Honours Thesis Template

An Honours Thesis

submitted in partial fulfilment of the requirements for the degree of

Bachelor of Science with Honours in Mathematics

presented to

the Department of Mathematics

Faculty of Science

National University of Singapore

Assistant Professor Subhroshekhar Ghosh, Supervisor

by

Ng Wei Le

February 28, 2021

This Honours Thesis represents my own work and due acknowledgement is given whenever information is derived from other sources. No part of this Honours Thesis has been or is being concurrently submitted for any other qualification at any other university.

Signed: _____

Acknowledgement

Special thanks to my research supervisor, Prof. Subhroshekhar Ghosh, for guiding me in the direction to take the research, and for the time he spent teaching me the concepts.

Abstract

The bias-variance tradeoff is a tenet of classical Machine Learning practice, where the algorithm tries to balance the error from the bias term and variance term, typically by means of controlling the richness of the model to find a sweet spot between underfitting and overfitting. However, in many modern Machine Learning practices, the model is made to highly overfit the data, even up till the point of near interpolation. This is done even under the conditions of large amounts of data and noise.

The papers discussed in this thesis show that this happens not only in deep learning, but also kernel machines with close to 0 training error, and why current generalization bounds do not explain such phenomenon well. Additionally, a paper proposes a possible performance curve to explain how overfitting could be beneficial, and some empirical evidence to support this.

Contents

1. Introduction

1.1 Supervised Learning

Given a set of (training) datapoints $\{(x_1, y_1), ..., (x_n, y_n)\}$ where $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$ for $1 \leq i \leq n$, and assuming (x_i, y_i) were independent and identically distributed variables drawn from a probability distribution Q, we want to find a predictor function $h_n : \mathbb{R}^d \to \mathbb{R}$ that predicts y "well" given x that has not been seen.

To represent how well this function predicts, let l represent a loss function. Examples include the squared-loss function $l(y, \hat{y}) = (\hat{y} - y)^2$, and the 0-1 loss, usually used for classification: $l(\hat{y}, y) = 1_{\hat{y} \neq y}$.

The general goal of (supervised) machine learning to find a function h that minimizes the expected loss: $\mathbb{E}_{x,y}[l(h(x), y)]$.

In Empirical Risk Minimization (ERM), given datapoints $\{(x_1, y_1), ..., (x_n, y_n)\}$, the goal is to find a function h_n in some class of function $\mathcal{H}$ to minimize the empirical risk:

$$L_{emp}(h_n) = \frac{1}{n} \sum_{i=1}^{n} l(h_n(x_i), y_i).$$

Intuitively, if a function h_n does not predict training data well (has a high $L_{emp}(h_n)$), this would tell us that it will not predict other (x, y) samples drawn randomly as well.

The empirical risk minimizer, $\hat{h} : \mathbb{R}^d \to \mathbb{R}$ is then defined by:

$$\hat{h_n} = \arg \min_{h_n \in \mathcal{H}} L_{emp}(h_n).$$

Classically, there are reasons why one does not tend to choose a function h_n that reduces the empirical loss to near zero values, typically due to bounds on the generalization gap.

1.2 Bias Variance Tradeoff

We define the generalization error (or sometimes known as the generalization gap) as difference between the empirical and expected classifier loss, i.e. $|\mathbb{E}_{x,y}[l(\hat{h_n}(x), y)] - L_{emp}(\hat{h_n})|$. It would make sense to decrease this gap to as little as possible.

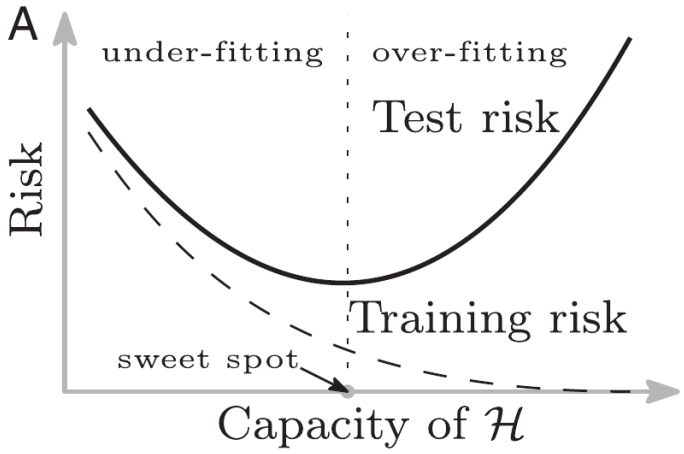

Figure 1.1: Curve showing how the training and test risk changes with respect to the capacity of $\mathcal{H}$. The test risk results from bias-variance tradeoff, and the Capacity of $\mathcal{H}$ is selected at the sweet spot.

Many classical bounds have this generalization gap is a form of:

$$\mathbb{E}_{x,y}[l(\hat{h_n}(x), y)] \leq L_{emp}(\hat{h_n}) + O(\sqrt{c/n}) \tag{1.1}$$

where c is some measure of the complexity of $\mathcal{H}$, for example the fat-shattering dimension, VC-dimension, Rademacher complexity, etc. The general result being that, with a greater complexity of the function class $\mathcal{H}$, the greater the upper bound of this generalization gap. However, if $\mathcal{H}$ is not complex enough, we may not find any function $\hat{h_n} \in \mathcal{H}$ that reduces the empirical risk sufficiently (underfitting). Hence, classical algorithms try to find a balance between over and underfitting, decreasing the expected loss by controlling $\mathcal{H}$, either explicitly or implicitly. For example, to change the complexity of $\mathcal{H}$ explicitly, one might choose a simpler or more complicated architecture for a neural network. To reduce the complexity of $\mathcal{H}$ implicitly, one might user regularization to penalize coefficients to limit the model complexity, or simply stop the training algorithm prematurely (early stopping).

This classical curve from Belkin et al. (2019) is shown in Figure 1.1.

1.3 Modern Machine Learning

In a paper by Zhang et al. (2016), examples were given where deep neural networks were trained to the point of little to no training error. Different architectures were tested on the

CIFAR10 dataset (Krizhevsky & Hinton (2009)) and ImageNet dataset (Russakovsky et al. (2015)). However, very accurate predictions on the new data was given.
In Canziani et al. (2016), the architectures used on ImageNet have large amounts of parameters, multiple times bigger than the number of training datapoints.

1.4 Research Question

Indeed, it is still unanswered as to why these overparameterized data do not seem to cause high test loss due to overfitting. The papers discussed further show empirically that this property is not exclusive to deep learning, but also seems to appear in learning for kernel machines as well. This will be followed by a plausible explanation on how the classical bias-variance tradeoff graph can be reconciled with modern methods.
We first give a brief introduction on kernels and Reproducing Kernel Hilbert Spaces.

2. Kernels

2.1 Notation

We use the symbol $\mathbb{K}$ when it can refer to both $\mathbb{R}$ or $\mathbb{C}$. Also, let z^* or $(z)^*$ denote the conjugate of z for any $z \in \mathbb{C}$. The sections covering Kernels and reproducing kernel Hilbert spaces are heavily referenced using Steinwart, Christman Steinwart & Christmann (2008).

2.2 Definition and Properties

Definition 1. For a non-empty set X, let $k : X \times X \to \mathbb{K}$ be known as a kernel if there exists a function $\phi : X \to \mathcal{H}$ (known as a feature map of k) where $\mathcal{H}$ is a $\mathbb{K}$-Hilbert space (known as a feature space of k) such that

$$k(x_1, x_2) = \langle \phi(x_2), \phi(x_1) \rangle_{\mathcal{H}}. \tag{2.1}$$

Lemma 1. *For any kernel k on X, $k(x_1, x_2) = k(x_2, x_1)^*$.*

From the properties of the inner product, we know that $k(x_1, x_2) = \langle \phi(x_1), \phi(x_2) \rangle^* = k(x_2, x_1)^*$. Therefore, for kernels on $\mathbb{R}$, the symmetric property: $k(x_1, x_2) = k(x_2, x_1)$ holds.

Lemma 2. *Let k_1, k_2 be kernels on a non-empty set X. Then $k_1 + k_2$ and $ak_1, a \in \mathbb{R}^+ \cup \{0\}$ are kernels.*

Below, we define the Gaussian RBF kernel:

Definition 2. Let the complex Gaussian RBF kernel (on $\mathbb{C}^d$)be defined as:

$$k_{\gamma, \mathbb{C}^d}(z, z') := e^{-\gamma^{-2} \sum_{i=1}^{d} (z_i - z_i'^*)^2}.$$

We then define the real Gaussian RBF kernel (or simply the Gaussian RBF kernel for short) acting on $\mathbb{R}^d$ as:

$$k_\gamma(x, x') = e^{-\gamma^{-2} \|x - x'\|_2^2}$$

It can be shown (Steinwart & Christmann (2008)) that the complex and real Gaussian RBF kernels are kernels.

Definition 3. For a non-empty set X, a function $k : X \times X \to \mathbb{R}$ is said to be a positive definite if, for any $m \in \mathbb{Z}^+ \cup \{0\}$ and all $x_1, ..., x_n \in X$, we have the following matrix (called the Gram matrix) being positive semi-definite:

$$K := (k(x_i, x_j))_{i,j}.$$

Equivalently: for all $a_1, ..., a_n \in \mathbb{R}$, we have:

$$\sum_{j=1}^{n} \sum_{i=1}^{n} a_j a_i k(x_j, x_i).$$

Definition 4. The positive definite function $k : X \times X \to \mathbb{R}$ is said to be symmetric if $k(x_1, x_2) = k(x_2, x_1)$ for all $x_1, x_2 \in X$

Theorem 1. *A real function $k : X \times X \to \mathbb{R}$ is a kernel if and only if k is a positive definite symmetric function (also known as a positive definite kernel).*

Proof. Suppose k is a kernel. Then there exists some feature map $\Phi : X \to \mathcal{H}$.

$$\begin{aligned}
\sum_{j=1}^{n} \sum_{i=1}^{n} a_j a_i k(x_j, x_i) &= \sum_{j=1}^{n} \sum_{i=1}^{n} a_j a_i \langle \phi(x_i), \phi(x_j) \rangle_{\mathcal{H}} \\
&= \langle \sum_{i=1}^{n} a_i \phi(x_i), \sum_{j=1}^{n} a_j \phi(x_j) \rangle_{\mathcal{H}} \\
&= \| \sum_{i=1}^{n} a_i \phi(x_i) \| \\
&\geq 0.
\end{aligned}$$

Also, from Lemma 1, we know that the real kernel k is symmetric, proving one side of the theorem. To prove the other side:
Given $k : X \times X \to \mathbb{R}$ a positive definite symmetric function, we prove that $\Phi : X \to H$ where $x \mapsto k(\cdot, x)$ is a valid feature map for some feature space H. First, we define

$$\hat{\mathcal{H}} := \sum_{i=1}^{n} a_i k(\cdot, x_i), n \in \mathbb{Z}^+ \cup \{0\}, a_i \in \mathbb{R} \text{ for all } i, x_i \in X \text{ for all } i.$$

For $f, g \in \hat{\mathcal{H}}$ where $f = \sum_{i=1}^{n} a_i k(\cdot, x_i)$ and $g = \sum_{j=1}^{m} b_j k(\cdot, y_j)$, we define the inner product

as such:

$$\begin{aligned}\langle f, g\rangle &:= \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j k(y_j, x_i) \\ &= \sum_{j=1}^{m} b_j f(y_j) \\ &= \sum_{i=1}^{n} a_i g(x_i)\end{aligned} \tag{2.2}$$

This definition is bilinear and symmetric.

We also have: $\langle f, f\rangle = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j k(x_j, x_i) \geq 0$ since k is a positive definite function. It can be shown that $\langle\cdot,\cdot\rangle$ follows Cauhy-Schwarz Inequality (Steinwart & Christmann (2008)), hence we have:

$$\begin{aligned}|f(x)|^2 &= |\sum_{i=1}^{n} a_i k(\cdot, x_i)|^2 \\ &= |\langle f, k(\cdot, x)\rangle|^2 \ (\because (2.2) \text{ with } g = \sum_{j=1}^{m} b_j k(\cdot, y_j) = k(\cdot, x) \text{ with } m = 1, b_1 = 1, y_1 = x) \\ &\leq \langle k(\cdot, x), k(\cdot, x)\rangle \ \langle f, f\rangle.\end{aligned}$$

Therefore, if $\langle f, f\rangle = 0$, then $f = 0$, hence showing that $\langle f, f\rangle > 0$ if and only if $f \neq 0$. Hence, $\langle\cdot,\cdot\rangle$ defines a proper inner product in $\hat{\mathcal{H}}$.

Let $\mathcal{H}$ be the completion of $\hat{\mathcal{H}}$ and the map $U : \hat{\mathcal{H}} \to \mathcal{H}$ be the map where $\langle Ux, Uy\rangle_{\mathcal{H}} = \langle x, y\rangle_{\hat{\mathcal{H}}}$ for all $x, y \in \hat{\mathcal{H}}$. Then we have, for all $x, x' \in X$:

$$k(x, x') = \langle k(\cdot, x'), k(\cdot, x)\rangle_{\hat{\mathcal{H}}} = \langle Uk(\cdot, x'), Uk(\cdot, x)\rangle_{\mathcal{H}}$$

. Thus we find a feature map of k, proving that k is a kernel. □

2.3 Reproducing Kernel Hilbert Spaces

Initially introduced by Stanislaw Zaremba, reproducing kernel Hilbert spaces have many applications in the fields such as Statistical Learning and complex analysis. An RKHS is a $\mathbb{K}$-Hilbert function space where point evaluation is continuous linear funcitonal.

Definition 5. (RKHS). Let $\mathcal{H}$ be a $\mathbb{K}$-Hilbert space of functions over a non-empty set X.

$\mathcal{H}$ is called an RKHS over X if the Dirac function $\delta_x : \mathcal{H} \to \mathbb{K}$ defined as:

$$\delta_x(f) := f(x),\ x \in X,\ f \in \mathcal{H}$$

is continuous. Equivalently, there exists $0 < M_x < \infty$ such that

$$\delta_x(f) \leq M_x \|f\|_{\mathcal{H}},\ \text{for all} f \in \mathcal{H}.$$

δ_x is called a bounded operator on $\mathcal{H}$.

This is not easy to put into practice, hence the reproducing kernel is defined.

Definition 6. (Reproducing Kernel). For a non-empty set X and a function $k : X \times X \to \mathbb{K}$ where $k(\cdot, x) \in \mathcal{H}$ for all $x \in X$ and the following property hold for all $xinX$ and $f \in \mathcal{H}$:

$$f(x) = \langle f, k(\cdot, x) \rangle \tag{2.3}$$

The condition in equation (2.3) is also known as the reproducing property.

Definition 7. (Canonical Feature Maps). Let $\mathcal{H}$ be an RKHS over X with reproducing kernel k. Let the function $\Phi : X \to \mathcal{H}$ be defined such that for all $x \in X$,

$$\Phi(x) = k(\cdot, x).$$

We call Φ the canonical feature map of k.

Lemma 3. ***(A reproducing kernel of an RKHS is a kernel).*** *Let $\mathcal{H}$ be an RKHS over X with reproducing kernel k. Then k is a kernel.*

Proof. We simply proof that Φ is a feature map of k.

$$\begin{aligned}\langle \Phi(x_2), \Phi(x_1) \rangle &= \langle k(\cdot, x_2), k(\cdot, x_1) \rangle \\ &= k(x_1, x_2)\ (\because \text{Reproducing Property (2.3)})\end{aligned}$$

So $\mathcal{H} is also a feature space of k.$ □

Lemma 4. *Let $\mathcal{H}$ be an $\mathbb{K}$-Hilbert functional RKHS over X with reproducing kernel k. Then H is a Reproducing Kernel Hilbert Space.*

Proof. Recall the Dirac functional $\delta_x : H \to \mathbb{K}$ where:

$$\delta_x(f) = f(x),\ x \in X,\ f \in H.$$

Then we have:

$$\begin{aligned} |\delta_x(f)| &= |f(x)| \\ &= |\langle f, k(\cdot, x)\rangle| \ (\because \text{Reproducing Property (2.3)}) \\ &\leq \|k(\cdot, x)\|_{\mathcal{H}} \|f\|_{\mathcal{H}} \ (\because \text{Cauchy-Schwarz Inequality}) \end{aligned}$$

This shows that the Dirac functionals are continuous. □

2.3.1 Representer Theorem

Representer Theorem ensures that the *argmin* of an empirical risk expression involving a function over an RKHS can be expressed as a linear combination of kernels applied on the training data points as proven in Schölkopf et al. (2001).

Theorem 2. *Given a non-empty set X, training data $\{(x_1, y_1), ...(x_n, y_n)\} \in X \times \mathbb{R}$, and RKHS $\mathcal{H}$ be an $\mathbb{R}$-Hilbert function space over X with reproducing kernel $k : X \times X \to \mathbb{R}$. Let g be a strictly increasing function $g : [o, \infty] \to \mathbb{R}$, and l be an arbitrary loss function, where $l : (X \times \mathbb{R}^2)^n \to \mathbb{R} \cup \{\infty\}$.*
We want to minimize the following empirical risk term:

$$E(f, (x_1, y_1), ..., (x_n, y_n)) \ := \ l((x_1, y_1, f(x_1)), ..., (x_n, y_n, f(x_n))) + g(\|f\|).$$

For $\hat{f} = argmin_{f \in \mathcal{H}} E(f, (x_1, y_1), ..., (x_n, y_n))$, $\hat{f}$ can be represented in the form:

$$\hat{f}(\cdot) = \sum_{i=1}^{n} a_i k(\cdot, x_i)$$

with $a_i \in \mathbb{R}$ for all i.

Proof. First we let Φ be the canonical feature map of k as defined in 7. Recall: function $\Phi : X \to \mathcal{H}$ where $\Phi(x)(\cdot) = k(\cdot, x)$. Due to the reproducing property where $\Phi(x)(x') =$

$\langle \Phi(x), k(\cdot, x') \rangle$, we have:

$$\begin{aligned}\Phi(x)(x') &= k(x', x) \\ &= \langle \Phi(x), k(\cdot, x') \rangle \\ &= \langle \Phi(x), \Phi(x') \rangle.\end{aligned}$$

So Φ is a feature space of k. Using orthogonal decomposition, we decompose $f \in \mathcal{H}$ into a component projected onto the span of $\Phi(x_i), ..., \Phi(x_n)$, and the other component orthogonal to this span. We will then prove this orthogonal component is 0 for any f that reduces the empirical risk term, hence completing the prove.

$$f = \sum_{i=1}^{n} a_i \Phi(x_i) + \gamma,$$

where $\gamma \in \mathcal{H}$, $\langle \Phi(x_i), \gamma \rangle = 0$ for all i.

Next, applying the reproducing property again,

$$\begin{aligned}f(x_j) &= \langle f, k(\cdot, x_j) \rangle \\ &= \langle \sum_{i=1}^{n} a_i \Phi(x_i) + \gamma, \Phi(x_j) \rangle \\ &= \langle \sum_{i=1}^{n} a_i \Phi(x_i), \Phi(x_j) \rangle + \langle \gamma, \Phi(x_j) \rangle \\ &= \sum_{i=1}^{n} a_i \langle \Phi(x_i), \Phi(x_j) \rangle.\end{aligned}$$

Now, consider:

$$\begin{aligned}\|f\|^2 &= \|\sum_{i=1}^{n} a_i \Phi(x_i) + \gamma\|^2 \ (\because \text{orthogonality}) \\ &= \|\sum_{i=1}^{n} a_i \Phi(x_i)\|^2 + \|\gamma\|^2 \\ &\geq \|\sum_{i=1}^{n} a_i \Phi(x_i)\|^2 \\ \Longrightarrow g(\|f\|) &\geq g(\|\sum_{i=1}^{n} a_i \Phi(x_i)\|)\end{aligned}$$

Therefore, if we have $\gamma = 0$, since $f(x_i)$ is unaffected by this for all i, $l((x_1, y_1, f(x_1)), ..., (x_n, y_n, f(x_n)))$ is also unaffected by γ. For the term $g(\|f\|)$, it decreases if we have $\gamma = 0$. Hence,

$\hat{f} = argmin_{f \in \mathcal{H}} E(f, (x_1, y_1), ..., (x_n, y_n))$, $\hat{f}$ must have $\gamma = 0$, and

$$\begin{aligned} \hat{f} &= \sum_{i=1}^{n} a_i \Phi(x_i) \\ &= \sum_{i=1}^{n} a_i k(\cdot, x_i) \end{aligned}$$

□

3. Overfitted and Interpolated Kernel Classifiers

We shall give experimental results performed in Belkin et al. (2018) that show the strong generalization performance also appears in kernel classifiers.

4. Approximation and Estimation

4.1 Notations

Let $\pi_m(\mathbb{R}^d)$ be a multivariate polynomial with d variables and degree $\leq m$, i.e.

$$\pi_m(\mathbb{R}^d) = \{p(x) = \sum_{k \leq m} c_k x^k\}$$

Let $C^k(X)$ be the set of functions on X that are k times continuously differentiable.

For a point $x \in \mathbb{R}^d$, it has the components of its coordinates $\chi_1, .., \chi_d$, whereas we represent n points in $\mathbb{R}^d$ as $x_1, .., x_n$.

We denote $\mathbb{N}_0$ as the set of non-negative integers. We denote the multi-index vector with its components as $\alpha = (\alpha_1, ..., \alpha_d)^{\mathrm{T}} \in \mathbb{N}_0^d$, and $|\alpha| := \|\alpha\|_1$ For $X \subseteq \mathbb{R}^d$, $f \in C^k(X)$, $|\alpha| \leq k$ and $x \in \mathbb{R}^d$, we denote:

$$D^\alpha f := \frac{\partial^{|\alpha|}}{\partial \chi_1^{\alpha_1} \cdots \partial \chi_d^{\alpha_d}} f$$

Definition 8. Consider the set of points $X = x_1, ..., x_n \subseteq \mathbb{R}$ and $\pi_m(\mathbb{R}^d)$ with $n \geq \dim(\pi_m(\mathbb{R}^d)$. X is called $\pi_m(\mathbb{R}^d)$-unisolvent if there is no polynomial in $\pi_m(\mathbb{R}^d)$ (besides the zero polynomial) that is zero on all the points.

4.2 Interpolation Estimates

Let $X = x_1, ..., x_n \subseteq \mathbb{R}$ be a set of points that is $\mathcal{P}$ unisolvent. For $p_1, .., p_Q$ that form a basis of $\mathcal{P}$, let $P = (p_j(x_i)) \in \mathbb{R}^{n \times Q}$. Let Φ be a positive definite kernel and $A = (\Phi(x_i, x_j)) \in \mathbb{R}^{n \times n}$. We let $e^{(j)}$ represent the jth unit vector. Consider the linear system:

$$\begin{pmatrix} A & P \\ P^{\mathrm{T}} & 0 \end{pmatrix} \begin{pmatrix} \alpha^{(j)} \\ \beta^{(j)} \end{pmatrix} = \begin{pmatrix} e^{(j)} \\ 0 \end{pmatrix}. \tag{4.1}$$

This linear system is uniquely solvable for $j \in \{1, 2, .., n\}$.

Proof. For $\begin{pmatrix} \alpha \\ \beta \end{pmatrix}$ that lies in the null space of the matrix $\begin{pmatrix} A & P \\ P^{\mathrm{T}} & 0 \end{pmatrix}$, we have the 2 equations:

$$A\alpha + P\beta = 0,$$
$$P^{\mathrm{T}}\alpha = 0.$$

From the first equation, multiplying both sides by α^{T}, we get:

$$\begin{aligned} &\alpha^{\mathrm{T}}A\alpha + \alpha^{\mathrm{T}}P\beta = 0 \\ \Longrightarrow &\alpha^{\mathrm{T}}A\alpha + (P^{\mathrm{T}}\alpha)^{\mathrm{T}}\beta = 0 \\ \Longrightarrow &\alpha^{\mathrm{T}}A\alpha = 0 \end{aligned}$$

Since we know Φ is a positive definite kernel, then $\alpha = 0$, so we know $P\beta = 0$. Since X is $\mathcal{P}$-unisolvent, so $\beta = 0$. Hence, the linear system (4.1) is uniquely solvable. □

Let u_j^*, V_X be defined such that:

$$\begin{aligned} u_j^* &:= \sum_{i=1}^{n} \alpha_i^{(j)} \Phi(\cdot, x_i) + \sum_{k=1}^{Q} \beta_k^j p_k \\ V_X &:= \{\sum_{i=1}^{n} \alpha_i \Phi(\cdot, x_i) \ : \sum_{i=1}^{n} \alpha_i p(x_i) = 0, p \in \mathcal{P}\} + \mathcal{P}. \end{aligned} \tag{4.2}$$

So we have $u_j^* \in V_X$ and

$$u_i^*(x_j) = \begin{cases} 1 & i = j \\ 0 & i \neq j. \end{cases}$$

Also, for all $f \in V_X$, we have

$$f = \sum_{i=1}^{n} u_i^* f(x_i). \tag{4.3}$$

For our next theorem, we further define:

$$\begin{aligned} R(\cdot) &:= (\Phi(\cdot, x_1), ..., \Phi(\cdot, x_n))^{\mathrm{T}} \in \mathbb{R}^n \\ S(\cdot) &:= (p_1(\cdot), ..., p_Q(\cdot))^{\mathrm{T}} \in \mathbb{R}^Q \end{aligned} \tag{4.4}$$

Theorem 3. *For Φ a positive definite kernel on $\Omega \subseteq \mathbb{R}^d$ and points $X = \{x_1, ..., x_n\} \subseteq \Omega$ is $\mathcal{P}$-unisolvent. There exists functions $v_i^*(\cdot)$ for $i = \{1, 2, .., Q\}$ such that for $u^*(x) =$*

$(u_1^*(x), ..., u_n^*(x)$ where u_i^* is defined in (4.2). we have:

$$\begin{pmatrix} A & P \\ P^T & 0 \end{pmatrix} \begin{pmatrix} u^*(x) \\ v^*(x) \end{pmatrix} = \begin{pmatrix} R(x) \\ S(x) \end{pmatrix} \tag{4.5}$$

Proof. One of the things to be proven is that $P^{\mathrm{T}}u^*(x) = S(x)$. Since $\mathcal{P} \subseteq V_X$, then $(p_1, ..., p_Q) \in V_X$, so by (4.3), we have: $p_i(x) = \sum p_i(x_j)u_i^*(x)$, showing $P^{\mathrm{T}}u^*(x) = S(x)$. Next, we need to show that there exists $v*$ such that $Au^*(x) + Pv^*(x) = R(x)$, so we need to show $Au^*(x) - R(x) \in P(\mathbb{R}^Q)$.
$(P(\mathbb{R}^Q))^\perp$ is the null space of its transpose. Consider $\omega \in \text{null}(P^{\mathrm{T}}) \subseteq \mathbb{R}^n$, so we need to show that $\omega^{\mathrm{T}}(Au^*(x) - R(x)) = 0$, i.e. $\omega^{\mathrm{T}}Au^*(x) = \omega^{\mathrm{T}}R(x)$.

$$\begin{aligned} P^{\mathrm{T}}\omega &= 0 \\ \Longrightarrow \omega^{\mathrm{T}}R(x) &\in V_X \\ \omega^{\mathrm{T}}R(\cdot) &= \sum_{i=1}^{n} u_i(\cdot)\omega^{\mathrm{T}}R(x_i) \quad (\because (4.3)) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} u_i^*(\cdot)\omega_j\Phi(x_i, x_j) \\ &= \gamma^{\mathrm{T}}Au^*(\cdot). \end{aligned}$$

□

Lemma 5. *v^* has the property where, for v_i with $1 \leq i \leq Q$, we have: $v_i(x_j) = 0$ for $1 \leq j \leq n$.*

Proof.

$$\begin{aligned} Au^*(x_i) &= Ae^{(i)} \\ &= \begin{pmatrix} \Phi(x_i, x_1) \\ \vdots \\ \Phi(x_i, x_n) \end{pmatrix} \\ Au^*(x_i) + Pv^*(x_i) &= R(x_i) = Au^*(x_i) \\ \Longrightarrow Pv^*(x_i) &= 0 \end{aligned}$$

Since X is $\mathcal{P}$-unisolvent, $v^*(x_i) = 0$. □

Also, from Theorem 11.1 of Wendland (2004), we are able to rewrite an interpolant as:

$$s_{f,X}(\cdot) = \sum_{i=1}^{n} f(x_i) u_i^*(\cdot). \tag{4.6}$$

By differentiating 4.5, we have:

$$\begin{pmatrix} A & P \\ P^{\mathrm{T}} & 0 \end{pmatrix} \begin{pmatrix} D^a u^*(x) \\ D^a v^*(x) \end{pmatrix} = \begin{pmatrix} D^a R(x) \\ D^a S(x) \end{pmatrix}. \tag{4.7}$$

We will define the power function, as defined in Definition 11.2 in Wendland (2004):

Definition 9. Suppose $X \in \mathbb{R}^d$ is open, with $k : X \times X \to \mathbb{R}$ be a positive definite kernel. For $\alpha \in \mathbb{N}_0^d$, $\hat{X} = x_1, x_2, ..., x_n \subseteq X$ the power function $P_{k,\hat{X}}^{(\alpha)}(x)$ is defined by:

$$\begin{aligned}(P_{k,\hat{X}}^{(\alpha)}(x))^2 := D_1^\alpha D_2^\alpha k(x,x) - 2\sum_{j=1}^{n} D^\alpha u_j^*(x) D_1^\alpha k(x, x_j) \\ + \sum_{i,j=1}^{n} D^\alpha u_i^*(x) D^\alpha u_j^*(x) k(x_i, x_j).\end{aligned}$$

In Theorem 11.4 of Wendland (2004), we have:

Theorem 4. *For an open set $\Omega \subseteq \mathbb{R}$ and a positive definite kernel $k \subseteq C^{2k}(\Omega \times \Omega)$, and the set of points $X = \{x_1, ..., x_n\} \subseteq \Omega$ is $\mathcal{P}$-unisolvent. Let $\mathcal{H}$ be RKHS corresponding to the kernel k, a function $f \in \mathcal{H}$ and its interpolant be $s_{f,X}$. For every $x \in \Omega, a \in \mathbb{N}_0^d, |a| \leq k$, we have:*

$$|D^a f(\cdot) - D^a s_{f,X}(\cdot)| \leq P_{k,X}^a(\cdot) \|f\|_{\mathcal{H}}. \tag{4.8}$$

Proof. We shall proof the case of $a = 0$, which is the case which will be used in a later theorem. First, we note that:

$$\begin{aligned}\|k(\cdot, x) - \sum_{i=1}^{n} u_i k(\cdot, x_i)\|_{\mathcal{H}}^2 &= k(x,x) - 2\sum_{i=1}^{n} u_i^* k(x, x_i) + \sum_{i,j=1}^{n} u_i^* u_j^* k(x_i, x_j) \\ &= (P_{k,X}^{(0)}(x))^2.\end{aligned}$$

Next, using (4.6), we have:

$$\begin{aligned}
s_{f,X}(x) &= \sum_{i=1}^{n} f(x_i)u_i^*(x) \\
&= \sum_{i=1}^{n} u_i^*(x)\langle f, k(\cdot, x_i)\rangle_{\mathcal{H}} \quad (\because \text{reproduction property}) \\
&= \langle f, \sum_{i=1}^{n} u_i^*(x)k(\cdot, x_i)\rangle \\
\implies |f(x) - s_{f,X}(x)| &= |\langle f, k(\cdot, x)\rangle_{\mathcal{H}} - \langle f, \sum_{i=1}^{n} u_i^*(x)k(\cdot, x_i)\rangle| \\
&= |\langle f, k(\cdot, x) - \sum_{i=1}^{n} u_i^*(x)k(\cdot, x_i)\rangle_{\mathcal{H}}| \\
&\leq \|f\|_{\mathcal{H}} \|k(\cdot, x) - \sum_{i=1}^{n} u_i^*(x)k(\cdot, x_i)\|_{\mathcal{H}} \quad (\because \text{Cauchy-Schwarz inequality}) \\
&= \|f\|_{\mathcal{H}} P_{k,X}^{(0)}(x).
\end{aligned}$$

□

Definition 10. The fill distance (or sometimes referred to as 'fill' for short) for a set of points $X = \{x_1, ..., x_N\} \subseteq \Omega$ for a bounded domain Ω is defined to be

$$h_{X,\Omega} \coloneqq \sup_{x \in \Omega} \min_{1 \leq j \leq N} \|x - x_j\|_2.$$

Theorem 11.22 in Wendland (2004):

Theorem 5. *Let Ω be a cube in $\mathbb{R}^d$ and $k = \phi(\|\cdot\|_2)$ be a positive definite function with $f = \phi(\cdot)$ satisfying the condition that there exists, l_0 and constant $M > 0$ such that for all $r > 0$ and $l > l_0$, $|f^{(l)}(r)| \leq l!M^l$. Then there exists a constant $c > 0$ such that the error between a function $f \in \mathcal{H}_\infty$ and its interpolant $s_{f,X}$ for all data points $X = \{x_1, ..., x_n\}$ can be bounded by:*

$$\|f - s_{f,X}\|_{L_\infty(\Omega)} \leq exp(-c/h_{X,\Omega})|f|_{\mathcal{H}_\infty}$$

with sufficiently small fill $h_{X,\Omega}$.

Proof. From Theorem 11.9 in Wendland (2004), we have:

$$P_{\Phi,X}^2(x) \leq [1 + c_1(2N)]^2 \|f - p\|_{L_\infty(G)}$$

Where G is on the interval $[0, 4(c_2(2N))^2h^2]$, $x \in \Omega$, $p \in \pi_n(\mathbb{R})$, $h = h_{X,\Omega}$.

From Theorem 11.21 in Wendland (2004),: we have for sufficiently small fill distance $h_{X,\Omega} \leq \frac{c_0}{2n}$, there exists a constant γ_d, where the constants c_1, c_2 can be replaced by:

$$c_1(2n) = \exp(2d\gamma_d(2n+1))$$
$$c_2(2n) = 2c_2n$$

So G is on the interval $[0, 16N^2c_2^2h^2]$ For p the Taylor series of f about 0, and up to the term t^N, we then have:

$$\begin{aligned} |f(t) - p(t)| &\leq t^{N+1}\frac{|f^{n+1}(t')|}{(n+1)!} \\ &\leq M^{N+1}t^{N+1} \quad \text{(By assumption)} \\ \implies \|f-p\|_{L_\infty(G)} &\leq (M \cdot 16N^2c_2^2h^2)^{N+1} \\ &= (C_0N^2h^2)^{N+1} \quad \text{for constant } C_0 = Mc_2^2 \end{aligned}$$

Also, we have:

$$\begin{aligned} [1 + c_1(2N)]^2 &= [1 + \exp(2d\gamma_d(2N+1))]^2 \\ &\leq [2\exp(2d\gamma_d(2N+1))]^2 \\ &= 4\exp(4d\gamma_d(2N+1)) \\ &= \exp(\log 4 + 4d\gamma_d(2N+1)) \\ &\leq \exp(C_1(N+1)) \quad \text{for sufficiently large } C_1. \end{aligned}$$

We then have:

$$\begin{aligned} P^2_{\Phi,X}(x) &\leq [1 + c_1(2N)]^2\|f-p\|_{L_\infty(G)} \\ &\leq \exp(C_1(N+1))(C_0N^2h^2)^{N+1} \\ &= (C_0N^2h^2\exp(C_1))^{N+1} \\ &= (C_2N^2h^2)^{N+1} \quad \text{for constant } C_2 = C_0\exp(C_1). \end{aligned} \tag{4.9}$$

For $C_3 = \min(\frac{c_0}{2}, \frac{1}{\sqrt{eC_2}})$, and N such that

$$\frac{C_3}{N+1} \leq h \leq \frac{C_3}{N}$$

, which gives us:

$$\begin{aligned} h &\leq \frac{c_0}{2N}, \\ -(N+1) &\leq -C_3/h, \\ N^2h^2 \leq & C_3^2 \leq \frac{1}{eC_2} \\ \Longrightarrow & C_2N^2h^2 \leq 1/e. \end{aligned}$$

We then have:

$$P^2_{\Phi,X}(x) \leq e^{-(N+1)} \leq e^{-C_3/h}.$$

Now, using $C = C_3/2$ and $|f(x) - s_{f,X}(x)| \leq P_{\Phi,X}(x)|f|_{\mathcal{H}_\infty}$ (from Theorem (4)), we have:

$$|f(x) - s_{f,X}(x)| \leq e^{-C/h_{X,\Omega}}|f|_{\mathcal{H}_\infty},$$

which is what we wanted to prove. □

Theorem 6. *With the same conditions as Theorem 5, except that f satisfies the stricter condition $|f^{(l)}(r)| \leq M^l$, we can get a better error bound of:*

$$\|f - s_{f,X}\|_{L_\infty(\Omega)} \leq \exp(\frac{c\log(h_{X,\Omega})}{h_{X,\Omega}})\|f\|_{\mathcal{H}_\infty}.$$

Proof. The inequality at 4.9 becomes:

$$P^2_{\Phi,X}(x) \leq \frac{(C_2N^2h^2)^{N+1}}{(N+1)!}.$$

Using Stirling's inequality $1/n! \leq (e/n)^n$, we have:

$$P^2_{\Phi,X}(x) \leq (eC_2Nh^2)^{N+1}.$$

Similarly, with $C_3 = \min(\frac{c_0}{2}, \frac{1}{eC_2})$, and N such that

$$\frac{C_3}{N+1} \leq h \leq \frac{C_3}{N},$$

we then get

$$eC_2Nh \leq 1,$$

which gives us:

$$P^2_{\Phi,X}(x) \leq h^{N+1} \leq h^{C_3/h} = e^{C_3 \log h/h}.$$

Following the steps of the previous theorem then gives us our result. □

4.3 Approximation Theorem

The below theorem gives us some justification as to why the minimum norm interpolating function was chosen, though this only works under noiseless conditions:

Theorem 7. *Fix $h^* \in \mathcal{H}_\infty$. Let $(x_1, y_1), ..., (x_n, y_n)$ be i.i.d. random variables where x_i drawn uniformly at random from a compact cube $\Omega \subseteq \mathbb{R}^d$, $y_i = h^*(x_i) \forall i$. There exists $A, B > 0$ such that for any interpolating $h \in \mathcal{H}_\infty$ with high probability*

$$\sup_{x\in\Omega} |h(x) - h^*(x)| A e^{-B(n/\log n)^{1/d}} (\|h^*\|_{\mathcal{H}_\infty} + \|h\|_{\mathcal{H}_\infty})$$

With $h_{X,\Omega}$ as the fill on the order of $O(n/\log n)^{-1/d}$ (using the theorem S1 in Belkin's paper TODO). We consider $f(x) := h(x) - h^*(x)$. Since h is interpolating, we have $f(x_i) = 0$ for all x_i. The interpolant of f, $s_{f,X}$ will then be the zero function. Thus, we have:

$$\begin{aligned}
\|f\|_{L_\infty(\Omega)} &= \sup_{x\in\Omega} |h(x) - h^*(x)| \\
&< \exp(-c(n/\log n)^{1/d}) \|f\|_{\mathcal{H}_\infty} \\
&\leq \exp(-c(n/\log n)^{1/d}) (\|h^*\|_{\mathcal{H}_\infty} + \|h\|_{\mathcal{H}_\infty})
\end{aligned}$$

Another form we can have is using Proposition 14.1 in Wendland (2004):

Proposition 1. *Let $\Omega \subseteq \mathbb{R}^d$ be bounded and measurable. Suppose $X = \{x_1, ..., x_N\} \subseteq \Omega$ is quasi-uniform with respect to $c_{qu} > 0$. Then there exists constants $c_1, c_2 > 0$ depending only on space dimension d, on Ω and on c_{qu} such that:*

$$c_1 N^{-1/d} \leq h_{X,\Omega} \leq c_2 N^{-1/d}.$$

With the definition of quasi-uniformness being:

Definition 11. For the separation distance of $X = \{x_1, ..., x_N\}$ being defined as $q_x := \frac{1}{2} \min_{i\neq j} \|x_i - x_j\|_2$.

We can then use the above proposition with n replacing $n/\log n$ in Theorem 7 for a better bound of:

$$\sup_{x\in\Omega}|h(x) - h^*(x)| < Ae^{-B(n)^{1/d}}(\|h^*\|_{\mathcal{H}_\infty} + \|h\|_{\mathcal{H}_\infty})$$

if the quasi-uniformness conditions are fulfilled. In either case, by choosing a the smallest norm for h, we can see that it corresponds to the smallest upperbound for $|h(x) - h^*(x)|$. In the next subsection, we have another rationale for choosing a smaller $\|h\|_{\mathcal{H}_\infty}$:

4.3.1 Regularity of Functions

Let the $\mathbb{R}$-RKHS $\mathcal{H}$ consist of functions over X. The functions fulfill a Lipschitz-like condition with the Lipschitz constant being the RKHS norm of the function $\|f\|_{\mathcal{H}_\infty}$.
For the positive definite symmetric kernel k corresponding to $\mathcal{H}$, we have for all $x, x' \in X$:

$$\begin{aligned}
|f(x) - f(x')| &= |\langle f, k(\cdot, x)\rangle_{\mathcal{H}_\infty} - \langle f, k(\cdot, x')\rangle_{\mathcal{H}_\infty}| \\
&= |\langle f, \Phi(x) - \Phi(x')\rangle_{\mathcal{H}_\infty}| \quad \text{where } \Phi(x)(\cdot) = k(\cdot, x) \\
&\leq \|f\|_{\mathcal{H}_\infty} d(x, x'),
\end{aligned}$$

where $d(x, x')$ is the pseudometric defined by:

$$d^2(x, x') = k(x, x) - 2k(x, x') + k(x', x').$$

Intuitively, we can see the "speed" of which the function can change is bounded by the norm of the function in the RKHS space.

5. Existing Bounds

5.1 Existing Bounds Provide No Guarantees for Interpolated Kernel Classifiers

Steps are:

- Find lower bound on function norm of t-overfitted classifiers in RKHS corresponding to Gaussian Kernels.
- Show loss for available bounds for kernel methods based on function norm (can perhaps use this to explain approximation theorem as well?)

Interpolation: 0 regression error. Overfitting: 0 classification error. Interpolation implies overfitting.

Definition 12. We say $h \in H$ t-overfits data, if it achieves zero classification loss (overfits) and $\forall_i y_i h(x_i) > t > 0$.

The below shows a theorem on how the function norm changes with respect to t-overfitting.

Theorem 8. *Let $(\mathbf{x}_i, y_i)$ be data sampled from P on $\Omega \times \{-1, 1\}$ for $i = 1, ..., n$. Assume that y is not a deterministic function of x on a subset of non-zero measure. Then, with high probability, any h that t-overfits the data, satisfies*

$$\|h\|_H > Ae^{Bn^{1/d}}$$

for some constants $A, B > 0$ depending on t.

We define the γ-shattering and fat-shattering dimension below:

Definition 13. Let F be a set of functions mapping from a domain X to $\mathbb{R}$. Suppose $S = \{x_1, x_2, ..., x_m\} \subseteq X$. Suppose also that γ is a positive real number. Then S is γ-shattered by F if there are real numbers $r_1, r_2, ..., r_m$, such that for each $b \in \{0, 1\}^m$ there

is a function f_b in F with

$$f_b(x_i) \geq r_i + \gamma \text{ if } b_i = 1, \text{ and } f_b(x_i) \leq r_i - \gamma \text{ if } b_i = 0, \text{ for } 1 \leq i \leq m.$$

We say $r = (r_1, r_2, ..., r_m)$ witnesses the shattering. Suppose that F is a set of functions from a domain X to $\mathbb{R}$ and that $\gamma > 0$. Then F has γ-dimension d if d is the maximum cardinality of a subset S of X that is γ-shattered by F. If no such maximum exists, we say that F has infinite γ-dimension. The γ-dimension of F is denoted $fat_F(\gamma)$. This defines a function $fat_F : \mathbb{R} \to N \cup \{0, \infty\}$, which we call the fat-shattering dimension of F.

Proof. Let $B_R = \{f \in \mathcal{H}, \|f\|_{\mathcal{H}} < R\}$ be a ball of radius R in RKHS $\mathcal{H}$. Suppose the data is γ-overfitted, Kégl et al. (2001) gives us a high probability of a bound of

$$L(f) < O(\frac{\ln(n)^2}{\sqrt{n}} \sqrt{fat_{B_R}(\gamma/8)})$$

for $L(f)$ the expected classification error. Also, from Belkin (2018) we have

$$fat_{B_R}(\gamma) < O((log(R/\gamma))^d)$$

. We then have B_R containing no function that γ overfits the data unless

$$(log(R/\gamma))^d > O(n) \Longrightarrow R > c_1 \exp(c_2(\frac{n}{\ln n})^{1/d})$$

for some positive constants c_1, c_2. □

Classical bounds for kernel methods (Belkin et al. (2018)) are in the form:

$$|\frac{1}{n}\sum_i l(f(x_i), y_i) - L(f)| \leq C\frac{\|f\|_{\mathcal{H}}^a}{n^b}, \quad C, a, b \geq 0$$

The right side on this will tend to infinity for bigger $\|f\|_{\mathcal{H}}$, which is suggested by Theorem 8.

6. Experiments

6.1 Random Fourier Features

For a feature map $\phi : \mathbb{R}^d \to \mathbb{R}^{d'}$ the kernel trick allows easy computation for positive definite kernel k where $k(x, y) =< \phi(x), \phi(y) >$. We want to find a randomized feature map $z : \mathbb{R}^d \to \mathbb{R}^{\bar{d}}$ such that

$$k(x, y) =< \phi(x), \phi(y) >\approx< z^{\mathrm{T}}(x), z(y) >$$

. As suggested by Rahimi & Recht (2008), for a shift-invariant kernel k: $k(x, y) = k(x - y))$, we consider the mapping $z(x) = cos(w^{\mathrm{T}}x + b)$, where w is drawn from the probability distribution p:

$$p(w) = \frac{1}{2\pi} \int k(h) \exp(-iw^{\mathrm{T}}h) \, \mathrm{d}h \tag{6.1}$$

when we compute the Fourier transform of the kernel k, and b is drawn from the uniform distribution on $[0, 2\pi]$.

We know that the fourier transform of $k(\cdot)$ is a probability distribution from Bochner's theorem:

Theorem 9. *(Bochner Rudin (1990)).For a continuous kernel $k(x - y)$ it is a positive definite kernel if and only if $k(\cdot)$ is the fourier transform of a non-negative measure.*

We now have:

$$k(x - y) = \int_{\mathbb{R}^d} p(w)\exp(iw^{\mathrm{T}}(x - y)) \, \mathrm{d}w = \mathbb{E}_w[e^{iw^{\mathrm{T}}x}(e^{iw^{\mathrm{T}}y})^*]$$

. Therefore, we can use $e^{iw^{\mathrm{T}}x}(e^{iw^{\mathrm{T}}y})^*$ as an estimate (unbiased) of $k(x, y)$. Let $\phi_w(x) = e^{iw^{\mathrm{T}}x}$ We can also use $z_w(x) = \sqrt{2}cos(w^{\mathrm{T}}x + b)$ instead of $\phi_w(x)$, as suggested by Rahimi & Recht (2008).

Proposition 2. *For $z_w(x) = \sqrt{2}cos(w^T x + b)$, where w is drawn from probability distribution*

p in (6.1) *and b drawn from a uniform random variable on* $[0, 2\pi]$.

$$E(z_w(x))z_w(y) = k(x, y)$$

Proof.

$$\begin{aligned} z_w(x) &= 2\ \frac{\sqrt{2}}{2} cos(w^{\mathrm{T}}x + b) \\ &= \frac{1}{\sqrt{2}}\ (e^{i(w^{\mathrm{T}}x+b)} + e^{-i(w^{\mathrm{T}}x+b)}) \\ &= \frac{1}{\sqrt{2}}\ (\phi_w(x)e^{ib} + \phi_w(x)^* e^{-ib}) \end{aligned}$$

Where $\phi_w(x) = e^{iw^{\mathrm{T}}x}$.

$$z_w(x)z_y(y) = \frac{1}{2}[\phi_w(x)\phi_w(y)e^{i2b} + \phi_w(x)^*\phi_w(y)^* e^{-i2b} + \phi_w(x)\phi_w(y)^* + \phi_w(x)^*\phi_w(y)]$$

$$\mathbb{E}[z_w(x)z_y(y)] = \frac{1}{2}\mathbb{E}[\phi_w(x)\phi_w(y)e^{i2b} + \phi_w(x)^*\phi_w(y)^* e^{-i2b}] + \frac{1}{2}\mathbb{E}[\phi_w(x)\phi_w(y)^*] + \frac{1}{2}\mathbb{E}[\phi_w(x)^*\phi_w(y)]$$

As mentioned earlier in Theorem 9, $\mathbb{E}_w[\phi_w(x)\phi_w(y)^*] = k(x - y)$. Also $\phi_w(x)\phi_w(y)^* = (\phi_w(x)^*\phi_w(y))^*$.

$$\begin{aligned} \mathbb{E}[z_w(x)z_y(y)] &= \frac{1}{2}\mathbb{E}[\phi_w(x)\phi_w(y)e^{i2b} + \phi_w(x)^*\phi_w(y)^* e^{-i2b}] + \frac{1}{2}k(x-y) + \frac{1}{2}[k(x-y)]^* \\ &= \frac{1}{2}\mathbb{E}[\phi_w(x)\phi_w(y)e^{i2b} + \phi_w(x)^*\phi_w(y)^* e^{-i2b}] + k(x-y) \end{aligned}$$

For real kernel , $k(x - y) = (k(x - y))^*$.

$$\begin{aligned} \mathbb{E}_{w,b}[\phi_w(x)\phi_w(y)e^{i2b}] &= \frac{1}{2\pi}\int_{\mathbb{R}^d}\int_0^{2\pi} p(w)\phi_w(x)\phi_w(y)e^{i2b}\mathrm{d}b\ \mathrm{d}w \\ &= \frac{1}{2\pi}\int_{\mathbb{R}^d} p(w)\phi_w(x)\phi_w(y)\int_0^{2\pi} e^{i2b}\mathrm{d}b\ \mathrm{d}w \\ &= 0 \end{aligned}$$

Since $\int_0^{2\pi} e^{i2b}\mathrm{d}b = 0$. Similarly, $\mathbb{E}_{w,b}[\phi_w(x)^*\phi_w(y)^* e^{-i2b}] = 0$.

$$\therefore \mathbb{E}[z_w(x)z_y(y)] = k(x - y).$$

□

As suggested by Rahimi & Recht (2008), the variance of the estimate is decreased by

using z, a D dimensional vector by concatenating D of z_w and normalizing by a constant $\sqrt{D}$. We let:

$$z(x) = \sqrt{\frac{2}{D}}[cos(w_1^{\mathrm{T}}x + b_1)...cos(w_D^{\mathrm{T}}x + b_D)]$$

with randomly drawn w_i and b_i as described previously.

Theorem 10. *For N the number of random features, and $x_1, x_2, ..., x_n$ the data points, when $N > n$ and as N increases, the norm of the minimizer tends to the norm of the minimum norm RKHS interpolant.*

Proof. Let $f(x)$ be the minimum norm RKHS interpolant function for the datapoints.

$$f(x) = \sum_i \alpha_i k(x_i, x) \approx \sum_i \alpha_i z(x_i)^{\mathrm{T}} z(x) = \beta^{\mathrm{T}} z(x) = \hat{f}(x)$$

(the first equality holds due to Representer Theorem) Where $\beta = \sum_i \alpha_i z(x_i)$. The norm of the function from the random fourier features approximation is:

$$\|\beta\| = \beta^{\mathrm{T}}\bar{\beta} = (\sum_i \alpha_i z^{\mathrm{T}}(x_i))(\sum_i \bar{\alpha}_i \bar{z}(x_i)) = \sum_i \sum_j \alpha_i \bar{\alpha}_j z^{\mathrm{T}}(x_i)\bar{z}(x_j) \approx \sum_i \sum_j \alpha_i \bar{\alpha_j} k(x_i, x_j) = \|f\|$$

□

7. Other Notes

7.1 Weaknesses

Huge sample size, density of adversarial examples.

A. Your first appendix

See `https://www.overleaf.com/read/qyhckhfyfvmb` on Overleaf for useful examples of formatted text, typing in IPA, etc.

Bibliography

Belkin, Mikhail. 2018. Approximation beats concentration? an approximation view on inference with smooth radial kernels. *arxiv:1801.03437* http://arxiv.org/abs/1801.03437v2.

Belkin, Mikhail, Daniel Hsu, Siyuan Ma & Soumik Mandal. 2019. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proc Natl Acad Sci USA* 116(32). 15849–15854. doi:10.1073/pnas.1903070116. https://doi.org/10.1073%2Fpnas.1903070116.

Belkin, Mikhail, Siyuan Ma & Soumik Mandal. 2018. To understand deep learning we need to understand kernel learning. *arxiv:1802.01396* http://arxiv.org/abs/1802.01396v3.

Canziani, Alfredo, Adam Paszke & Eugenio Culurciello. 2016. An analysis of deep neural network models for practical applications .

Kégl, Balázs, Tamás Linder & Gábor Lugosi. 2001. Data-dependent margin-based generalization bounds for classification. In *Lecture notes in computer science*, 368–384. Springer Berlin Heidelberg. doi:10.1007/3-540-44581-1_24. https://doi.org/10.1007%2F3-540-44581-1_24.

Krizhevsky, A & G Hinton. 2009. Learning multiple layers of features from tiny images .

Rahimi, Ali & Benjamin Recht. 2008. Random features for large-scale kernel machines. In J. Platt, D. Koller, Y. Singer & S. Roweis (eds.), *Advances in neural information processing systems*, vol. 20, 1177–1184. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2007/file/013a006f03dbc5392effeb8f18fda755-Paper.pdf.

Rudin, Walter. 1990. *Fourier analysis on groups*. John Wiley & Sons, Inc. doi:10.1002/9781118165621. https://doi.org/10.1002%2F9781118165621.

Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg & Li Fei-Fei. 2015. ImageNet large scale visual recognition challenge. *Int J Comput Vis* 115(3). 211–252. doi:10.1007/s11263-015-0816-y. https://doi.org/10.1007%2Fs11263-015-0816-y.

Schölkopf, Bernhard, Ralf Herbrich & Alex J. Smola. 2001. A generalized representer

theorem. In *Lecture notes in computer science*, 416–426. Springer Berlin Heidelberg. doi:10.1007/3-540-44581-1_27. https://doi.org/10.1007%2F3-540-44581-1_27.

Steinwart, Ingo & Andreas Christmann. 2008. *Support vector machines*. Springer New York. doi:10.1007/978-0-387-77242-4. https://doi.org/10.1007%2F978-0-387-77242-4.

Wendland, Holger. 2004. *Scattered data approximation*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press. doi:10.1017/CBO9780511617539.

Zhang, Chiyuan, Samy Bengio, Moritz Hardt, Benjamin Recht & Oriol Vinyals. 2016. Understanding deep learning requires rethinking generalization .