Bias-Variance Tradeoff, Overfitting and the Double Descent Curve

April 1, 2021

Main Papers

- *Reconciling modern machine-learning practice and the classical bias-variance trade-off*, Belkin et al
- *To undersand deep learning we need to understand kernel learning*, Belkin et al

General Aim

Given training sample

$$\{(x_1, y_1), ..., (x_n, y_n)\} \in \mathbb{R}^d \times \mathbb{R}$$

where (x_i, y_i) are i.i.d. variables drawn from probability distribution P, learn a predictor $h_n : \mathbb{R}^d \to \mathbb{R}$ that predicts y "well" given unseen x.

Loss function l

Minimize loss:
squared-loss function $l(y, \hat{y}) = (\hat{y} - y)^2$
0-1 loss/ classification loss: $l(\hat{y}, y) = 1_{\hat{y} \neq y}$

Generalization

- Find h_n that performs well on unseen data.
- Minimize true risk: $h^*(x) = \arg\min_h \mathbb{E}[\ell(h(x), y)]$ where (x, y) drawn independently from P.
- Empirical Risk Minimization (ERM) goal: Minimize training risk. $L_{emp}(h_n) = \frac{1}{n}\sum_{i=1}^{n} l(h_n(x_i), y_i)$.

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

- $\mathcal{H}$ is a function class that contains functions approximating h^*.
- Classically, we do not choose function h_n that reduces the empirical loss to near zero values, typically due to bounds on the generalization gap.

Bias Variance Tradeoff

- Finding a balance between underfitting and overfitting.
- "Bias-Variance Tradeoff"
- 0 training error does not tend to generalize well.

Generalization Error

Difference Between empirical and expected classifier loss:
$\mathbb{E}_{x,y}[l(\hat{h}_n(x), y)] \leq L_{emp}(\hat{h}_n) + O(\sqrt{c/n})$
where c is some measure of the complexity of $\mathcal{H}$, for example the fat-shattering dimension, VC-dimension, etc.

- To control L_{emp} and c, control $\mathcal{H}$ implicitly or explicitly.
- Examples: Changing NN architectures, regularization, early stopping.

Modern practice

- Modern ML methods such as large neural networks and other non-linear predictors have very low to no training risk
- NN architectures chosen such that interpolation can be achieved.
- Works even when training data have high levels of noise.

# params	random crop	weight decay	train accuracy	test accuracy
1,649,402	yes	yes	100.0	89.05
	yes	no	100.0	89.31
	no	yes	100.0	86.03
	no	no	100.0	85.75

Understanding Deep Learning Requires Rethinking Generalization, Zhang et al

Question

- Unanswered as to why these overparameterized data do not seem to cause high test loss due to overfitting.
- 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.

Short introduction to Kernels and RKHS

- Where $\mathbb{K}$ may refer to either $\mathbb{R}$ or $\mathbb{C}$.
- For kernel $k : X \times X \to \mathbb{K}$, there exists $\mathbb{K}$-Hilbert space $\mathcal{H}$ and map $\psi : X \to \mathcal{H}$ such that for all $x_1, x_2 \in X$,

$$k(x_1, x_2) = \langle \psi(x_2), \psi(x_1) \rangle.$$

- ψ is called a feature map of k.
- (Real) Gaussian RBF kernel with width γ:

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

where $x, x' \in \mathbb{R}^d$.

Positive Definite Functions

For a non-empty set X, a function $k : X \times X \to \mathbb{R}$ is said to be a positive definite if, for all $x_1, ..., x_n \in X$, for all $a_1, ..., a_n \in \mathbb{R}$, we have:

$$\sum_{j=1}^{n} \sum_{i=1}^{n} a_j a_i k(x_j, x_i) \geq 0.$$

Positive Definite Symmetric Functions are Kernels

A real function $k : X \times X \to \mathbb{R}$ is a kernel if and only if k is a positive definite symmetric function.

Short introduction to Kernels and RKHS

Reproducing kernel Hilbert spaces have many applications in the fields such as Statistical Learning and complex analysis. An RKHS is a K-Hilbert function space where point evaluation is continuous linear funcitonal.

Reproducing Kernel Hilbert Spaces (**RKHS**) Definition

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.

Reproducing Kernels

For a non-empty set X and a function $k : X \times X \to \mathbb{K}$ k is called a reproducing kernel of $\mathcal{H}$ (a Hilbert function space) if $k(\cdot, x) \in \mathcal{H}$ for all $x \in X$ and the following property hold for all $x \in X$ and $f \in \mathcal{H}$:

$$f(x) = \langle f, k(\cdot, x) \rangle$$

This condition is also known as the reproducing property.

- It can be shown that k is a kernel.
- $\mathcal{H}$ is then an RKHS.

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 Scholkopf et al.

Representer Theorem

Training data $(x_1, y_1), ...(x_n, y_n) \in X \times \mathbb{R}$, and RKHS $\mathcal{H}$ a $\mathbb{R}$-Hilbert function space over X with reproducing kernel $k : X \times X \to \mathbb{R}$.
Let g be a strictly increasing function $g : [0, \infty] \to \mathbb{R}$, and $I : (X \times \mathbb{R}^2)^n \to \mathbb{R} \cup \{\infty\}$ be an arbitrary loss function.
We want to minimize the following empirical risk term:
$L_{emp}(f, (x_1, y_1), ..., (x_n, y_n)) := I((x_1, y_1, f(x_1)), ..., (x_n, y_n, f(x_n))) + g(\|f\|)$
For $\hat{f} = \arg\min_{f \in \mathcal{H}} L_{emp}(f, (x_1, y_1), ..., (x_n, y_n))$, $\hat{f}$ 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 $1 \leq i \leq n$.

Proof Sketch

- Let $\Phi(x)(\cdot) = k(\cdot, x)$.
- $f = \sum_{i=1}^{n} a_i \Phi(x_i) + \gamma$, where $\langle \Phi(x_i), \gamma \rangle = 0$
- $f(x_j) = \sum_{i=1}^{n} a_i \langle \Phi(x_i), \Phi(x_j) \rangle$
- So $f(x_j)$ is unaffected by γ.
- $\|f\|^2 = \|\sum_{i=1}^{n} a_i \Phi(x_i) + \gamma\|^2 \geq \|\sum_{i=1}^{n} a_i \Phi(x_i)\|^2$
- $g(\|f\|) \geq g(\|\sum_{i=1}^{n} a_i \Phi(x_i)\|)$
- So $\gamma = 0$ and $\hat{f} = \sum_{i=1}^{n} a_i k(\cdot, x_i)$

Overfitted and Interpolated Kernel Classifiers

- Performed in Belkin et al. (2018b) that show the strong generalization performance also appears in kernel classifiers.
- The aim is to have interpolated solutions which fits the data perfectly, thus having no regularization.
- Though no finite number of functions in the RKHS are able to fit the training data, minimum norm RKHS solutions can be obtained using Representer Theorem.

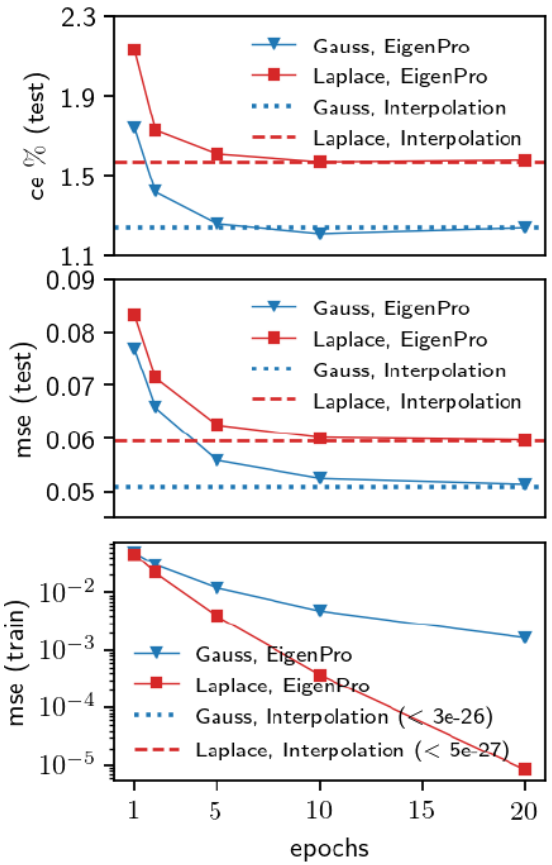
Training

- Training datapoints $(x_1.y_1), ..., (x_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$.
- Let $\mathcal{H}$ denote the RKHS corresponding to the kernel k.
- Let f^* denote the minimum norm interpolant given the datapoints.
- $f^*(\cdot) = \sum_{i=1}^n a_i^* k(x_i, \cdot)$.
- Due to its interpolating properties, we know that for $a^* := (a_1^*, ..., a_n^*)^\mathsf{T}$.
- $\|f^*\|_\mathcal{H} = \sum_{i,j=1}^n a_i^* a_j^* k(x_i, x_j) = a^{*\mathsf{T}} K a^*$
- $a^* = \arg\min_a \sum_i l((\sum_j a_i k(x_j, x_i)), y_i)$ for some convex loss l.

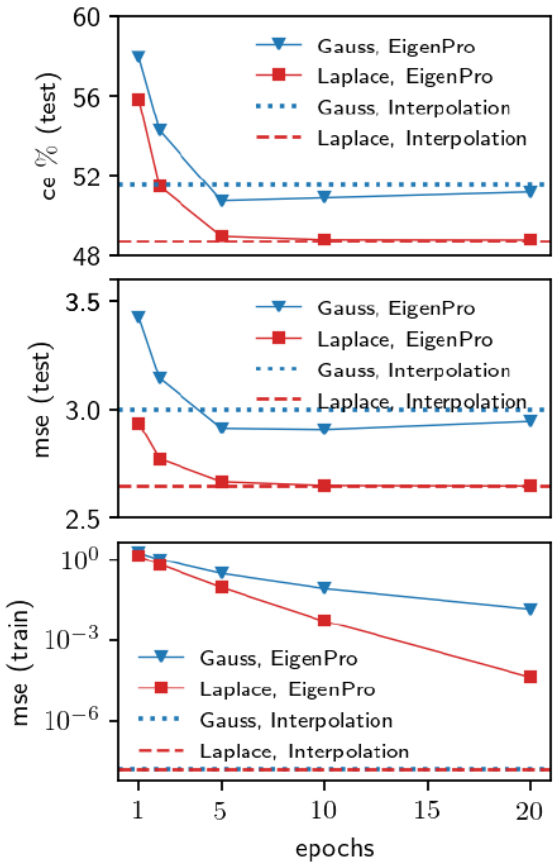
Results

- Models were trained till mean square loss on training dataset approaches zero.
- The interpolated solution performed close to optimal.
- The benefits of early stopping regularization were little to none in terms of test regression or classification error.

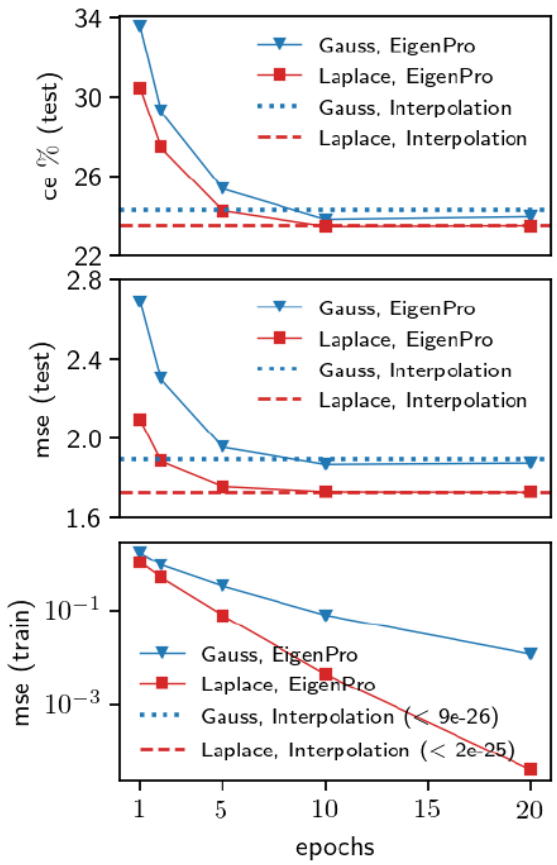
Interpolated Kernels Performance



(a) MNIST



(b) CIFAR-10



(c) SVHN ($2 \cdot 10^4$ subsamples)

Existing Bounds

- How the norm of the classifiers change w.r.t. data size.
- classifiers are called overfitted if the classification loss for the training set is 0 or near 0, and classifiers are called interpolated when the mean square error is 0 or near 0.

t-overfitting

For a function $h \in \mathcal{H}$, and $t \in \mathbb{R}^+$, it **t-overfits** the data if it is overfitted with 0 classification loss, and $0 < t < y_i h(x_i)$ for $1 \leq i \leq n$.

Introduced to prevent arbitrary scaling of solution.

Bounds of t-overfitters

Let $(x_1, y_1), ..., (x_n, x_n) \in \Omega \times \{-1, 1\}$ be data sampled from probability distribution P, and y is not a deterministic function of x. With high probability, for any h that t-overfits the data, there exists some constants $A, B > 0$ depending on t, that satisfies

$$\|h\|_{\mathcal{H}} > Ae^{Bn^{1/d}}.$$

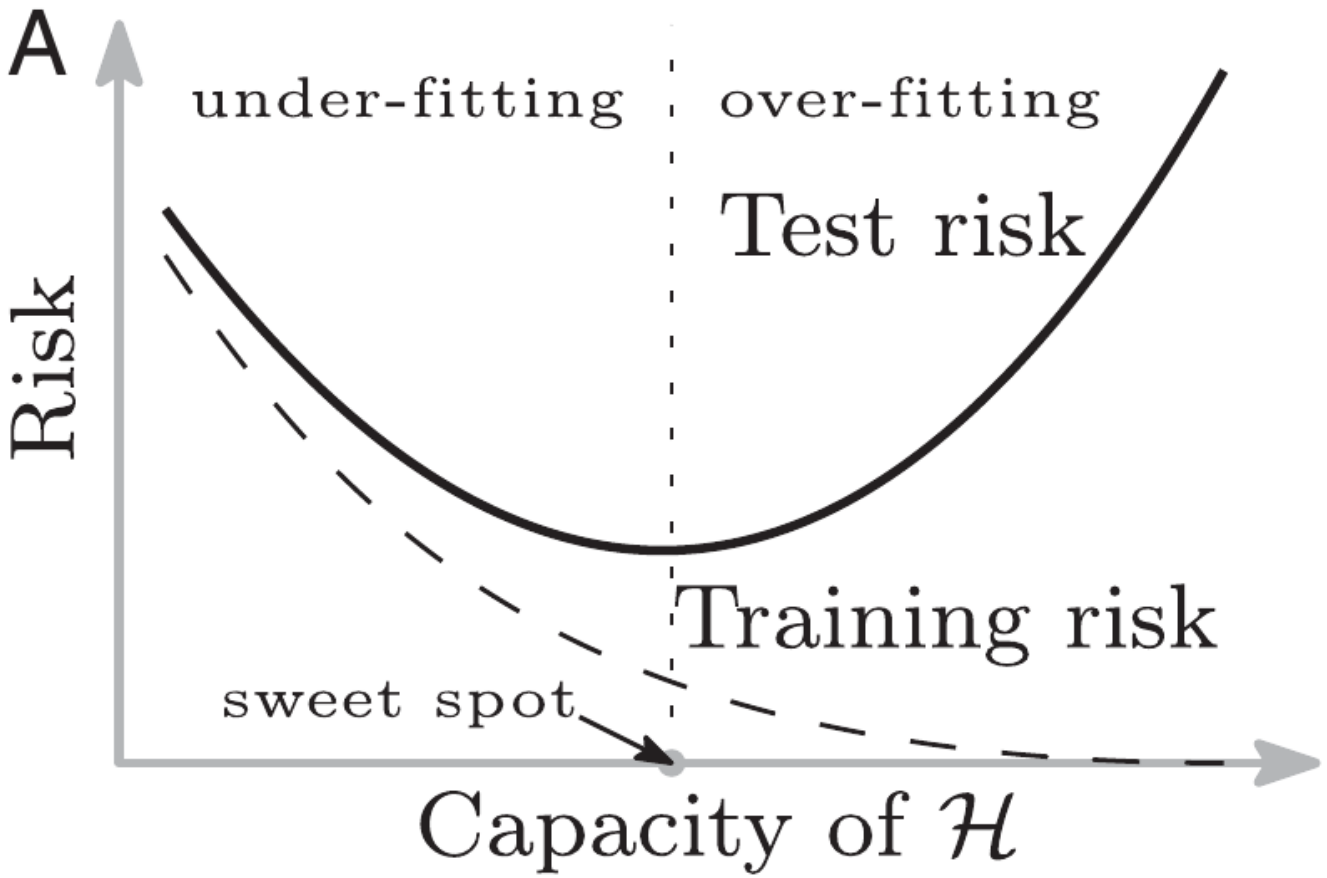
Classical Bounds

Classical bounds for kernel methods are in the form:

$$|\frac{1}{n}\sum_{i=1}^{n} l(f(x_i), y_i) - L(f)| \leq C_1 \frac{\|f\|_{\mathcal{H}}^a}{n^b}, \quad C_1, a, b \geq 0$$

The right side on this will tend to infinity for bigger $\|f\|_{\mathcal{H}}$.

Classical Descent Curve

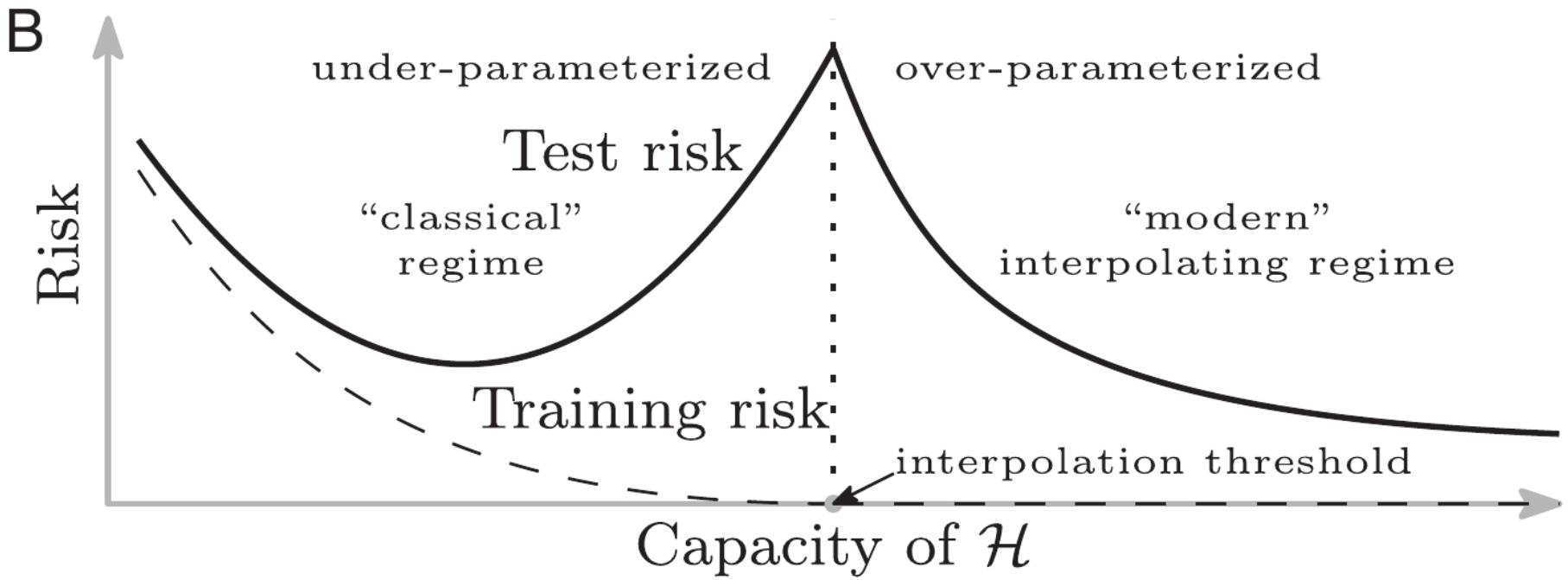


Classical curve from bias variance tradeoff.

"Double Descent"

- "Double Descent" curve proposed and empirically observed to some extent.
- Curve the extends beyond the point of interpolation
- Risk decreases beyond this point, typically surpassing performance of classical stopping point.

Double Descent Curve



Random Fourier Features

We want to find a randomized feature map $z : \mathbb{R}^d \to \mathbb{R}^D$ such that: $k(x, y) = \langle \phi(x), \phi(y) \rangle \approx \langle z(x), z(y) \rangle$.

(Bochner Rudin (1990)).For a continuous (properly scaled) 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.

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

Random Fourier Features

Let the function class $\mathcal{H}_\infty$ be the Reproducing Kernel Hilbert Space (RKHS) corresponding to the Gaussian kernel.
We consider the following non-linear parametric model:

Random Fourier Features (RFF)

Random Features for Large-Scale Kernel Machines (Rahimi et al)

Let the function class $\mathcal{H}_N$ consist of functions $h_n : \mathbb{R}^d \to \mathbb{C}$ of the form:

$$h(\cdot) = \sum_{k=1}^{N} a_k \phi(\cdot, v_k)$$

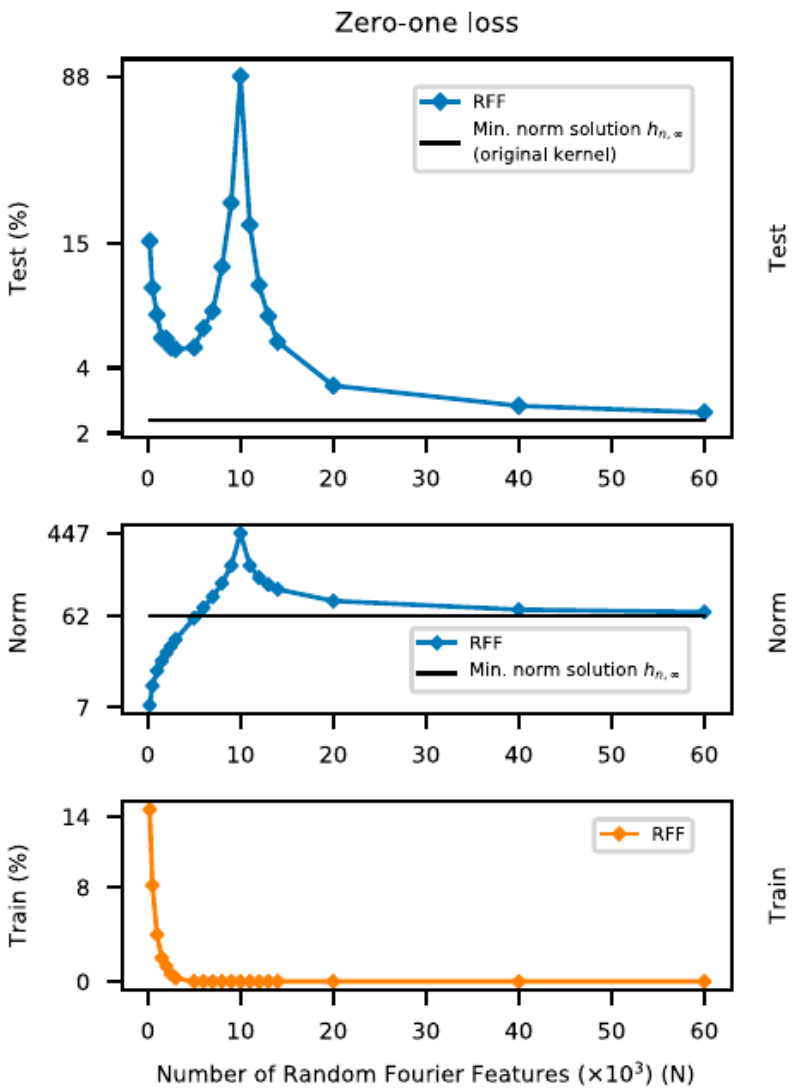
where $\phi(\cdot, v_k) := e^{i\langle v_k, \cdot \rangle}$ vectors $v_1, ..., v_N$ are sampled independently from the standard normal distribution in $\mathbb{R}^d$.

H_N has N parameters in $\mathbb{C}$, $\{ a_1, ..., a_N \}$.
As $N \to \infty$, H_N becomes a closer approximation to $\mathcal{H}_\infty$

Empirical Evidence - Learning Procedure

- Given training sample $(x_1, y_1), ..., (x_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$.
- Minimize empirical risk: $\frac{1}{n} \sum_{j=1}^{n} (h(x_j) - y_j)^2$ for $h \in \mathcal{H}_N$.
- When minimizer not unique ($N > n$), choose the minimizer with the coefficients $(a_1, ..., a_N)$ that have the smallest ℓ_2 norm.
- Let this predictor be: $h_{n,N} \in \mathcal{H}_N$.

Empirical Evidence - Results



Explanations on Double Descent

Why should the test risk decrease even when empirical risk stays the same?

Classical Bounds

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

- The inductive bias of a learning algorithm is the set of assumptions that the learner uses to predict outputs of given inputs that it has not encountered
- Capacity of function class needs not suit the appropriate inductive bias for the problem.
- By having a larger function class, may find a function that matches the inductive bias better. Eg., smoother function, smaller norm, larger margin.
- In this case, a smaller norm.

Min Norm RKHS solution $h_{n,\infty}$

Fill Distance

The fill distance for a set of points $X = \{x_1, ..., x_N\} \subseteq \Omega$ for a bounded domain Ω is defined to be

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

Approximation theorem

For $\mathcal{H}$ an RKHS corresponding to the Gaussian RBF kernel, fix $h^* \in \mathcal{H}$. 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 $h \in \mathcal{H}$ with high probability

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

Proof Sketch

Theorem (Wendland (2004))

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 $r \geq 0$ and $l \geq l_0$, $|f^{(l)}(r)| \leq l! M^l$. Then there exists a constant $c > 0$ such that the error between a function $f \in \mathcal{H}$ (where $\mathcal{H}$ is the RKHS corresponding to the kernel k), and its interpolant $s_{f,X}$ for all data points $X = \{x_1, ..., x_n\}$ can be bounded by:

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

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

- Let $f := h - h^*$. So $\|f\|_{\mathcal{H}} \leq \|h\|_{\mathcal{H}} + \|h^*\|_{\mathcal{H}}$.
- Use the result (Belkin et al (2019)) that for the points from the compact cube Ω, the fill distance has a high probability to be $O(n/\log n)^{-1/d}$.

Other Rationale

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 $\|f\|_{\mathcal{H}}$.

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

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.