_1\cdots r_d}\alpha_{r_1+\cdots+r_d}}\sqrt{c_{r_1'\cdots r_d'}a_{r_1'+\cdots+r_d'}}\frac{p_{r_1\cdots r_d}(x_i)p_{r_1'\cdots r_d'}(x_i)}{d^{(r_1+\cdots+r_n+r_1'+\cdots+r_n')/2}} \quad . \tag{20}$$

It would be hard to directly work with Φ to analyze the eigenvalues of the random matrix because of complicated correlation structure in the entries. Instead, we define another matrix Ψ such that

1. $\Psi^\top\Psi$ is easier to analyse from the probabilistic point of view;
2. there is a linear transformation Λ with $\|\Lambda\|_{\ell_2\to\ell_2}, \|\Lambda^{-1}\|_{\ell_2\to\ell_2}$ bounded such that

$$\Phi = \Psi\Lambda. \tag{21}$$

With such Λ, we have

$$\Theta = \Lambda^\top\Psi^\top\Psi\Lambda. \tag{22}$$

Such Λ can be obtained through the Gram-Schimdt process on polynomial basis, which we will elaborate on next.

3.2 Gram-Schimdt Process

We now proceed in the following steps:

Step 1. Define Ψ

Definition 1. Given a distribution $\mathcal{P}$ over $\mathbb{R}$, define $q_0, q_1, \cdots, q_k, \cdots$, to be the sequence of polynomials obtained by the Gram-Schmidt process on the basis $\{1, x, \cdots, x^k, \cdots\}$ w.r.t. the inner product of space $L^2(\mathcal{P})$. Define the $n \times \binom{\iota+d}{\iota}$ matrix Ψ as

$$\Psi_{i,(r_1\cdots r_d)_\iota} = \sqrt{c_{r_1\cdots r_d}\alpha_{r_1+\cdots+r_d}}q_{r_1\cdots r_d}(x_i)/d^{(r_1+\cdots+r_d)/2} \quad . \tag{23}$$

To be concrete, one can see that

$$q_0 = 1, q_1 = (x - m_1)/\sqrt{(m_2 - m_1^2)}, \cdots, \tag{24}$$

where m_i is the ith moment of $\mathcal{P}$, and for $i < j$

$$\mathbb{E}_{X\sim\mathcal{P}}X^i q_j(X) = 0. \tag{25}$$

The following lemmas on properties of the polynomial basis will be useful.

Lemma 1. Suppose that $f(x) = \sum_{r_i;} u_{r_i;} q_{r_i;}(x)$, then

$$u_{r_i;} = \mathbb{E}_{x\sim\mathcal{P}^d} f(x) q_{r_i;}(x) \tag{26}$$

where $q_{r_i;}(x) := \prod_{i=1}^d q_{r_i}(x[i])$.

Lemma 2. $\mathbb{E}_{x\sim\mathcal{P}^d} q_{r_i;}(x)p(x) = 0$ if there exists one i such that the degree of $p(x)$ as a polynomial of x_i is less than r_i.

Lemma 3 (Triangle condition)**.** $\mathbb{E}^d_{x\in\mathcal{P}} q_{r_i;}(x) q_{r'_i;}(x) q_{r''_i;}(x) = 0$ if there exists one i, such that $r_i > r'_i + r''_i$.

Step 2. Existence of Λ.

Lemma 4. There is a $\binom{\iota+d}{\iota} \times \binom{\iota+d}{\iota}$ invertible matrix Λ such that

$$\Phi = \Psi\Lambda. \tag{27}$$

Proof of Lemma 4. For any $a \in \mathbb{R}^{\binom{\iota+d}{\iota}}$, there is a unique $b \in \mathbb{R}^{\binom{\iota+d}{\iota}}$ such that

$$\begin{aligned}
&\sum_{r_1+\cdots+r_d\le\iota, r_i\ge 0} a_{(r_1\cdots r_d)_\iota} \sqrt{c_{r_1\cdots r_d}\alpha_{r_1+\cdots+r_d}} p_{r_1\cdots r_d}/d^{(r_1+\cdots+r_d)/2} \\
=&\sum_{r_1+\cdots+r_d\le\iota, r_i\ge 0} b_{(r_1\cdots r_d)_\iota} \sqrt{c_{r_1\cdots r_d}\alpha_{r_1+\cdots+r_d}} q_{r_1\cdots r_d}/d^{(r_1+\cdots+r_d)/2}.
\end{aligned} \tag{28}$$

As a result

$$\Phi a = \Psi b. \tag{29}$$

Choose Λ to be the linear mapping that maps a to b

$$\Phi a = \Psi\Lambda a. \tag{30}$$

This holds for any a, so we have

$$\Phi = \Psi\Lambda. \tag{31}$$

□

Step 3. Boundedness of Λ and Λ^{-1}. For a vector $v \in \mathbb{R}^{\binom{\iota+d}{\iota}}$, let $v_{\ge\iota'}$ be the vector such that

$$(v_{\ge\iota'})_{(r_1\cdots r_d)_\iota} = v_{(r_1\cdots r_d)_\iota} 1_{\{r_1+\cdots+r_d\ge\iota'\}}. \tag{32}$$

Define similarly $v_{\iota'}, v_{>\iota'}, \Lambda_{\ge\iota'}, \Lambda_{\ge\iota',<\iota'}$, etc.

Lemma 5. There is a constant $C(\iota)$ independent of d such that

$$\|\Lambda\|_{\ell_2\to\ell_2}, \|\Lambda^{-1}\|_{\ell_2\to\ell_2} \le C(\iota) \tag{33}$$

Proof of Lemma 5. We start with few claims about the Gram-Schimidt process and the structure of Λ.

Claim 1: $\Lambda_{\geq \iota', <\iota'} = 0$ for any k. Or alternatively, if $b = \Lambda a$ then

$$b_{\geq \iota'} = \Lambda_{\geq \iota'} a_{\geq \iota'}. \tag{34}$$

Proof of Claim 1. We need only to show that if $a_{\geq \iota'} = 0$, then $b_{\geq \iota'} = 0$. Observe that $a_{\geq \iota'} = 0$ means that the left hand of equation (28) is of degree less than ι'. Since this is an equality, the right hand side must be of degree less than ι'. Note that this implies that $b_{\geq \iota'} = 0$. □

Claim 2: $\Lambda_{\iota', \iota'}$ is diagonal with

$$c \leq \lambda_{\min}(\Lambda_{\iota', \iota'}) \ \leq \lambda_{\max}(\Lambda_{\iota', \iota'}) \leq C \tag{35}$$

where c, C depends only on $\iota', \mathcal{P}$

Proof of Claim 2. Given r_i; and r_i'; with $\sum_i r_i = \sum_i r_i' = \iota'$, we have

$$\Lambda_{(r_i;)_\iota, (r_i';)_\iota} = \mathbb{E}_{x \sim \mathcal{P}^d} q_{r_i;}(x) p_{r_i';}(x) \tag{36}$$

If $r_i; \neq r_i';$, there is at least one i such that $r_i > r_i'$, then according to Lemma 2, we have

$$\Lambda_{(r_i;)_\iota, (r_i';)_\iota} = 0 \tag{37}$$

Then $\Lambda_{\iota', \iota'}$ is diagonal. Now we have

$$\begin{aligned} \Lambda_{(r_i;)_\iota, (r_i;)_\iota} &= \mathbb{E}_{x \sim \mathcal{P}^d} q_{r_i;}(x) p_{r_i;}(x) \\ &= \prod_{i=1}^{d} \left(\mathbb{E}_{X \sim \mathcal{P}} q_{r_i}(X) p_{r_i}(X) \right) \end{aligned} \tag{38}$$

Note that $q_0 \equiv 1$. Since the set $I := \{1 \leq i \leq d : r_i \text{ is nonzero}\}$ is of size at most ι, $\Lambda_{(r_i;)_\iota, (r_i;)_\iota}$ is uniformly bounded by the following constant (depending on ι and $\mathcal{P}$):

$$\left(\max_{r_i} \mathbb{E}_{X \sim \mathcal{P}} q_{r_i}(X) p_{r_i}(X) \right)^\iota \ . \tag{39}$$

□

Claim 3: Let $b_{\iota'} = \Lambda_{\iota' \iota'} a_{\iota'} + \Lambda_{\iota', >\iota'} a_{>\iota'}$. Then $\|\Lambda_{\iota', >\iota'}\|_{\ell_2 \to \ell_2}$ has an upper bound that only depends on ι.

Proof of Claim 3. An entry $(\Lambda_{\iota',>\iota'})_{(r_1\cdots r_d)_\iota,(r'_1\cdots r'_d)_\iota}$ for $\sum_i r_i = \iota'$ can be obtained by

$$\begin{aligned}
&(\Lambda_{\iota',>\iota'})_{(r_1\cdots r_d)_\iota,(r'_1\cdots r'_d)_\iota} \\
&= \sqrt{\frac{c_{r'_1\cdots r'_d}\alpha_{r'_1+\cdots+r'_d}}{c_{r_1\cdots r_d}\alpha_{r_1+\cdots+r_d}}} d^{(r_1+\cdots+r_d-r'_1-\cdots-r'_d)/2}\mathbb{E}_x q_{r_1\cdots r_d}(x) p_{r'_1\cdots r'_d}(x) \\
&= \sqrt{\frac{c_{r'_1\cdots r'_d}\alpha_{r'_1+\cdots+r'_d}}{c_{r_1\cdots r_d}\alpha_{r_1+\cdots+r_d}}} d^{(r_1+\cdots+r_d-r'_1-\cdots-r'_d)/2}\prod_i \left(\mathbb{E}_{X\sim\mathcal{P}} q_{r_i}(X) p_{r'_i}(X)\right) \quad .
\end{aligned}$$

Only when $\forall i, r_i \le r'_i$, the above term is nonzero and scales with d in the order of $d^{(r_1+\cdots+r_d-r'_1-\cdots-r'_d)/2}$. As a result,

$$\begin{aligned}
b^2_{(r_1\cdots r_d)_\iota} &= \left(\sum_{r'_1,\cdots,r'_d} (\Lambda_{\iota',>\iota'})_{(r_1\cdots r_d)_\iota,(r'_1\cdots r'_d)_\iota} a_{(r'_1\cdots r'_d)_\iota}\right)^2 \\
&= \left(\sum_{r_i\le r'_i,\sum_i r'_i>\iota'} (\Lambda_{\iota',>\iota'})_{(r_1\cdots r_d)_\iota,(r'_1\cdots r'_d)_\iota} a_{(r'_1\cdots r'_d)_\iota}\right)^2 \\
&\le \left(\sum_{r_i\le r'_i,\sum_i r'_i>\iota'} (\Lambda_{\iota',>\iota'})^2_{(r_1\cdots r_d)_\iota,(r'_1\cdots r'_d)_\iota}\right)\left(\sum_{r_i\le r'_i,\sum_i r'_i>\iota'} a^2_{(r'_1\cdots r'_d)_\iota}\right) \\
&\le \left(\sum_{l\ge 1}\sum_{\substack{r_i\le r'_i\\ \sum_i r'_i=\iota'+l}} (\Lambda_{\iota',>\iota'})^2_{(r_1\cdots r_d)_\iota,(r'_1\cdots r'_d)_\iota}\right)\left(\sum_{r_i\le r'_i,\sum_i r'_i>\iota'} a^2_{(r'_1\cdots r'_d)_\iota}\right) \\
&\lesssim \left(\sum_{l\ge 1}\sum_{\substack{r_i\le r'_i\\ \sum_i r'_i=\iota'+l}} (d^{-l/2})^2\right)\left(\sum_{r_i\le r'_i,\sum_i r'_i>\iota'} a^2_{(r'_1\cdots r'_d)_\iota}\right) \\
&\lesssim \left(\sum_{l\ge 1} d^l \times (d^{-l/2})^2\right)\left(\sum_{r_i\le r'_i,\sum_i r'_i>\iota'} a^2_{(r'_1\cdots r'_d)_\iota}\right) \\
&\lesssim \sum_{r_i\le r'_i,\sum_i r'_i>\iota'} a^2_{(r'_1\cdots r'_d)_\iota} \quad .
\end{aligned}$$

Then we have

$$\begin{aligned}
&\sum_{\sum_i r_i = \iota'} b^2_{(r_1 \cdots r_d)_\iota} \\
&\lesssim \sum_{\sum_i r_i = \iota'} \sum_{r_i \le r'_i, \sum_i r'_i > \iota'} a^2_{(r'_1 \cdots r'_d)_\iota} \\
&\lesssim \sum_{\sum_i r'_i > \iota'} a^2_{(r'_1 \cdots r'_d)_\iota}
\end{aligned}$$

where the last inequality holds because for a fixed $r'_1, \cdots, r'_d$ with $r'_1 + \cdots + r'_d = \iota$, there is at most 2^ι choices of $r_1, \cdots, r_d$ such that $\forall i, 0 \le r_i \le r'_i$. □

Assume that $\|b_{>\iota'}\| \sim \|a_{>\iota'}\|$, then since

$$\|b_{\iota'}\| \le \|\Lambda_{\iota'\iota'} a_{\iota'}\| + \|\Lambda_{\iota', >\iota'} a_{>\iota'}\| \lesssim \|a_{\ge \iota'}\|, \tag{40}$$

we have

$$\|b_{\ge\iota'}\| \le \|b_{\iota'}\| + \|b_{>\iota'}\| \lesssim \|a_{\ge\iota'}\|. \tag{41}$$

For the other direction, we have

$$\|a_{\iota'}\| = \|\Lambda^{-1}_{\iota'\iota'}(b_{\iota'} - \Lambda_{\iota', >\iota'} a_{>\iota'})\| \lesssim \|b_{\iota'} - \Lambda_{\iota', >\iota'} a_{>\iota'}\| \lesssim \|b_{\iota'}\| + \|a_{>\iota'}\| \lesssim \|b_{\iota'}\| + \|b_{>\iota'}\| \lesssim \|b_{\ge\iota'}\|. \tag{42}$$

Then

$$\|a_{\ge\iota'}\| \lesssim \|b_{\ge\iota'}\|. \tag{43}$$

Using induction on $\iota' \le \iota$ backwards concludes the proof. □

Now we can proceed to analyse $\Psi^\top \Psi$, given the boundedness of Λ.

3.3 Small Ball Property

Define the following function over $\mathbb{R}^d$:

$$f_u(x) = \sum_{r_1, \cdots, r_d \ge 0, \sum_i r_i \le \iota} u_{(r_1 \cdots r_d)_\iota} \sqrt{c_{r_1 \cdots r_d} \alpha_{r_1 + \cdots + r_d}} q_{r_1 \cdots r_d}(x) / d^{(r_1 + \cdots + r_d)/2} \quad . \tag{44}$$

In this section, we will prove that there exist θ, δ, such that for any u

$$\mathbb{P}(f_u(x)^2 > \theta \mathbb{E}[f_u(x)^2]) \ge \delta \quad . \tag{45}$$

Claim 1: $\forall u \in \mathbb{R}^{\binom{\iota+d}{\iota}}$ with $\|u\|_2 = 1$, there are constants $\delta, \epsilon > 0$ depending only on $\iota, \mathcal{P}$ such that with probability at least $1-\delta$

$$f_u(x_i)^2 > \epsilon d^{-\iota}. \tag{46}$$

Proof of Claim 1. First, according to the Paley-Zygmund Inequality for $x \sim \mathcal{P}^d$ and any $0 \le \theta \le 1$,

$$\mathbb{P}(f_u(x)^2 > \theta \mathbb{E}[f_u(x)^2]) \ge (1-\theta)^2 \frac{\mathbb{E}[f_u(x)^2]^2}{\mathbb{E}[f_u(x)^4]} \tag{47}$$

Therefore we just need to show that

$$\mathbb{E}[f_u(x_i)^2] \gtrsim d^{-\iota} \tag{48}$$

and

$$\mathbb{E}[f_u(x_i)^4] \lesssim \mathbb{E}[f_u(x_i)^2]^2. \tag{49}$$

We have by orthogonality of $q_{;r_i}$,

$$\begin{aligned}\mathbb{E}[f_u(x_i)^2] &= \sum_{r_1,\cdots,r_d \ge 0, \sum_i r_i \le \iota} u^2_{(r_1\cdots r_d)_\iota} c_{r_1\cdots r_d} \alpha_{r_1+\cdots+r_d} / d^{r_1+\cdots+r_d} \\ &\gtrsim \sum_{r_1,\cdots,r_d \ge 0, \sum_i r_i \le \iota} u^2_{(r_1\cdots r_d)_\iota} / d^{r_1+\cdots+r_d}\end{aligned} \tag{50}$$

Hence, it is clear that equation (48) holds.

Equation (49) is more difficult to prove. We need the following lemma:

Lemma 6. Let $f_\gamma(x) := \sum_{\le \iota} \gamma_{r_1\cdots r_d} q_{r_1\cdots r_d}(x)$, then

$$\mathbb{E} f_\gamma(x)^4 \lesssim (\mathbb{E} f_\gamma(x)^2)^2 = \left(\sum_{\sum_i r_i \le \iota} \gamma^2_{r_1\cdots r_d} \right)^2 \tag{51}$$

Proof of Lemma. Write

$$f_\gamma(x)^2 = \sum_{\le 2\iota} \theta_{r_1\cdots r_d} q_{r_1\cdots r_d}(x) \ \ . \tag{52}$$

Since

$$f_\gamma(x)^2 = \sum_{r'_i; r''_i;} \gamma_{r'_i;} \gamma_{r''_i;} q_{r'_i;}(x) q_{r''_i;}(x) \ \ , \tag{53}$$

then by triangle condition there are coefficients $T^{r_i;}_{r'_i;r''_i;}, M^{r_i;}_{r'_i;r''_i;}$ such that

$$
\begin{aligned}
\theta_{r_1\cdots r_d} &= \mathbb{E}_{x\sim\mathcal{P}^d} q_{r_i;}(x) f_\gamma(x)^2 \\
&= \mathbb{E}_{x\sim\mathcal{P}^d} q_{r_i;}(x) \sum_{r'_i;r''_i;} \gamma_{r'_i;}\gamma_{r''_i;} q_{r'_i;}(x) q_{r''_i;}(x) \\
&= \sum_{r'_i;r''_i;} \gamma_{r'_i;}\gamma_{r''_i;} \left(\mathbb{E}_{x\sim\mathcal{P}^d} q_{r_i;}(x) q_{r'_i;}(x) q_{r''_i;}(x)\right) \\
&= \sum_{\forall i, r'_i+r''_i\geq r_i} T^{r_i;}_{r'_i;r''_i;} \gamma_{r'_i;}\gamma_{r''_i;} \\
&= \sum_{s'_i+s''_i=r_i} \sum_{\substack{s'_i\leq r'_i, s''_i\leq r''_i \\ T^{r_i}_{r'_i;r''_i;}\neq 0}} T^{r_i;}_{r'_i;r''_i;} \gamma_{r'_i;}\gamma_{r''_i;} / M^{r_i;}_{r'_i;r''_i;}
\end{aligned}
\tag{54}
$$

where the coefficients are given by

$$
T^{r_i;}_{r'_i;r''_i;} = \mathbb{E}_{x\sim\mathcal{P}^d} q_{r_i;}(x) q_{r'_i;}(x) q_{r''_i;}(x) \tag{55}
$$

and

$$
M^{r_i;}_{r'_i;r''_i;} = \#\{(s'_i;, s''_i;) : s'_i + s''_i = r_i, s'_i \leq r'_i, s''_i \leq r''_i\}. \tag{56}
$$

Now we seek to upper bound $T^{r_i;}_{r'_i;r''_i;}$

$$
\begin{aligned}
&|T^{r_i;}_{r'_i;r''_i;}| \\
&= |\mathbb{E}_{x\sim\mathcal{P}^d} q_{r_i;}(x) q_{r'_i;}(x) q_{r''_i;}(x)| \\
&= |\mathbb{E}_{x\sim\mathcal{P}^d} \prod_{i=1}^d q_{r_i}(x[i]) q_{r'_i}(x[i]) q_{r''_i}(x[i])| \\
&= \left|\prod_{i=1}^d \mathbb{E}_{x_i\sim\mathcal{P}} q_{r_i}(x[i]) q_{r'_i}(x[i]) q_{r''_i}(x[i])\right|
\end{aligned}
$$

Note that $q_0 \equiv 1$. Since the set $I := \{1 \leq i \leq d : \text{one of } r_i, r'_i, r''_i \text{ is nonzero}\}$ is of size at most 3ι, $T^{r_i;}_{r'_i;r''_i;}$ is uniformly bounded by the following constant (depending on ι and $\mathcal{P}$) for $\sum r_i, \sum r'_i, \sum r''_i \leq \iota$:

$$
\left(\max_{r_i, r'_i, r''_i} \mathbb{E}_{X\sim\mathcal{P}} q_{r_i}(X) q_{r'_i}(X) q_{r''_i}(X)\right)^{3\iota} . \tag{57}
$$

As a result, we have

$$
\begin{aligned}
&\theta^2_{r_1\cdots r_d} \\
&= \left(\sum_{s'_i + s''_i = r_i} \sum_{s'_i \le r'_i, s''_i \le r''_i} T^{r_i;}_{r'_i;r''_i;} \gamma_{r'_i;} \gamma_{r''_i;} / M^{r_i;}_{r'_i;r''_i;} \right)^2 \\
&\lesssim \sum_{s'_i + s''_i = r_i} \left(\sum_{s'_i \le r'_i, s''_i \le r''_i} T^{r_i;}_{r'_i;r''_i;} \gamma_{r'_i;} \gamma_{r''_i;} / M^{r_i;}_{r'_i;r''_i;} \right)^2 \quad \text{(Ignoring Constants depending only on } \iota) \\
&\lesssim \sum_{s'_i + s''_i = r_i} \left(\sum_{\substack{s'_i \le r'_i, s''_i \le r''_i \\ T^{r_i;}_{r'_i;r''_i;} \ne 0}} (T^{r_i;}_{r'_i;r''_i;} / M^{r_i;}_{r'_i;r''_i;})^2 \gamma^2_{r'_i;} \right) \left(\sum_{\substack{s'_i \le r'_i, s''_i \le r''_i \\ T^{r_i;}_{r'_i;r''_i;} \ne 0}} \gamma^2_{r''_i;} \right) \quad \text{(Cauchy Inequality)} \\
&\lesssim \sum_{s'_i + s''_i = r_i} \left(\sum_{\substack{s'_i \le r'_i, s''_i \le r''_i \\ T^{r_i;}_{r'_i;r''_i;} \ne 0}} \gamma^2_{r'_i;} \right) \left(\sum_{\substack{s'_i \le r'_i, s''_i \le r''_i \\ T^{r_i;}_{r'_i;r''_i;} \ne 0}} \gamma^2_{r''_i;} \right)
\end{aligned}
\tag{58}
$$

Note that for $T^{r_i;}_{r'_i;r''_i;} \ne 0$, we must have by triangle condition

$$\forall i, r''_i \le r'_i + r_i \tag{59}$$

which means that for any $r''_i \ne 0$, either $r'_i \ne 0$ or $r_i \ne 0$. Then

$$\{i \in [d] : r''_i \ne 0\} \subset \{i \in [d] : r'_i \ne 0\} \cup \{i \in [d] : r_i \ne 0\}. \tag{60}$$

As a result, for fixed $r'_i;$ and $r_i;$, there is less than constantly many $r''_i;$ such that $T^{r_i;}_{r'_i;r''_i;} \ne 0$. Similarly, for fixed (r''_i) and (r_i), there is less than constantly many (r'_i) such that $T^{r_i;}_{r'_i;r''_i;} \ne 0$.

We then have

$$\begin{aligned}
\theta^2_{r_1\cdots r_d} &\lesssim \sum_{s_i'+s_i''=r_i}\left(\sum_{\substack{s_i'\le r_i', s_i''\le r_i''\\ T^{r_i;}_{r_i';r_i'';}\neq 0}} \gamma^2_{r_i';}\right)\left(\sum_{\substack{s_i'\le r_i', s_i''\le r_i''\\ T^{r_i;}_{r_i';r_i'';}\neq 0}} \gamma^2_{r_i'';}\right)\\
&\lesssim \sum_{s_i'+s_i''=r_i}\left(\sum_{s_i'\le r_i'}\gamma^2_{r_i';}\right)\left(\sum_{s_i''\le r_i''}\gamma^2_{r_i'';}\right)\\
&= \sum_{s_i'+s_i''=r_i}\sum_{s_i'\le r_i'}\sum_{s_i''\le r_i''}\gamma^2_{r_i';}\gamma^2_{r_i'';}
\end{aligned} \tag{61}$$

As a result,

$$\sum_{r_i}\theta^2_{r_1\cdots r_d} \lesssim \sum_{r_i}\sum_{s_i'+s_i''=r_i}\sum_{s_i'\le r_i'}\sum_{s_i''\le r_i''}\gamma^2_{r_i';}\gamma^2_{r_i'';} \tag{62}$$

Note that in the RHS term $\gamma^2_{r_i'}\gamma^2_{r_i''}$ appears constantly many times. So we have

$$\mathbb{E}f_\gamma(x)^4 = \sum_{r_i}\theta^2_{r_1\cdots r_d} \lesssim \sum_{r_i',r_i''}\gamma^2_{r_i';}\gamma^2_{r_i'';} = \left(\mathbb{E}f_\gamma(x)\right)^2 \ . \tag{63}$$

□

3.4 Lower Isometry

We now proceed to prove to lower bound the smallest eigenvalue for $\Psi^\top\Psi$, based on the small-ball property established.

Lemma 7. With probability at least blabla, the smallest eigenvalue of $\Psi^\top\Psi$ is larger than $Cd^{-\iota}$.

Step 1. We will first prove: there is $\epsilon > 0$ such that for any $u \in \mathbb{R}^{\binom{\iota+d}{\iota}}$ with $\|u\|_2 = 1$,

$$\mathbb{P}\left(u^\top\Psi^\top\Psi u \ge \frac{p\epsilon}{2d^\iota}\right) \ge 1 - e^{-p^2n/2}. \tag{64}$$

Since

$$u^\top\Psi^\top\Psi u = \frac{1}{n}\sum_{i=1}^n f_u(x_i)^2 \tag{65}$$

with $f_u(x_i)^2 \ge 0$ i.i.d. drawn. Define $Z_i = 1_{\{f_u(x_i)^2\ge\epsilon d^{-\iota}\}}$. According to Claim 1, we can choose ϵ so that $\mathbb{E}Z_i > C(\iota,\mathcal{P}) > 0$. Denote this $C(\iota,\mathcal{P})$ as p. Now we have

$$u^\top\Psi^\top\Psi u \ge \epsilon d^{-\iota}\frac{1}{n}\sum_{i=1}^n Z_i. \tag{66}$$

Using the Hoeffding's inequality,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i \le p-\delta\right) \le e^{-2\delta^2 n}. \tag{67}$$

Take $\delta = \frac{p}{2}$, we get that

$$\mathbb{P}\left(u^\top \Psi^\top \Psi u \ge \frac{p\epsilon}{2d^\iota}\right) \ge 1 - e^{-p^2 n/2}. \tag{68}$$

Step 2. Now we show that there is constant C such that for any $t > 0$, $u^\top \Psi^\top \Psi u$ is $n^2 d^{4\iota} t^{4\iota}$-Lipschitz with probability at least $1 - Cndt^{-\nu}$. In fact, for $\|u\| = \|v\| = 1$, we have

$$\begin{aligned}
&\|u^\top \Psi^\top \Psi u - v^\top \Psi^\top \Psi v\| \\
&= \|u^\top \Psi^\top \Psi u - v^\top \Psi^\top \Psi u + u^\top \Psi^\top \Psi v - v^\top \Psi^\top \Psi v\| \\
&= \|(u-v)^\top \Psi^\top \Psi u + (u-v)^\top \Psi^\top \Psi v\| \\
&= \|(u-v)^\top \Psi^\top \Psi u\| + \|(u-v)^\top \Psi^\top \Psi v\| \\
&\le \|u-v\| \|\Psi^\top \Psi\|_{L^2\to L^2} \|u\| + \|u-v\| \|\Psi^\top \Psi\| \|v\| \\
&= 2\|u-v\| \|\Psi^\top \Psi\|_{L^2\to L^2},
\end{aligned} \tag{69}$$

then $u \to u^\top \Psi^\top \Psi u$ is $2\|\Psi^\top \Psi\|_{L^2\to L^2}$-Lipschitz.

Now we need only to bound the spectral norm of $\Psi^\top \Psi$. We have

$$\begin{aligned}
\|\Psi^\top \Psi\|_{L^2\to L^2}^2 &= \|\Lambda^{-1\top} \Phi^\top \Phi \Lambda^{-1}\|_{L^2\to L^2}^2 \\
&\lesssim \|\Phi^\top \Phi\|_{L^2\to L^2}^2 \\
&= \|\Phi \Phi^\top\|_{L^2\to L^2}^2 \\
&= \|\mathbf{K}^{[\le \iota]}\|_{L^2\to L^2}^2 \\
&\le \|\mathbf{K}\|_F^2 \\
&= \frac{1}{n^2} \sum_{1\le ij\le n} \left(f^{[\le \iota]}(x_i^\top x_j)\right)^2
\end{aligned}$$

The last quantity is at most

$$
\begin{aligned}
&\frac{1}{n^2}\sum_{1\le ij\le n}\left(\sum_{i=0}^{\iota}\alpha_i|x_i^\top x_j|^i\right)^2\\
&\lesssim \frac{1}{n^2}\sum_{ij}(1+|x_i^\top x_j|)^{2\iota})\\
&\le \frac{1}{n^2}\sum_{ij}(1+\|x_i\|^2)^{\iota}(1+\|x_j\|^2)^{\iota}\\
&\le \frac{1}{n^2}\frac{1}{2}\sum_{ij}\left((1+\|x_i\|^2)^{2\iota}+(1+\|x_j\|^2)^{2\iota}\right)\\
&\lesssim d^{2\iota} \text{ with high probability}
\end{aligned}
$$

More specifically, since

$$
\forall t>0, \mathbb{P}\left(\frac{1}{n}\sum_i(1+\|x_i\|^2)^{2\iota}\ge(1+dt^2)^{2\iota}\right)\le\mathbb{P}(\exists i,j,|x_{ij}|\ge t)\le Cnd(1+t)^{-\nu}\ . \tag{70}
$$

$$
\forall t>1, \mathbb{P}\left(\|\Psi^\top\Psi\|^2_{L^2\to L^2}\ge d^{4\iota}t^{4\iota}\right)\le\mathbb{P}(\exists i,j,|x_{ij}|\ge t)\le Cndt^{-\nu}\ . \tag{71}
$$

Step 3. Suppose that $\frac{d^\iota\log d}{n}=o(1)$, then with probability at least $1-e^{-\Omega(n/d^\iota)}$,

$$
\Psi^\top\Psi\succ\Omega(d^{-\iota})\ . \tag{72}
$$

Now for $t=(4\log n)^{1/\nu}>1$, make covering with balls of radius $r=\frac{p\epsilon}{4}n^{-2}d^{-5\iota}t^{-4\iota}$. We have with probability at least $1-\frac{p\epsilon}{4}ndt^{-\nu}$

$$
\|\Psi^\top\Psi\|^2_{L^2\to L^2}<d^{4\iota}t^{4\iota}\ . \tag{73}
$$

Suppose the center of the covering balls are $u_1,\cdots,u_N$ with $N=r^{-O(d^\iota)}$

And for a single u_i, we have with probability at least $1-e^{-p^2n/2}$,

$$
u_i^\top\Psi^\top\Psi u_i\ge\frac{p\epsilon}{2d^\iota}\ . \tag{74}
$$

Then we have with probability at least $1-Ne^{-p^2n/2}$,

$$
\forall i, u_i^\top\Psi^\top\Psi u_i\ge\frac{p\epsilon}{2d^\iota}\ . \tag{75}
$$

Since $u \to u^\top \Psi^\top \Psi u$ is Lipschitz, we have with probability at least

$$1 - \frac{p\epsilon}{4} n d t^{-\nu} - N e^{-p^2 n/2} = 1 - \frac{p\epsilon}{4} n d t^{-\nu} - r^{-O(d^\iota)} e^{-p^2 n/2} \tag{76}$$

for any $\|u\| = 1$ (suppose u_i is the closest center)

$$u^\top \Psi^\top \Psi u \geq u_i^\top \Psi^\top \Psi u_i - \left(d^{4\iota} t^{4\iota}\right) \times r \geq \frac{p\epsilon}{4d^\iota} \ . \tag{77}$$

i.e.,

$$\Psi^\top \Psi \succ \frac{p\epsilon}{4d^\iota} \ . \tag{78}$$

Now to make the probability bound simpler, we choose $t = e^{\epsilon' n/d^\iota}$ with ϵ' to be determined. Then the tail probability is

$$\begin{aligned}
&\frac{p\epsilon}{4} n d t^{-\nu} + N e^{-p^2 n/2} \\
&\sim \frac{p\epsilon}{4} n d e^{-\nu\epsilon' n/d^\iota} + e^{O(\epsilon' n/d^\iota)\cdot O(d^\iota)} e^{-p^2 n/2} \\
&\sim n d e^{-\nu\epsilon' n/(2d^\iota)} + e^{-p^2 n/4} \\
&\sim n d e^{-\Omega(n/d^\iota)} \\
&\sim e^{-\Omega(n/d^\iota)}
\end{aligned} \tag{79}$$

where ϵ' is small enough and $\frac{d^\iota \log d}{n} = o(1)$.

Step 4.

$$\Theta = \Phi^\top \Phi / n = \Lambda^\top \Psi^\top \Psi \Lambda / n \succ C d^{-\iota}. \tag{80}$$

As a result $\mathbf{K}^{[\iota]} = \Phi\Phi^\top / n$ has $\binom{\iota+d}{\iota}$ number of nonzero eigenvalues, all of them larger than $C d^{-\iota}$.

4 Applications to Kernel Ridgeless Regression

4.1 Variance

Theorem 2. Suppose $x \sim \mathcal{P}^d$ and $\boldsymbol{X} \sim \mathcal{P}^{n\times d}$ and $\mathcal{P}$ is subGaussian and $d^\iota \log d = o(n)$.

(i) Suppose that the Taylor expansion coefficients for $k(x, x')$ satisfies:

- $\alpha_1, \cdots, \alpha_\iota > 0$;

- there is $\iota' > 2\iota + 2$ such that $\alpha_{\iota'} > 0$.

Then if $n = o(d^{\iota+1})$ with probability at least $1 - e^{-\Omega(n/d^{\iota})}$ w.r.t. $\boldsymbol{X}$,

$$\text{Variance} := \mathbb{E}_{x\sim\mathcal{P}_X}\|k(\boldsymbol{X},\boldsymbol{X})^{-1}k(\boldsymbol{X},x)\|^2 \leq C\left(\frac{d^{\iota}}{n} + \frac{n}{d^{\iota+1}}\right) \ . \tag{81}$$

(ii) Suppose that the Taylor expansion coefficients for $k(x, x')$ satisfies for some $\iota' \leq \iota$:

- $\alpha_1, \cdots, \alpha_{\iota'} > 0$;
- $\forall \iota'' > \iota', \alpha_{\iota''} = 0$, i.e. k is a polynomial kernel.

Then with probability at least $1 - e^{-\Omega(n/d^{\iota'})}$ w.r.t. $\boldsymbol{X}$,

$$\text{Variance} := \mathbb{E}_{x\sim\mathcal{P}_X}\|k(\boldsymbol{X},\boldsymbol{X})^{-1}k(\boldsymbol{X},x)\|^2 \leq C\frac{d^{\iota'}}{n} \ . \tag{82}$$

Proof. We need only prove (i) because (ii) shall easily follow from the proof of (i).

Claim: with high probability,

$$n\mathbf{K} \succ \Omega(1) \tag{83}$$

To show this, we just need to show that

$$n\mathbf{K}^{[\iota']} \succ \Omega(1) \tag{84}$$

Break $\mathbf{K}^{[\iota']}$ into

$$n\mathbf{K}^{[\iota']} = A + B \tag{85}$$

where A is the diagonal terms and B is the non-diagonal terms.
With probability at least $e^{-\Omega(n)}$, we shall have

$$A \succ \Omega(1) \tag{86}$$

and

$$\|B\|_F \leq O\left(\sqrt{n^2 \times \left(\frac{1}{\sqrt{d}}\right)^{\iota'}}\right) \leq O\left(\frac{1}{\sqrt{d}}\right) \tag{87}$$

Now back to the proof of (i).

Recall the normalized kernel matrix $\mathbf{K} = k(\boldsymbol{X}, \boldsymbol{X})/n$.

$$
\begin{aligned}
&\mathbb{E}_x \|k(\boldsymbol{X}, \boldsymbol{X})^{-1} k(\boldsymbol{X}, x)\|^2 \\
&\lesssim \sum_{\iota'=1}^{\iota} \mathbb{E}_x \|\mathbf{K}^{-1} \frac{1}{n} (\boldsymbol{X} x)^{\iota'} / d^{\iota'} \|^2 + \mathbb{E}_x \|\mathbf{K}^{-1} \frac{1}{n} \sum_{\iota'=\iota+1}^{\infty} (\boldsymbol{X} x)^{\iota'} / d^{\iota'} \|^2 \\
&\lesssim \frac{1}{n^2} \sum_{i=1}^{\iota} \mathbb{E}_x \|\mathbf{K}^{-1} (\boldsymbol{X} x)^{i} / d^{i} \|^2 + \mathbb{E}_x \| \sum_{\iota'=\iota+1}^{\infty} (\boldsymbol{X} x)^{\iota'} / d^{\iota'} \|^2 \\
&\lesssim \frac{1}{n^2} \sum_{\iota'=1}^{\iota} \mathbb{E}_x \|(\mathbf{K}^{[\leq \iota']})^{-1} (\boldsymbol{X} x)^{\iota'} / d^{\iota'} \|^2 + \frac{n}{d^{\iota+1}} \\
&\lesssim \frac{1}{n^2} \sum_{\iota'=1}^{\iota} \mathbb{E}_x \|d^{\iota'} (\boldsymbol{X} x)^{\iota'} / d^{\iota'} \|^2 + \frac{n}{d^{\iota+1}} \quad \text{use Proposition 1} \\
&\lesssim \frac{1}{n^2} \sum_{\iota'=1}^{\iota} \mathbb{E}_x \|(\boldsymbol{X} x)^{\iota'} \|^2 + \frac{n}{d^{\iota+1}} \\
&\lesssim \frac{1}{n^2} \sum_{\iota'=1}^{\iota} (n d^{\iota'}) + \frac{n}{d^{\iota+1}} \text{ with probability at least } 1 - e^{-\Omega(n)} \\
&\leq \frac{d^{\iota}}{n} + \frac{n}{d^{\iota+1}} \ .
\end{aligned}
\tag{88}
$$

□

4.2 Bias

In this section, we bound the bias part for the min-norm interpolated solution.

Theorem 3 (Bias bound)**.** Consider that the target function $f_*(x) = \mathbb{E}[Y|X=x]$ can be represented as

$$
f_*(x) = \int k(x, z) \rho_*(z) dz \tag{89}
$$

with $\rho_*(z)$ bounded. Assume that the probability density of $\mathcal{P}_X$ is bounded away from zero and $\sup_{x \in \mathcal{X}} k(x, x) \leq M$, then we have

$$
\begin{aligned}
\text{Bias} &:= \mathbb{E}_{x \sim \mathcal{P}_X} \left(k(x, \boldsymbol{X}) k(\boldsymbol{X}, \boldsymbol{X})^{-1} f_*(\boldsymbol{X}) - f_*(x) \right)^2 && (90) \\
&\leq C_1(\boldsymbol{X}) \cdot \mathbb{E}_x \left\| k(\boldsymbol{X}, \boldsymbol{X})^{-1} k(\boldsymbol{X}, x) \right\|^2 + \frac{C_2(\boldsymbol{X})}{n} && (91)
\end{aligned}
$$

where the scalar random variables are bounded in ℓ_2-sense $\mathbb{E}_{\boldsymbol{X}}[C_1(\boldsymbol{X})]^2, \mathbb{E}_{\boldsymbol{X}}[C_2(\boldsymbol{X})]^2 \lesssim 1$.

Remark 1. *The statement can be strengthened to in probability statement, as follows*

$$\textit{Bias} := \|k(x, \boldsymbol{X})k(\boldsymbol{X}, \boldsymbol{X})^{-1} f_*(\boldsymbol{X}) - f_*(x)\|_\mu^2 \tag{92}$$

$$\leq \frac{1}{\sqrt{\delta}} \cdot \left(\mathbb{E}_x \left\| k(\boldsymbol{X}, \boldsymbol{X})^{-1} k(\boldsymbol{X}, x) \right\|^2 \vee \frac{1}{n} \right) \tag{93}$$

with probability $1 - \delta$ *on* $\boldsymbol{X}$*. See Proposition 4 for details. We emphasize that here the factor* $\frac{1}{\delta}$ *has* ***no dependence on the dimension*** d*.*

Remark 2. *Note* $\mathbb{E}_x \left\| k(\boldsymbol{X}, \boldsymbol{X})^{-1} k(\boldsymbol{X}, x) \right\|^2$ *is the expression for the Variance as in Sec. 4.1.*

Proof of Theorem 3. Denote $x_i \sim \mathcal{P}_X$ to be i.i.d. samples, and use μ to refer to the probability density of $\mathcal{P}$. Define the following "surrogate" function for analyzing the bias term

$$\tilde{f}_n(x) := \frac{1}{n} \sum_{i=1}^n k(x, x_i) \frac{\rho_*(x_i)}{\mu(x_i)} \tag{94}$$

The proof uses the property of the data-dependent "surrogate" function $\tilde{f}_n(x)$. We start with inserting the $k(x, X)k(X, X)^{-1} \tilde{f}_n(X)$ in the expression

$$\|k(x, \boldsymbol{X})k(\boldsymbol{X}, \boldsymbol{X})^{-1} f_*(\boldsymbol{X}) - f_*(x)\|_\mu^2 \tag{95}$$

$$\precsim \left\| k(x, \boldsymbol{X})k(\boldsymbol{X}, \boldsymbol{X})^{-1} \left[f_*(\boldsymbol{X}) - \tilde{f}_n(\boldsymbol{X}) \right] \right\|_\mu^2 + \|k(x, \boldsymbol{X})k(\boldsymbol{X}, \boldsymbol{X})^{-1} \tilde{f}_n(\boldsymbol{X}) - f_*(x)\|_\mu^2 \tag{96}$$

For the first term

$$= \int \left\langle k(\boldsymbol{X}, \boldsymbol{X})^{-1} k(\boldsymbol{X}, x), f_*(\boldsymbol{X}) - \tilde{f}_n(\boldsymbol{X}) \right\rangle^2 \mu(x) dx \tag{97}$$

$$\leq \int \left\| k(\boldsymbol{X}, \boldsymbol{X})^{-1} k(\boldsymbol{X}, x) \right\|^2 \mu(x) dx \cdot \|f_*(\boldsymbol{X}) - \tilde{f}_n(\boldsymbol{X})\|^2 \quad \text{(Cauchy-Schwarz)} \tag{98}$$

$$= \|f_*(\boldsymbol{X}) - \tilde{f}_n(\boldsymbol{X})\|^2 \cdot \mathbb{E}_{x \sim \mathcal{P}_X} \left\| k(\boldsymbol{X}, \boldsymbol{X})^{-1} k(\boldsymbol{X}, x) \right\|^2 \quad \text{which is the } \underline{\text{Variance}} \text{ in Sec. 4.1.} \tag{99}$$

By Proposition 2,

$$\mathbb{E}_{\boldsymbol{X}} \|f_*(\boldsymbol{X}) - \tilde{f}_n(\boldsymbol{X})\|^2 \precsim 1.$$

For the second term, we know for vector $\tilde{V} = [\frac{\rho_*(x_1)}{n\mu(x_1)}, \ldots, \frac{\rho_*(x_n)}{n\mu(x_n)}] \in \mathbb{R}^n$

$$\tilde{f}_n(\boldsymbol{X}) = k(\boldsymbol{X}, \boldsymbol{X}) \tilde{V}$$

$$\|k(x, \boldsymbol{X})k(\boldsymbol{X}, \boldsymbol{X})^{-1} \tilde{f}_n(\boldsymbol{X}) - f_*(x)\|_\mu^2 \tag{100}$$

$$= \|k(x, \boldsymbol{X})k(\boldsymbol{X}, \boldsymbol{X})^{-1} k(\boldsymbol{X}, \boldsymbol{X}) \tilde{V} - f_*(x)\|_\mu^2 \tag{101}$$

$$= \|\tilde{f}_n(x) - f_*(x)\|_\mu^2. \tag{102}$$

By the Proposition 3, we know $\mathbb{E}_{\boldsymbol{X}} \|f_*(x) - \tilde{f}_n(x)\|_\mu^2 \precsim \frac{1}{n}$. ◻

Proposition 2 (Leave-one-out)**.**

$$\mathbb{E}_{\boldsymbol{X}} \|f_*(\boldsymbol{X}) - \tilde{f}_n(\boldsymbol{X})\|^2 \lesssim 1 \tag{103}$$

Proof of Proposition 2. We claim that,

$$|f_*(x_j) - \tilde{f}_n(x_j)|^2 \tag{104}$$

$$\leq \frac{1}{n^2}\left|f_*(x_j) - k(x_j,x_j)\frac{\rho_*(x_j)}{\mu(x_j)}\right|^2 + \frac{(n-1)^2}{n^2}\left|\frac{1}{n-1}\sum_{i\neq j} k(x_j,x_i)\frac{\rho_*(x_i)}{\mu(x_i)} - f_*(x_j)\right|^2 \tag{105}$$

We know that for any $x, z \in \mathcal{X}$

$$\left|k(z,x)\frac{\rho_*(x)}{\mu(x)}\right| \leq C,$$

due to $k(z,x) \leq \sqrt{k(z,z)k(x,x)} \leq C'$ and the boundedness of $\rho_*(x)/\mu(x)$. It is easy to verify the boundedness of $|f_*(x)| \leq \sup_z |k(z,x)|$. Therefore we know

$$\mathbb{E}_{\boldsymbol{X}} \frac{1}{n^2}\left|f_*(x_j) - k(x_j,x_j)\frac{\rho_*(x_j)}{\mu(x_j)}\right|^2 \leq \frac{1}{n^2}. \tag{106}$$

For the leave-on-out term,

$$\mathbb{E}_{\boldsymbol{X}} \frac{(n-1)^2}{n^2} |\frac{1}{n-1}\sum_{i\neq j} k(x_j,x_i)\frac{\rho_*(x_i)}{\mu(x_i)} - f_*(x_j)|^2 \tag{107}$$

$$\lesssim \frac{(n-1)^2}{n^2}\mathbb{E}_{x_j}\left[\mathbb{E}_{\boldsymbol{X}\setminus x_j}[|\frac{1}{n-1}\sum_{i\neq j} k(x_j,x_i)\frac{\rho_*(x_i)}{\mu(x_i)} - f_*(x_j)|^2 | x_j]\right] \tag{108}$$

$$\lesssim \frac{(n-1)^2}{n^2}\mathbb{E}_{x_j}\left[\frac{1}{n-1}\int k^2(x_j,x)\frac{\rho_*^2(x)}{\mu^2(x)}\mu(x)dx\right] \lesssim \frac{1}{n}. \tag{109}$$

Therefore we have

$$\mathbb{E}_{\boldsymbol{X}} \|f_*(\boldsymbol{X}) - \tilde{f}_n(\boldsymbol{X})\|^2 = \mathbb{E}_{\boldsymbol{X}} \sum_{j=1}^n |f_*(x_j) - \tilde{f}_n(x_j)|^2 \leq n\left(\frac{1}{n^2} + \frac{1}{n}\right) \lesssim 1 \tag{110}$$

□

Proposition 3 (Variance)**.**

$$n\mathbb{E}_{\boldsymbol{X}} \|f_*(x) - \tilde{f}_n(x)\|_\mu^2 \lesssim 1 \tag{111}$$

Proof of Proposition 3.

$$\mathbb{E}_{\boldsymbol{X}} \int \left(f_*(x) - \tilde{f}_n(x)\right)^2 \mu(x)dx \leq \frac{1}{n} \int \int k^2(x, x') \frac{\rho_*^2(x')}{\mu^2(x')} \mu(x')\mu(x)dx'dx. \tag{112}$$

□

Proposition 4 (In probability bounds)**.** The following bounds hold simultaneously with probability at least $1 - \delta$ on $\boldsymbol{X}$,

$$C_1(\boldsymbol{X}) := n\|f_*(x) - \tilde{f}_n(x)\|_\mu^2 \precsim \frac{1}{\sqrt{\delta}}, \tag{113}$$

$$C_2(\boldsymbol{X}) := \|f_*(\boldsymbol{X}) - \tilde{f}_n(\boldsymbol{X})\|^2 = \sum_{j=1}^n \left(f_*(x_j) - \tilde{f}_n(x_j)\right)^2 \precsim \frac{1}{\sqrt{\delta}} \tag{114}$$

Proof. We have shown that $\mathbb{E}_{\boldsymbol{X}} C_1(\boldsymbol{X}) \precsim 1/n$, and $\mathbb{E}_{\boldsymbol{X}} C_2(\boldsymbol{X}) \precsim 1$. Let's use the second moment method to show the in probability bounds for both terms. Define $\tilde{h}(x, x_i) := k(x, x_i)\frac{\rho_*(x_i)}{\mu(x_i)} - f_*(x)$. It is clear that $\mathbb{E}_{x_i \sim \mu}[\tilde{h}(x, x_i)] = 0$ for any fixed x.

Second moment method on $C_1(\boldsymbol{X})$.

$$\mathbb{E}\left[C_1(\boldsymbol{X})\right]^2 = n^2 \mathbb{E}\left[\frac{1}{n^2}\sum_{i,j}\int (k(x, x_i)\frac{\rho_*(x_i)}{\mu(x_i)} - f_*(x))(k(x, x_j)\frac{\rho_*(x_j)}{\mu(x_j)} - f_*(x))\mu(x)dx\right]^2 \tag{115}$$

$$= \frac{1}{n^2}\sum_{i,j,k,l} \mathbb{E}\left[\int \tilde{h}(x, x_i)\tilde{h}(x, x_j)\mu(x)dx\right]\left[\int \tilde{h}(x, x_k)\tilde{h}(x, x_l)\mu(x)dx\right] \tag{116}$$

Clearly the only nonzero terms on the RkS must be either (1) $(i, j) = (k, l)$, or (2) $(i, j) \neq (k, l)$ but $i = j$ and $k = l$. In case (1), we know

$$\sum_{i,j} \mathbb{E}\left[\int \tilde{h}(x, x_i)\tilde{h}(x, x_j)\mu(x)dx\right]^2 \precsim n^2.$$

In case (2), we know

$$\sum_{i \neq k} \mathbb{E}\left[\int \tilde{h}(x, x_i)^2\mu(x)dx\right]\left[\int \tilde{h}(x, x_k)^2\mu(x)dx\right] \precsim n^2.$$

All other terms, must have form $\mathbb{E}\left[\int \tilde{h}(x, x_i)\tilde{h}(x, x_j)\mu(x)dx\right]\left[\int \tilde{h}(x, x_k)^2\mu(x)dx\right] = 0$ for $i \neq j$, or $\mathbb{E}\left[\int \tilde{h}(x, x_i)\tilde{h}(x, x_j)\mu(x)dx\right]\left[\int \tilde{h}(x, x_k)\tilde{h}(x, x_l)\mu(x)dx\right] = 0$ for $i \neq j, k \neq l, (i, j) \neq (k, l)$. Therefore, we have the second moment bound $\mathbb{E}\left[C_1(\boldsymbol{X})\right]^2 \precsim 1$, by Chebyshev's inequality, we have the desired bound.

Second moment method on $C_2(\boldsymbol{X})$.

$$\begin{aligned}\mathbb{E}\left[C_2(\boldsymbol{X})\right]^2 &= \sum_{i,j} \mathbb{E}\left[\left(f_*(x_i) - \tilde{f}_n(x_i)\right)^2 \left(f_*(x_j) - \tilde{f}_n(x_j)\right)^2\right] \\ &\leq \sum_{i,j} \left[\mathbb{E}\left(f_*(x_i) - \tilde{f}_n(x_i)\right)^4\right]^{1/2} \left[\mathbb{E}\left(f_*(x_j) - \tilde{f}_n(x_j)\right)^4\right]^{1/2}\end{aligned}$$

We know that

$$\mathbb{E}\left(f_*(x_i) - \tilde{f}_n(x_i)\right)^4 = \frac{1}{n^4} \sum_{j_1, j_2, j_3, j_4} \mathbb{E}[\tilde{h}(x_i, x_{j_1})\tilde{h}(x_i, x_{j_2})\tilde{h}(x_i, x_{j_3})\tilde{h}(x_i, x_{j_4})].$$

Divide into two case, (1) some j equals i, (2) all j's do not equal i. In the first case, the only nonzero terms are, of the form $\tilde{h}(x_i, x_i)^2 \tilde{h}(x_i, x_j)^2$ with $j \neq i$, or $\tilde{h}(x_i, x_i)\tilde{h}(x_i, x_j)^3$ for $j \neq i$. In both cases there are at most $O(n)$ such terms.

In the second case, the only nonzero terms are $\tilde{h}(x_i, x_j)^4$ with $j \neq i$ (at most $O(n)$ such terms), and $\tilde{h}(x_i, x_j)^2 \tilde{h}(x_i, x_k)^2$, with unique $k, j \neq i$ (at most $O(n^2)$).

Therefore, we know

$$\mathbb{E}\left(f_*(x_i) - \tilde{f}_n(x_i)\right)^4 \precsim \frac{n^2}{n^4} = \frac{1}{n^2}, \tag{117}$$

which implies that

$$\begin{aligned}\mathbb{E}\left[C_2(\boldsymbol{X})\right]^2 &\leq \sum_{i,j} \left[\mathbb{E}\left(f_*(x_i) - \tilde{f}_n(x_i)\right)^4\right]^{1/2} \left[\mathbb{E}\left(f_*(x_j) - \tilde{f}_n(x_j)\right)^4\right]^{1/2} \\ &\precsim n^2 \frac{1}{n^2} = 1.\end{aligned}$$

By Chebyshev's inequality, again we have the desired bound.

□

5 Applications to Wide Neural Networks

5.1 Regression of NTK-type Kernel

Before discussing neural networks, we need some preparation by studying more general kernels. Specifically, we consider kernels of the following form:

$$k(x, x') = \|\tilde{x}\| \|\tilde{x}'\| \sum_{\iota=0}^{\infty} \alpha_\iota \cos^{2\iota} \langle \tilde{x}, \tilde{x}' \rangle \tag{118}$$

where $\tilde{x} = (x^\top, \sqrt{d})^\top$ and $\tilde{x}' = (x'^\top, \sqrt{d})^\top$ and the expansion is guaranteed to converge everywhere uniformly and $\alpha_i > 0$ for all i. Recall here we assume that $X \sim \mathcal{P}^{n\times d}$.

In order to bound generalization, we need only to bound the following quantity:

$$\mathbb{E}_{x\sim\mathcal{P}_X}\|k(\mathbf{X},\mathbf{X})^{-1}k(\mathbf{X},x)\|^2$$

Proposition 5.

$$\text{Variance} := \mathbb{E}_{x\sim\mathcal{P}_X}\|k(\boldsymbol{X},\boldsymbol{X})^{-1}k(\boldsymbol{X},x)\|^2 \le C\left(\frac{d^\iota}{n} + \frac{n}{d^{\iota+1}}\right) \ . \tag{119}$$

Proof of Proposition. We shall proceed in the following steps. First, make approximation of the given kernel by a weighted sum of polynomial-type inner product kernels. Note that

$$\cos\langle\tilde{x},\tilde{x}'\rangle = \frac{d + x^\top x'}{\|\tilde{x}\|\|\tilde{x}'\|},$$

which is an inner product kernel divided by $\|\tilde{x}\|\|\tilde{x}'\|$. Therefore, we write kernel k as

$$k(x,x') = \sum_{\iota=0}^{\infty}\alpha_\iota\|\tilde{x}\|^{1-\iota}\|\tilde{x}'\|^{1-\iota}(d + x^\top x')^\iota$$

For a constant ι_0 large enough, we would have $k^{[\le\iota_0]}$ so close to k so that

$$\left|\mathbb{E}_{x\sim\mathcal{P}^d}\|k(\mathbf{X},\mathbf{X})^{-1}k(\mathbf{X},x)\|^2 - \mathbb{E}_{x\sim\mathcal{P}^d}\|k^{[\le\iota_0]}(\mathbf{X},\mathbf{X})^{-1}k^{[\le\iota_0]}(\mathbf{X},x)\|^2\right|$$

is ignorable. Then we need only to upper bound

$$\mathbb{E}_{x\sim\mathcal{P}^d}\|k^{[\le\iota_0]}(\mathbf{X},\mathbf{X})^{-1}k^{[\le\iota_0]}(\mathbf{X},x)\|^2.$$

Define kernels h_i as

$$h_\iota(x,x') := (d + x^\top x')^\iota,$$

and we have

$$k(x,x') = \sum_{\iota=0}^{\infty}\alpha_\iota\|\tilde{x}\|^{1-\iota}\|\tilde{x}'\|^{1-\iota}h_\iota(x,x').$$

Lemma 8. With high probability on the choice of $x_1,\cdots,x_n,x$,

$$\|k^{[\iota]}(\mathbf{X},\mathbf{X})^{-1}k^{[\iota]}(\mathbf{X},x)\| \sim \|h_\iota(\mathbf{X},\mathbf{X})^{-1}h_\iota(\mathbf{X},x)\| \ .$$

Proof of Lemma. Define a diagonal matrix

$$A := \operatorname{diag}(\|\tilde{x}_1\|,\cdots,\|\tilde{x}_n\|). \tag{120}$$

Then we have

$$k^{[\iota]}(\mathbf{X}) = \alpha_\iota A^{1-\iota} h_\iota(\mathbf{X}) A^{1-\iota} \tag{121}$$

and

$$k^{[\iota]}(\mathbf{X}, x) = \alpha_i A^{1-\iota} h_\iota(\mathbf{X}, x) \|\tilde{x}\|^{1-\iota}. \tag{122}$$

Now we have

$$k^{[\iota]}(\mathbf{X})^{-1} k^{[\iota]}(\mathbf{X}, x) = A^{\iota-1} h_\iota(\mathbf{X}, x) \|\tilde{x}\|^{1-\iota}, \tag{123}$$

and then the conclusion follows directly from

$$A \sim \sqrt{d} I_d, \|\tilde{x}\| \sim \sqrt{d}. \tag{124}$$

□

Now using this Lemma on the kernel, we have

$$\begin{aligned}
&\mathbb{E}_{x\sim\mathcal{P}^d} \|k^{[\leq \iota_0]}(\mathbf{X}, \mathbf{X})^{-1} k^{[\leq \iota_0]}(\mathbf{X}, x)\|^2 \\
\lesssim &\sum_{\iota=0}^{\iota_0} \mathbb{E}_{x\sim\mathcal{P}^d} \|k^{[\leq \iota_0]}(\mathbf{X}, \mathbf{X})^{-1} k^{[\iota]}(\mathbf{X}, x)\|^2 \\
\leq &\sum_{\iota=0}^{\iota_0} \mathbb{E}_{x\sim\mathcal{P}^d} \|k^{[\iota]}(\mathbf{X}, \mathbf{X})^{-1} k^{[\iota]}(\mathbf{X}, x)\|^2 \\
\sim &\sum_{\iota=0}^{\iota_0} \mathbb{E}_{x\sim\mathcal{P}^d} \|h_\iota(\mathbf{X})^{-1} h_\iota(\mathbf{X}, x)\|^2.
\end{aligned} \tag{125}$$

Note that h is an inner product kernel, we can directly use the main proposition to get the desired result.

□

5.2 Generalization Error of Wide Neural Networks

In this section, we show that the neural tangent kernel for wide neural networks are of the form (118) studied in this paper. Here we consider a one-hidden-layer neural network defined as follows:

$$f(x; W, a) := \frac{1}{\sqrt{m}} \sum_{j=0}^{m} a_j \sigma(w_j^\top \tilde{x}),$$

where the input x is a d-dimensional vector $W = (w_1, \cdots, w_m)$ is a $(d+1) \times m$ matrix and $a = (a_1, \cdots, a_m)$ is a m-dimensional vector and $\tilde{x} = (x^\top, \sqrt{d})^\top$.

Suppose that for n labeled datapoints $(x_1, y_1), \cdots, (x_n, y_n)$, the loss is given by

$$L = \frac{1}{2} \sum_{i=1}^{n} (f(x_i; W, a) - y_i)^2$$

Then the full gradient will be given by

$$\frac{\partial L}{\partial a_j} = \sum_{i=1}^{n} \frac{\sigma(w_j^\top \tilde{x}_i)}{\sqrt{m}}(f(x_i; W, a) - y_i),$$

and

$$\frac{\partial L}{\partial w_j} = \sum_{i=1}^{n} \frac{a_j \tilde{x}_i \sigma'(w_j^\top \tilde{x}_i)}{\sqrt{m}}(f(x_i; W, a) - y_i).$$

With a continuous gradient flow

$$\frac{da_j}{dt} = -\frac{\partial L}{\partial a_j}, \frac{dw_j}{dt} = -\frac{\partial L}{\partial w_j},$$

the prediction value of the neural network at a fixed point x will change according to

$$\begin{aligned}\frac{df(x; W(t), a(t))}{dt} &= \sum_{j=1}^{m} \frac{\partial f}{\partial a_j}\frac{da_j}{dt} + \sum_{j=1}^{m} \frac{\partial f}{\partial w_j}^\top \frac{dw_j}{dt} \\ &= -\sum_{j=1}^{m} \frac{\partial f}{\partial a_j}\frac{\partial L}{\partial a_j} - \sum_{j=1}^{m} \frac{\partial f}{\partial w_j}^\top \frac{\partial L}{\partial w_j} \\ &= -\sum_{i=1}^{n} h^m(x, x_i)(f(x_i; W, a) - y_i)\end{aligned}$$

where the neural tanget kernel h^m is defined by

$$h^m(x, x') := \frac{1}{m}\left(\sum_{j=1}^{m} \sigma(w_j^\top \tilde{x})\sigma(w_j^\top \tilde{x}') + x^\top x' \sum_{j=1}^{m} a_i^2 \sigma'(w_j^\top \tilde{x})\sigma'(w_j^\top \tilde{x}')\right) .$$

Assume that the parameters are initialized according to i.i.d. $\mathcal{N}(0, 1)$, then the above kernel converges pointwise to the following kernel as $m \to \infty$:

$$\begin{aligned}h^\infty(x, x') :&= \mathbb{E}_{w \sim \mathcal{N}(0, I_{d+1})}\left(\sigma(w^\top \tilde{x})\sigma(w^\top \tilde{x}') + \tilde{x}^\top \tilde{x}' \sigma'(w^\top \tilde{x})\sigma'(w^\top \tilde{x}')\right) \\ &= \frac{1}{4\pi}\|\tilde{x}\|\|\tilde{x}'\|((\pi - \theta_{\tilde{x}, \tilde{x}'}) \cos\theta_{\tilde{x}, \tilde{x}'} + \sin\theta_{\tilde{x}, \tilde{x}'}) + \frac{1}{2\pi}\tilde{x}^\top \tilde{x}'(\pi - \theta_{\tilde{x}, \tilde{x}'}) \\ &= \frac{1}{4\pi}\|\tilde{x}\|\|\tilde{x}'\| U(\cos\theta_{\tilde{x}, \tilde{x}'})\end{aligned}$$

where $\theta_{\tilde{x}, \tilde{x}'}$ is the angle between $\tilde{x}$ and $\tilde{x}'$ and U is the following function

$$\begin{aligned}
U(t) &:= 3t(\pi - \arccos(t)) + \sqrt{1-t^2} \\
&= 1 + \sum_{i=0}^{\infty} \left(\frac{3(\frac{1}{2})_i}{(1+2i)i!} - \frac{1}{2}\frac{(\frac{1}{2})_i}{(i+1)!} \right) t^{2i+2} \\
&= 1 + \sum_{i=0}^{\infty} \left(\frac{3}{2i+1} - \frac{1}{2(i+1)} \right) \frac{(\frac{1}{2})_i t^{2i+2}}{i!} \\
&= 1 + \sum_{i=0}^{\infty} \frac{(4i+5)(\frac{1}{2})_i t^{2i+2}}{2(2i+1)(i+1)i!}
\end{aligned}$$

where $(\frac{1}{2})_i = \frac{1}{2} \times \frac{3}{2} \times \cdots \times (\frac{1}{2} + i - 1)$ is the Pochhammer symbol. Now we have verified that the neural tangent kernel h are of the form (118).

It is not difficult to prove that for multilayer fully connectedly neural network the NTK is also of this form with all positive Taylor coefficients if seen as a function of $\cos^2 \theta_{x,x'}$.

References

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *arXiv preprint arXiv:1906.11300*, 2019.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine learning and the bias-variance trade-off. *arXiv preprint arXiv:1812.11118*, 2018a.

Mikhail Belkin, Daniel Hsu, and Partha Mitra. Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate. *arXiv preprint arXiv:1806.05161*, 2018b.

Mikhail Belkin, Siyuan Ma, and Soumik Mandal. To understand deep learning we need to understand kernel learning. *arXiv preprint arXiv:1802.01396*, 2018c.

Mikhail Belkin, Alexander Rakhlin, and Alexandre B Tsybakov. Does data interpolation contradict statistical optimality? *arXiv preprint arXiv:1806.09471*, 2018d.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. *arXiv preprint arXiv:1904.12191*, 2019.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

Vladimir Koltchinskii and Shahar Mendelson. Bounding the smallest singular value of a random matrix without concentration. *International Mathematics Research Notices*, 2015(23):12991–13008, 2015.

Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel" ridgeless" regression can generalize. *arXiv preprint arXiv:1808.00387*, 2018.

Shahar Mendelson. Learning without concentration. In *Conference on Learning Theory*, pages 25–39, 2014.

Alexander Rakhlin and Xiyu Zhai. Consistency of interpolation with laplace kernels is a high-dimensional phenomenon. *arXiv preprint arXiv:1812.11167*, 2018.