Reconciling modern machine learning practice and the bias-variance trade-of

J. Wilder Lavington

December 4, 2019

University of British Columbia: Department of Computer Science

General Overview (TLDR)

1. Classic statistical modeling assumptions don't hold for high capacity model regimes.
2. Not only do these high capacity models reach performance comparable to the "sweet spot" found in classical machine learning, but often achieves even better generalization performance.
3. This performance is correlated with high capacity, interpolating models, that yield low RKHS norm solutions in cases where we arrive at an RKHS model class in the limit of the model class capacity.
4. This phenomena seems to also hold for other classes of parametric models with adjustable model capacity such as random forests.

Problem Statement

General Framework

Given a sample of training examples $(x_1, y_1), ..., (x_n, y_n)$ from $\mathbb{R}^d \times \mathbb{R}$, we learn a predictor $h_n : \mathbb{R}^d \to \mathbb{R}$ that is used to predict the label of a new data point x.

Problems:

- Binary Classification
- Multi-Class Classification

Model $h_n \in \mathcal{H}$:

- Kernel Machines
- Neural Networks
- Random Forests
- Decision Trees
- Boosting algorithms

Empirical Risk Minimization

How do we Choose h_n?

The model h_n chosen from $\mathcal{H}$ is determined using empirical risk minimization (ERM).

What is empirical risk minimization?

The predictor h is taken to be a function $h \in \mathcal{H}$ that minimizes the empirical risk $\frac{1}{n}\sum_{i=1}^{n} l(h(x_i, y_i))$ where l is a loss function such as squared loss for regression, or zero one loss for classification.

What is a desirable h?

We hope that the model h not only minimizes the empirical risk, but but the test risk as well.

Classical Machine Learning Wisdom

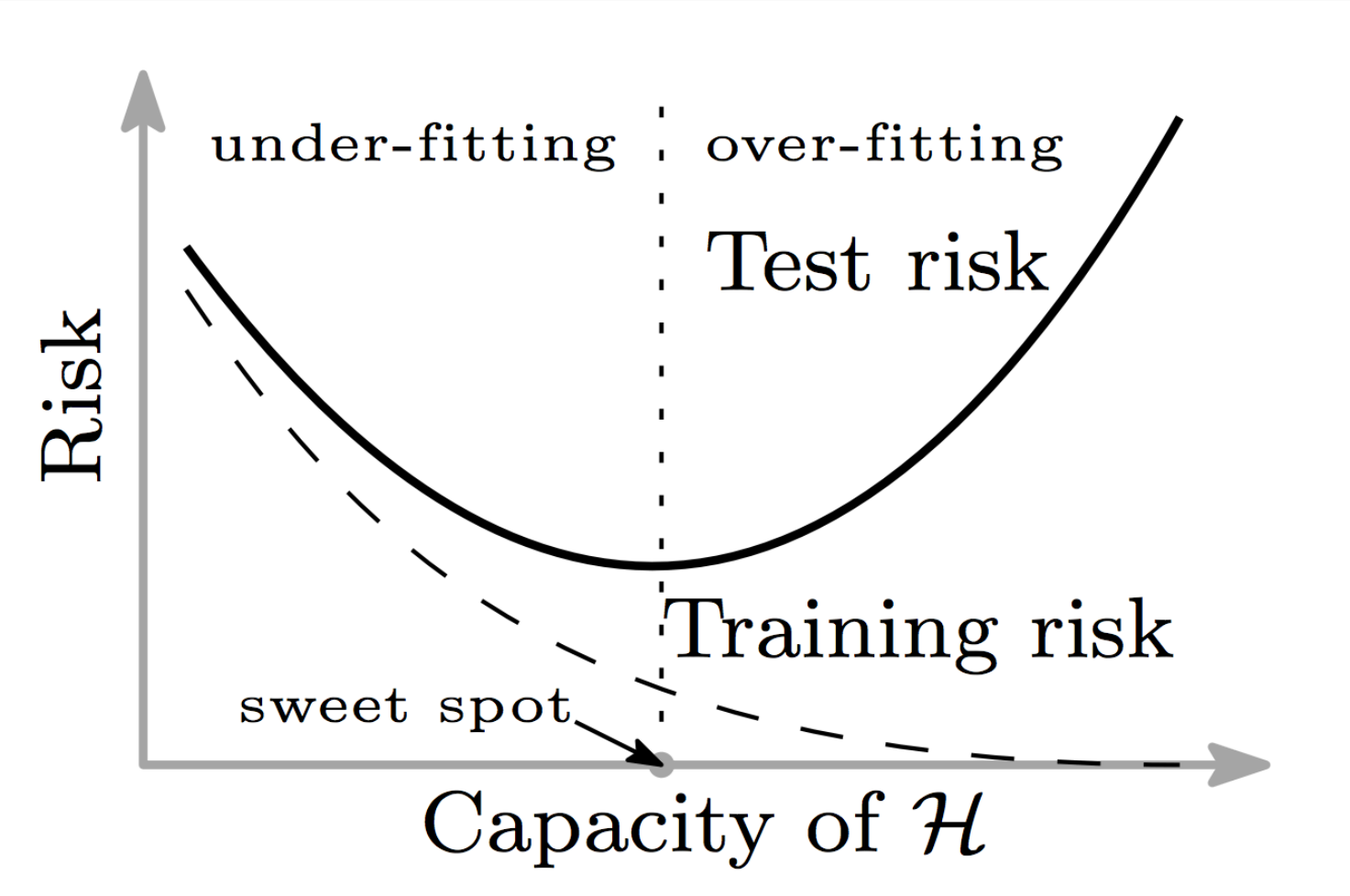


Figure 1: Classically, we controlled the models capacity to ensure $\mathcal{H}$ was not too large so as to lead to over fitting, and not too small to lead to under fitting.

Inductive Bias

Definition: Inductive bias is the set of assumptions a learner uses to predict results given inputs it has not yet encountered.

Classical Inductive Bias: Classical models carry an implicit inductive bias towards simpler models that restricts the space of functions that we can realize.

Yet increasingly, more complicated models out perform simpler ones!

The Double Decent Curve

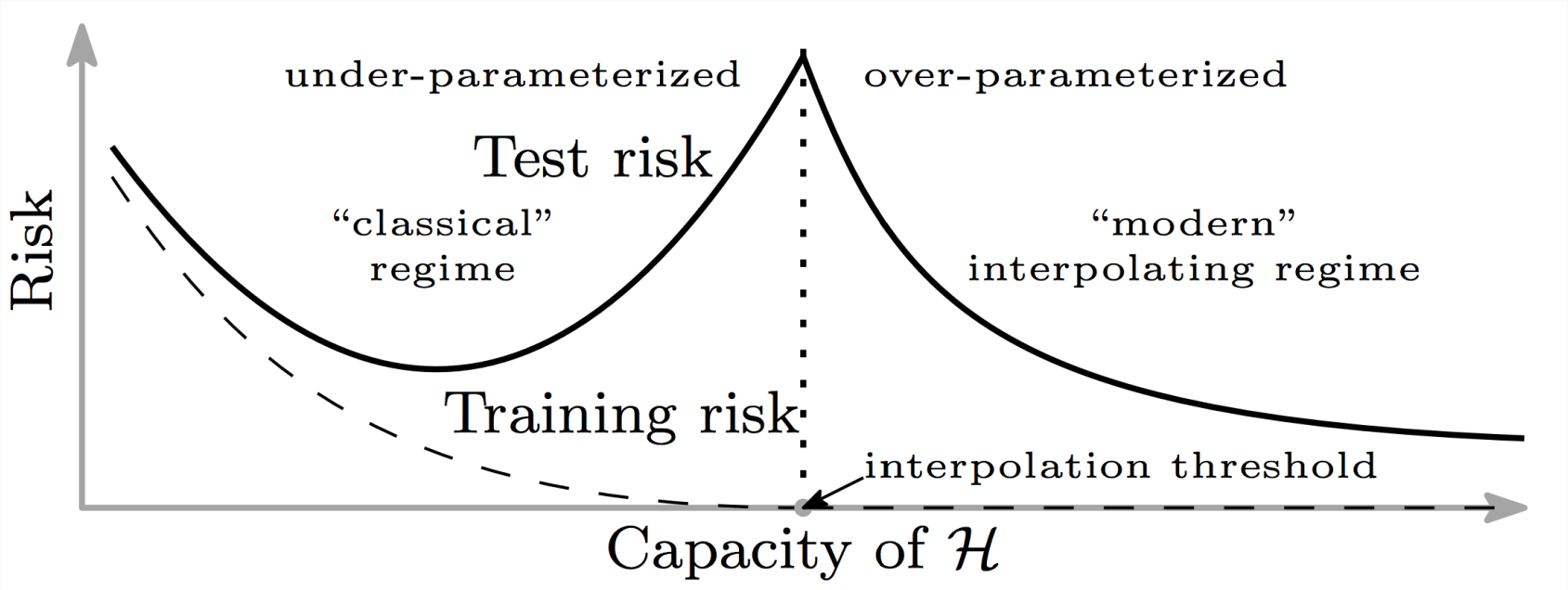


Figure 2: While in classical regimes we see that using a inductive bias towards simple models does allow us to reach a "sweet spot" with respect to test risk, however in the interpolating regime, we can often do better by further increasing the model capacity.

Smooth Functions as an Inductive Bias

Why is this possible?

While the classic assumption of low capacity can improve performance on test data, it is does not necessarily lead to the correct inductive bias for the problem at hand.

What is the correct inductive bias?

Belkin et. al. posits that the appropriate inductive bias is connected to the regularity or smoothness of a function under some function space norm.

Larger Function Classes and Smoothness

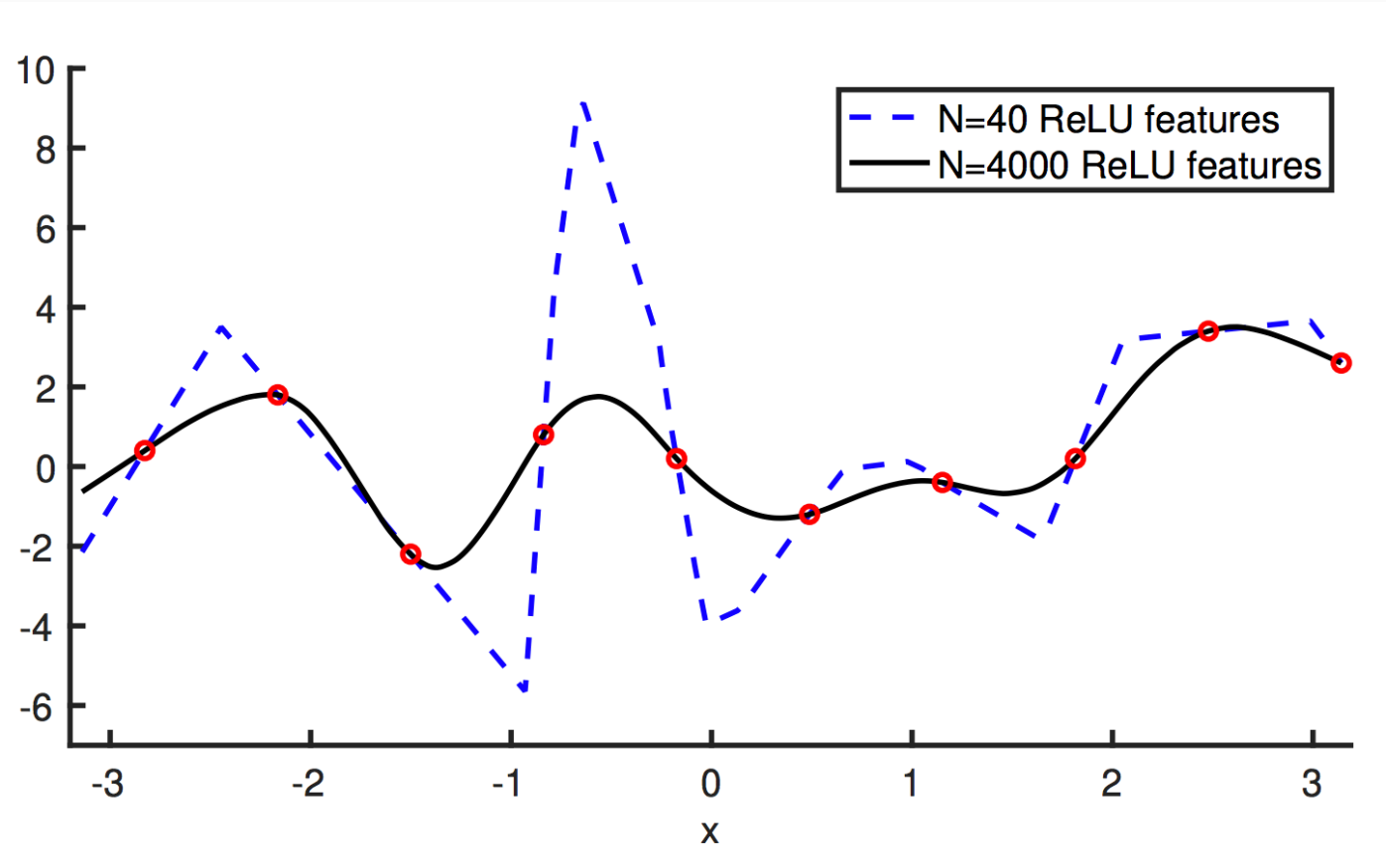


Figure 3: By considering larger function classes which contain more candidate predictors, we are able, for whatever reason to find interpolating functions that have smaller norms and are thus, smoother.

The Reproducing Kernal Hilbert Space (RKHS)

THE UNIVERSITY OF BRITISH COLUMBIA

What is RKHS?

The RKHS is a set of functions that satisfy specific properties when evaluated as inner products with kernels that are also from the RKHS.

What is a low norm RKHS solution?

A function drawn from the RKHS that when evaluated as an inner product with itself has a small magnitude.

Why do we want to use this metric?

Low RKHS norm solutions represent solutions considered over a class of functions that are as “simple” as possible. This as we will see is a useful inductive bias when we are already complex enough to interpolate.

Why is this confusing?

Because it represents a norm over function spaces instead of vector spaces, so our usual notion of closeness doesn’t quite work.

Data-set Information

THE UNIVERSITY OF BRITISH COLUMBIA

Dataset	Size of full training set	Feature dimension (d)	Number of classes
CIFAR-10	$5 \cdot 10^4$	1024	10
MNIST	$6 \cdot 10^4$	784	10
SVHN	$7.3 \cdot 10^4$	1024	10
TIMIT	$1.1 \cdot 10^6$	440	48
20-Newsgroups	$1.6 \cdot 10^4$	100	20

Random Fourier Features (RFF)

The RFF model family $\mathcal{H}_N$ with N complex valued parameters consists of functions $h : \mathbb{R}^d \to \mathbb{C}$ of the form:

$$h(x) = \sum_{k=1}^{N} a_k \phi(x, v_k) \quad \text{where } \phi(x, v) := e^{\sqrt{-1}\langle v, x \rangle} \tag{1}$$

Where the vectors v are sampled from a standard multivariate normal distribution.

Note: As $N \to \infty$, the function class becomes closer and closer to approximation to the RKHS ($\mathcal{H}_\infty$)

RFF Experiment Setup

1. Given data $(x_1, y_1), ..., (x_N, y_N)$ from $\mathbb{R}^d \times \mathbb{R}$, they find a predictor $h_{n,N} \in \mathcal{H}_N$ that minimizes MSE over all functions $h \in \mathcal{H}_N$.
2. When the mimizer is not unique (as is the case for the interpolation regime), they use the minimizer who's coefficients have the lowest l2 norm.
3. These minimizers are can be solved for directly.

Note: This norm is used as a surrogate for $\mathcal{H}_\infty$, which for general functions is difficult to compute.

RFF Results: Part 1

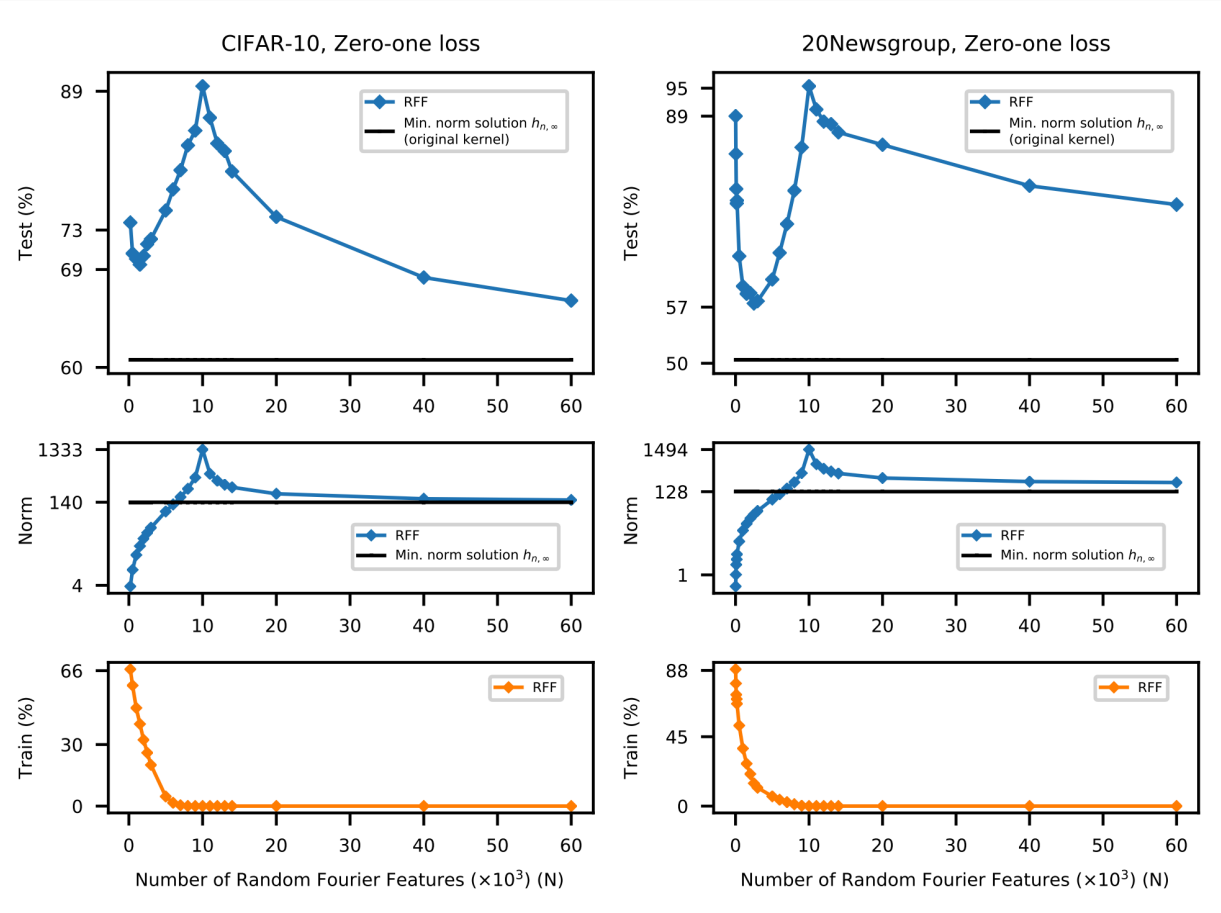


Figure 4: Dual decent curves for the RFF model on the CIFAR and 20Newsgroup data sets.

RFF Results: Part 2

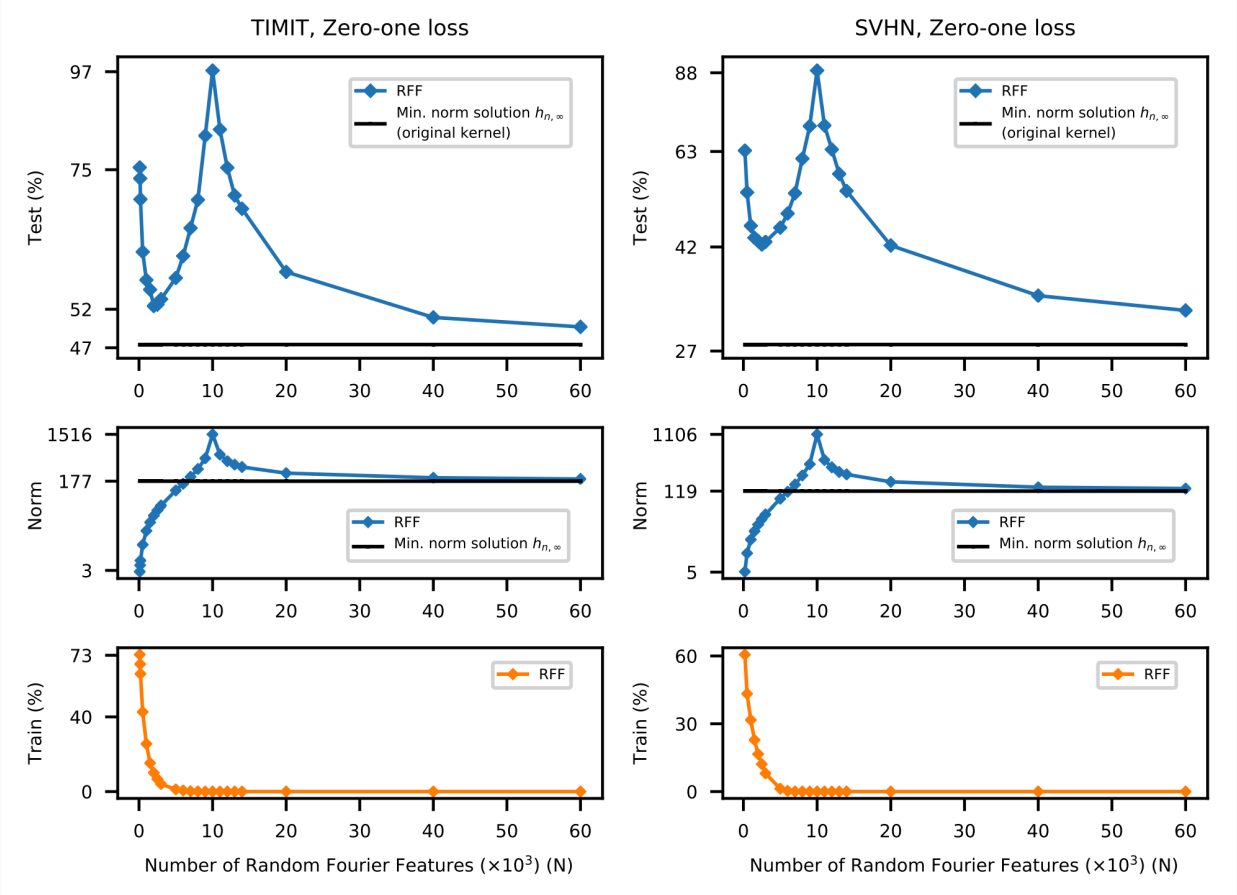


Figure 5: Dual decent curves for the RFF model on the TIMIT and SVHN data sets.

RFF Results: Part 3

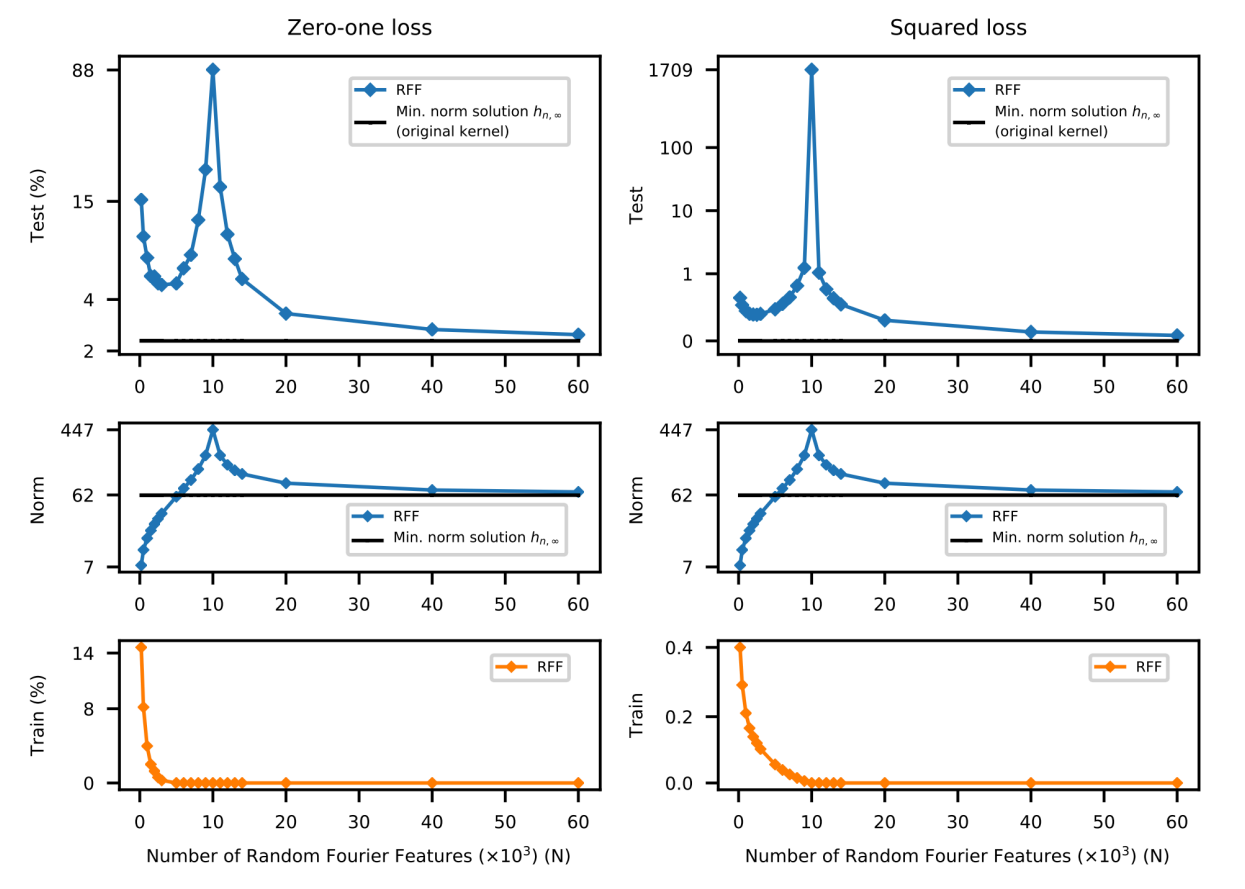


Figure 6: Dual decent curves for the RFF model on the MNIST data set.

Random Relu Features (RRF)

The RFF model family $\mathcal{H}_N$ with N real valued parameters consists of functions $h : \mathbb{R}^d \to \mathbb{R}$ of the form:

$$h(x) = \sum_{k=1}^{N} a_k \phi(x, v_k) \quad \text{where } \phi(x, v) := \min(0, \langle v, x \rangle) \tag{2}$$

Where the vectors v are sampled from a uniform distribution over the surface of the unit sphere in $\mathbb{R}^d$.

Note: Again, as $N \to \infty$, the function class becomes closer and closer to approximation to the RKHS ($\mathcal{H}_\infty$)

RRF Results

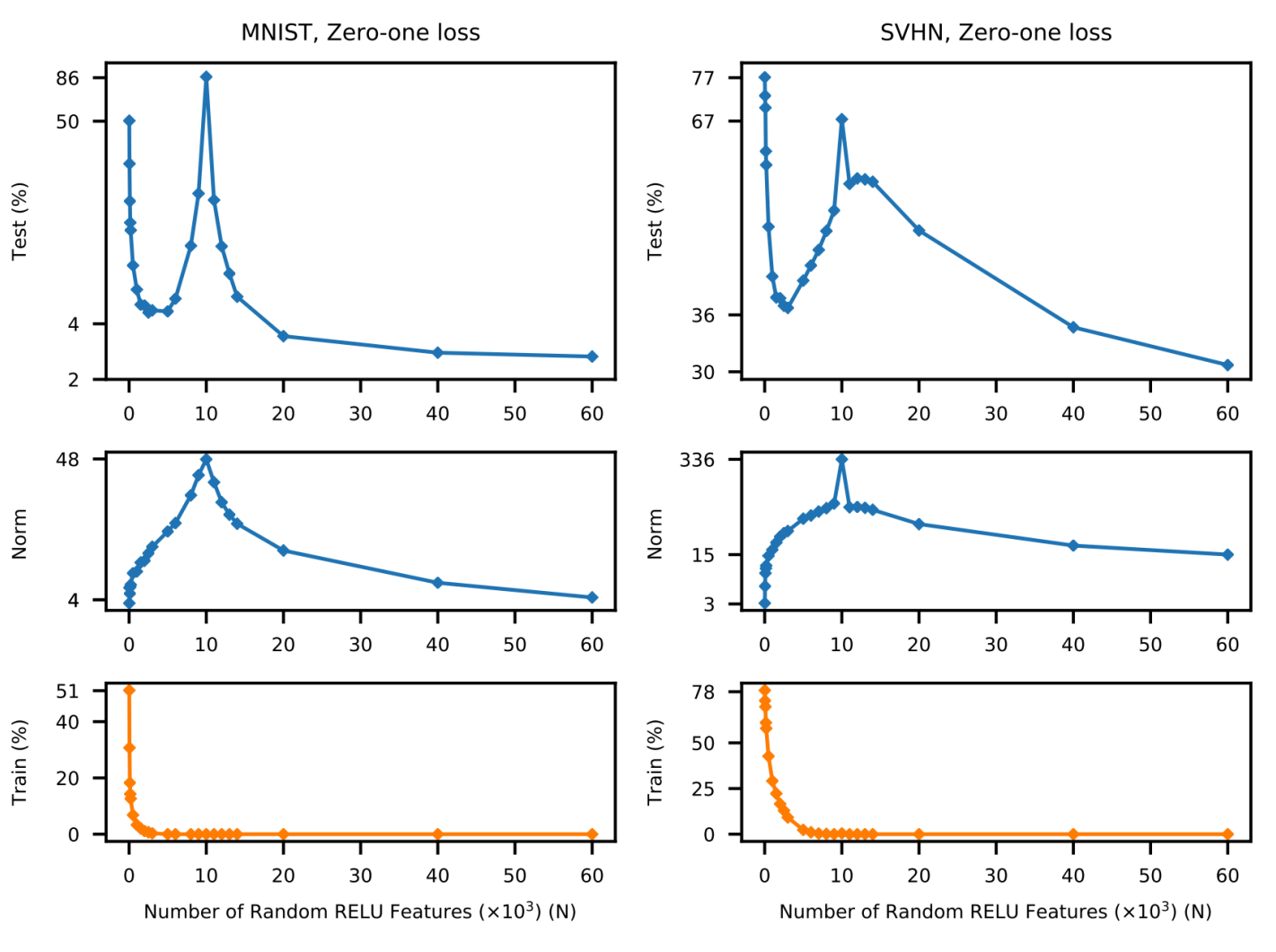
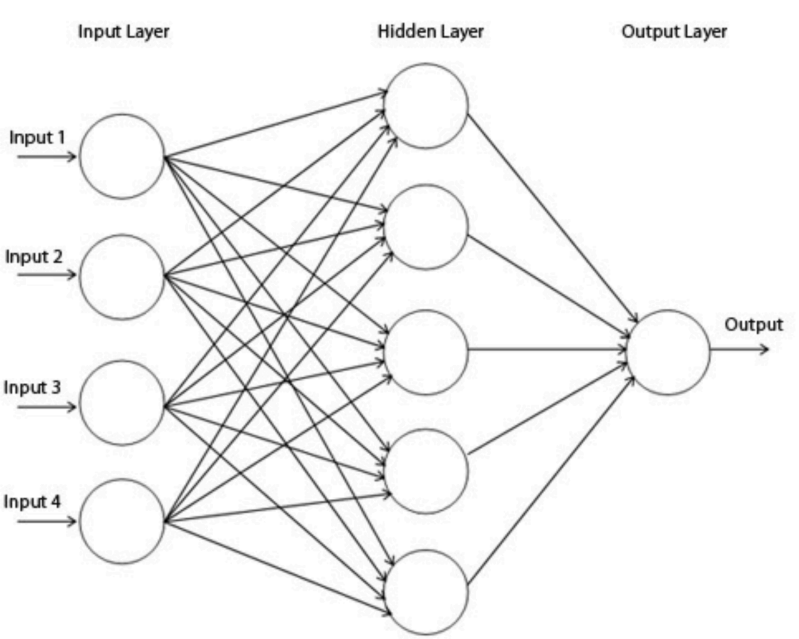


Figure 7: Dual decent curves for RRF model class on MNIST and SVHN.

Fully Connected Neural Networks

1. To alleviate sensitivity to initial conditions, they use a weight re-use scheme for the under parameterized regime, where parameters used for training smaller networks are used in initialization of progressively larger networks.
2. For over parameterized they use standard initialization.



Single layer hidden Neural Network

Single Layer Network Results: Part 2

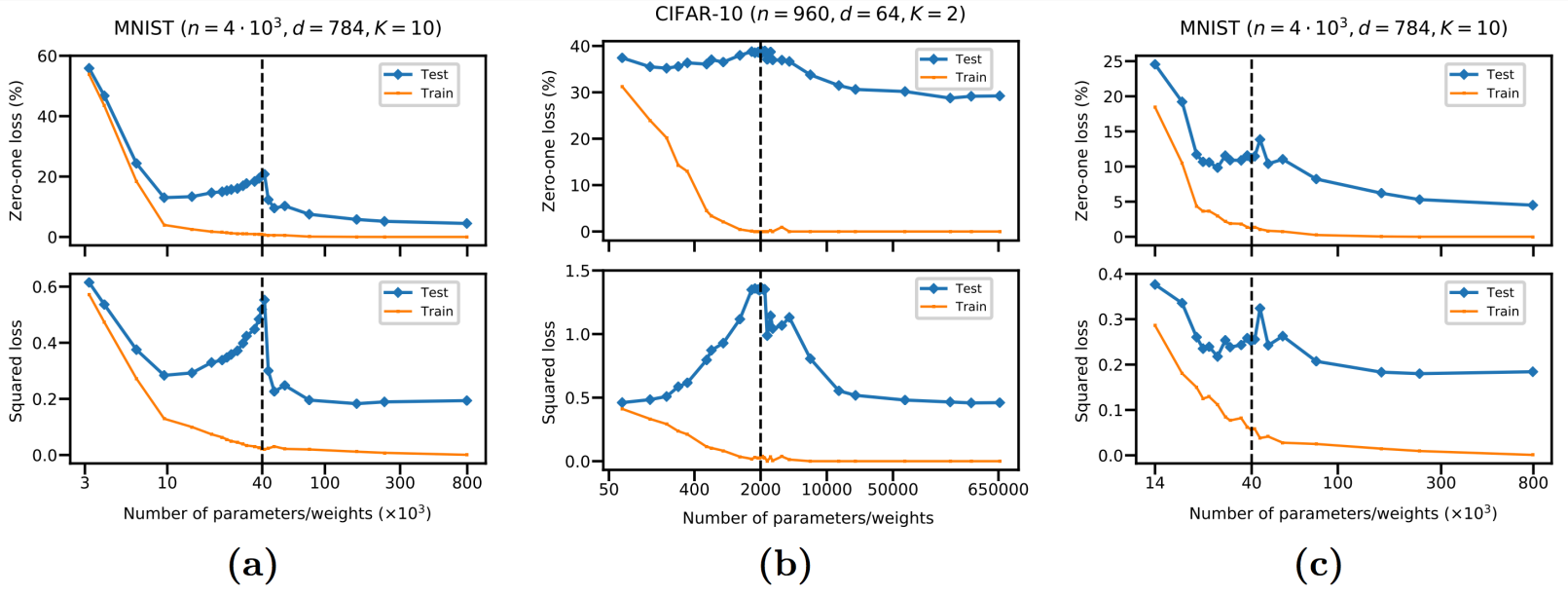
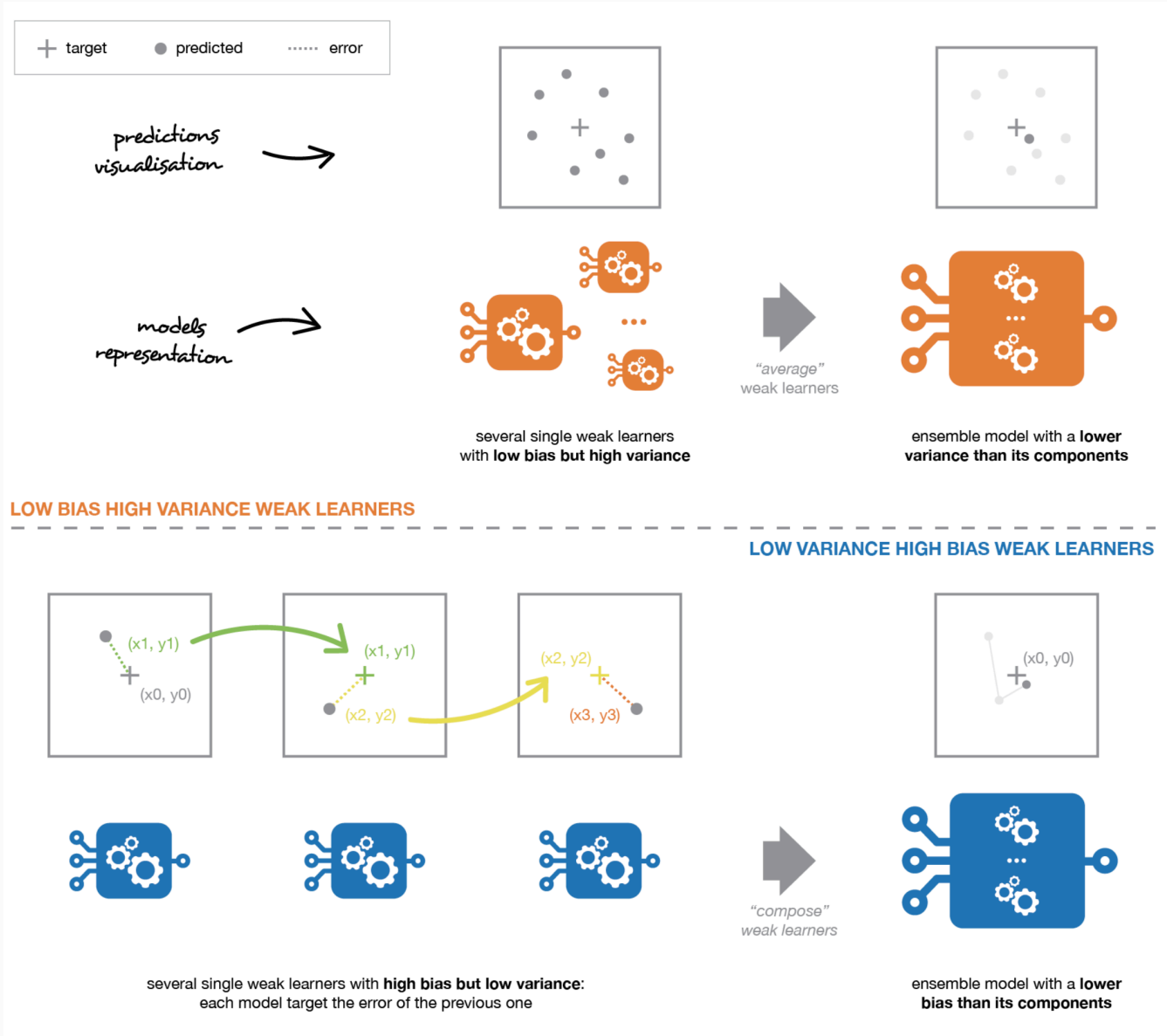


Figure 8: (a),(b) fully connected single layer neural network dual decent curves with weight reuse in non interpolating regime for CIFAR and SVHN. (c) Fully connected single layer neural network dual decent curves with no weight reuse in non interpolating regime for MNIST.

Bagging and Boosting



Experiment Setup

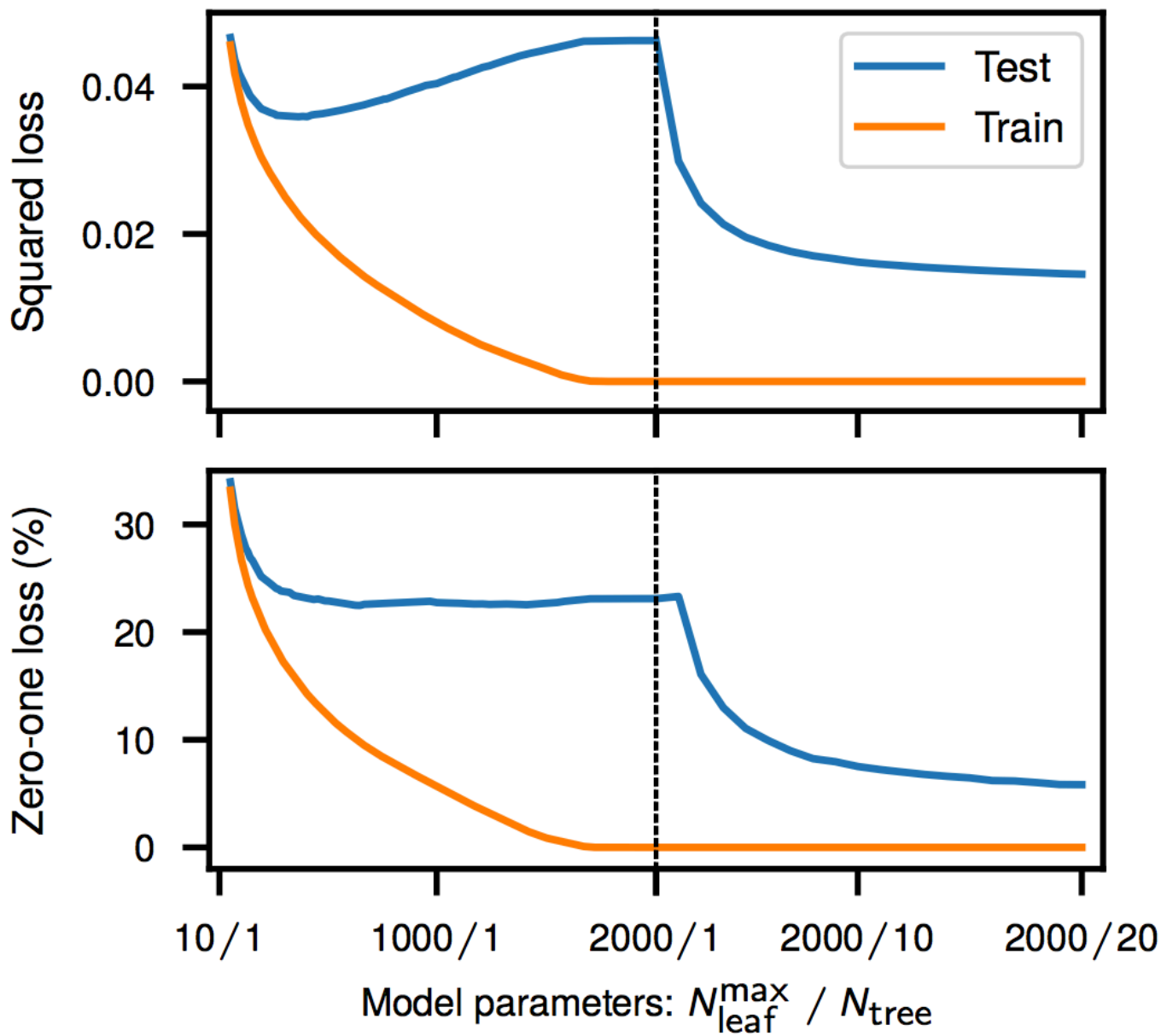
Random Forests (RF)

1. Model capacity is controlled by number of trees averaged, as well as the depth of each tree.
2. In the non-interpolating regime they increase both, in the interpolating regime they increase the averaging.

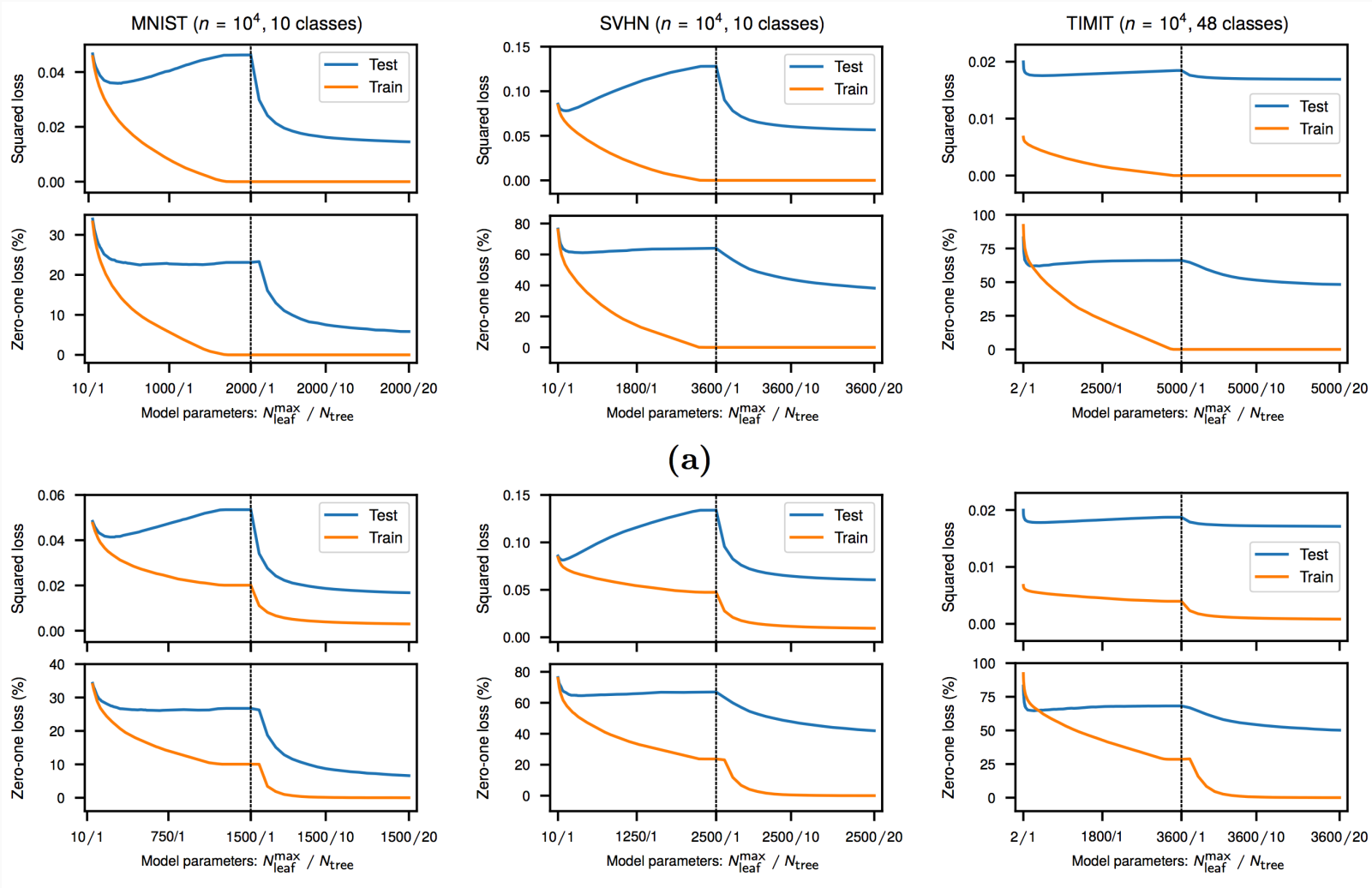
L2 Boosting

1. Constrain Each tree to have a small number of leaves (≤ 10)
2. The metric for capacity is the number of consecutive trees used in the boosting algorithm
3. They use low shrinkage to speed up interpolation
4. They then average runs over multiple runs of these boosted models to further increase capacity after interpolation.

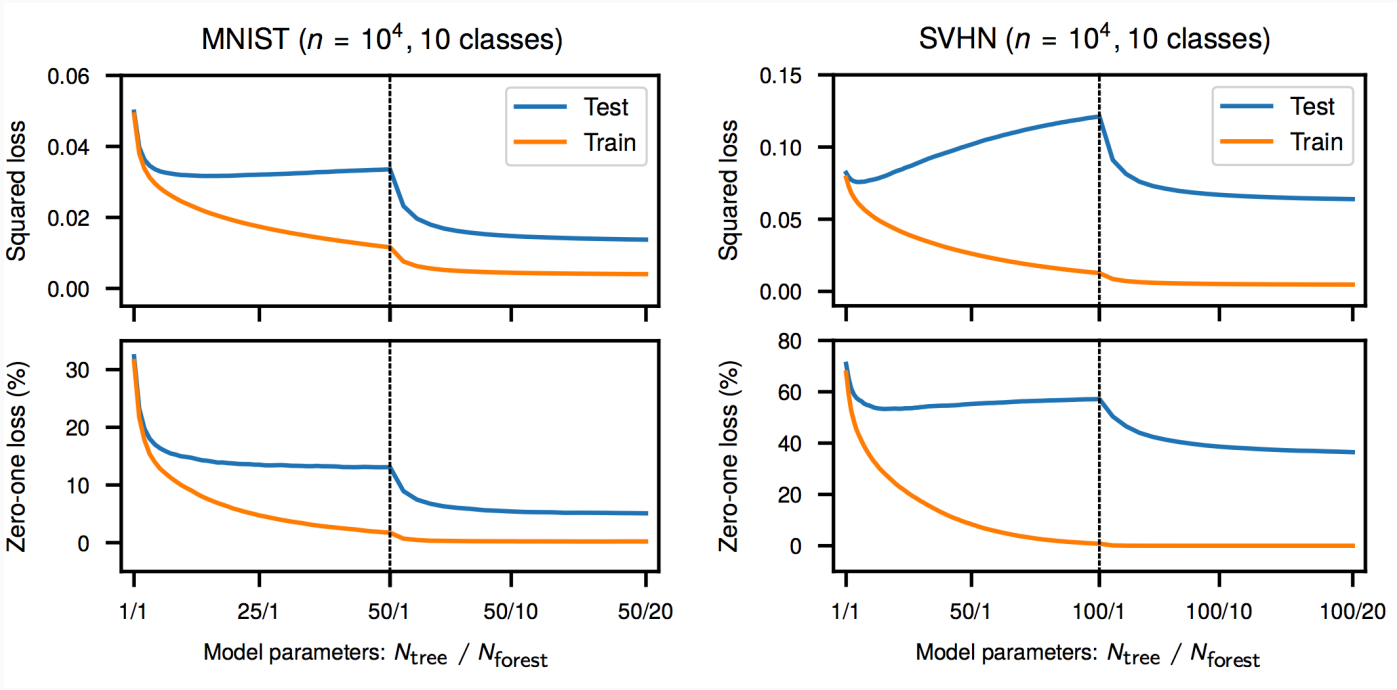
Results: RF MNIST



Results: RF (all datasets)



Results: L2 Boosting



Historical Absence

Why was this knowledge historically overlooked?

1. observing the double decent curve requires a parametric family of function spaces with elements that can achieve arbitrary complexity.
2. In non-parametric settings regularization is generally employed, and can limit the model capacity.
3. The computational advantage for kernel methods only holds in the non-interpolating regime.
4. The peak that is given in the dual decent curves is easy to miss in the multi-layer nueral networks case, as it can effectively be "missed" do to a higher starting model capacity.
5. In other cases, tricks like drop-out and early stopping can change the behavior of these models, indicating that the training methodology is just as important as the model itself (high capacity and low norm).

Small Norm Solutions and Optimization Considerations

A few closing comments:

1. Both the RFF and the RRF converge to the minimum functional norm solution in the RKHS.
2. This solution maximizes smoothness, subject to interpolation constraints.
3. For more general networks, it is still not clear that we are converging to a similar solution, however some work has been done showing that a similar inductive bias is present [5].
4. While an analysis similar to (1) isn't cited, its clear that averaging in RF and Boosting algorithms produces smoother classifiers as well, which also empirically perform better.

Approximation Theorem

Theorem 1. *Fix any $h^* \in \mathcal{H}_\infty$. Let $(x_1, y_1), \ldots, (x_n, y_n)$ be independent and identically distributed random variables, where x_i is drawn uniformly at random from a compact cube[2] $\Omega \subset \mathbb{R}^d$, and $y_i = h^*(x_i)$ for all i. There exists absolute constants $A, B > 0$ such that, for any interpolating $h \in \mathcal{H}_\infty$ (i.e., $h(x_i) = y_i$ for all i), so that with high probability*

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

Figure 9: The following theorem gives bounds on the error of approximating a function in the RKHS h^* with any other interpolating function h also found in the RKHS.

References i


M. Belkin, D. J. Hsu, and P. Mitra.
Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate.
In *Advances in Neural Information Processing Systems*, pages 2300–2311, 2018.


M. Belkin, S. Ma, and S. Mandal.
To understand deep learning we need to understand kernel learning.
arXiv preprint arXiv:1802.01396, 2018.


M. Belkin, A. Rakhlin, and A. B. Tsybakov.
Does data interpolation contradict statistical optimality?
arXiv preprint arXiv:1806.09471, 2018.

References ii

H. Daumé III.
From zero to reproducing kernel hilbert spaces in twelve pages or less.
University of Maryland, 2004.

S. Gunasekar, B. E. Woodworth, S. Bhojanapalli, B. Neyshabur, and N. Srebro.
Implicit regularization in matrix factorization.
In *Advances in Neural Information Processing Systems*, pages 6151–6159, 2017.

A. Montanari, F. Ruan, Y. Sohn, and J. Yan.
The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime.
arXiv preprint arXiv:1911.01544, 2019.

References iii

C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals.
Understanding deep learning requires rethinking generalization.
arXiv preprint arXiv:1611.03530, 2016.

Hilbert Spaces

Definition (Inner product)

Let $\mathcal{H}$ be a vector space over $\mathbb{R}$. A function $\langle \cdot, \cdot \rangle_{\mathcal{H}} : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ is an inner product on $\mathcal{H}$ if

1. Linear: $\langle \alpha_1 f_1 + \alpha_2 f_2, g \rangle_{\mathcal{H}} = \alpha_1 \langle f_1, g \rangle_{\mathcal{H}} + \alpha_2 \langle f_2, g \rangle_{\mathcal{H}}$
2. Symmetric: $\langle f, g \rangle_{\mathcal{H}} = \langle g, f \rangle_{\mathcal{H}}$
3. $\langle f, f \rangle_{\mathcal{H}} \geq 0$ and $\langle f, f \rangle_{\mathcal{H}} = 0$ if and only if $f = 0$.

Norm induced by the inner product: $\|f\|_{\mathcal{H}} := \sqrt{\langle f, f \rangle_{\mathcal{H}}}$

Definition (Hilbert space)

Inner product space containing Cauchy sequence limits.

Reproducing Kernal Hilbert Spaces (RKHS)

Definition

$\mathcal{H}$ a Hilbert space of $\mathbb{R}$-valued functions on non-empty set $\mathcal{X}$. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a reproducing kernel of $\mathcal{H}$, and $\mathcal{H}$ is a reproducing kernel Hilbert space, if

- $\forall x \in \mathcal{X}, \quad k(\cdot, x) \in \mathcal{H}$,
- $\forall x \in \mathcal{X}, \forall f \in \mathcal{H}, \quad \langle f(\cdot), k(\cdot, x) \rangle_{\mathcal{H}} = f(x)$ (the reproducing property).