Reconciling modern machine-learning practice and the classical bias–variance trade-off (PNAS 19)

Mikhail Belkin, Daniel Hsuc, Siyuan Maa, and Soumik Mandal

Departments at The Ohio State University, Columbia University

Presented by Manuel Burger, ETH Zürich

Abstract – Model Selection

- Breakthroughs in machine learning
- Lack of rigorous understanding
- Classical model selection by bias-variance trade-off
- Recent evidence suggests a new approach to model selection

https://www.flaticon.com/authors/pixel-perfect

Outline

- Definitions and Introduction
- The ”Double-Descent”-curve
- Empirical Evidence
  - Random Fourier Features
  - General Neural Networks
  - Decision Trees and Ensembles
- Conclusion
- Critique

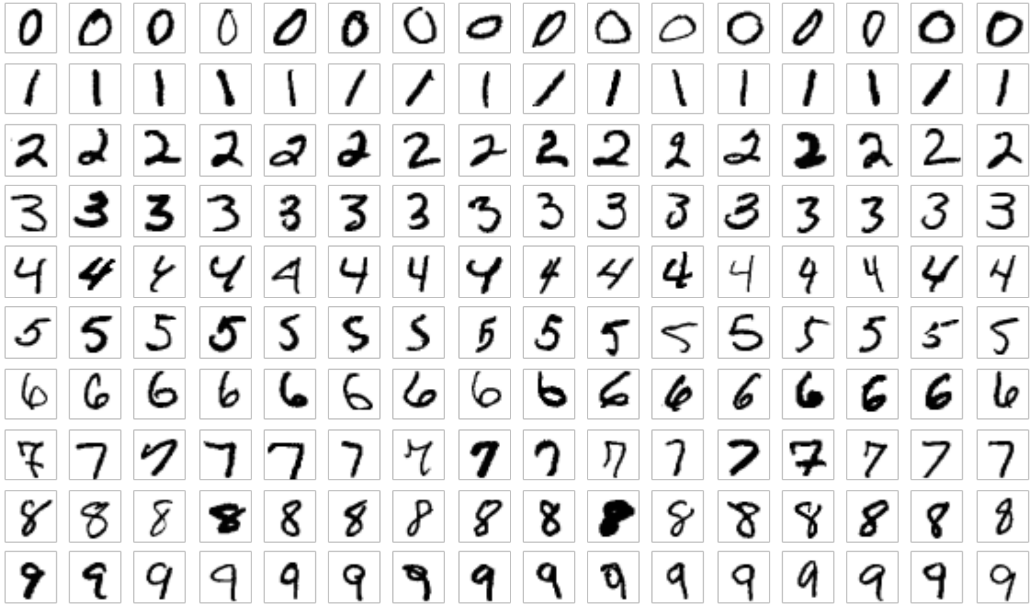
https://www.flaticon.com/authors/vaadin

Definitions

- Classical ERM:
  - $D = \{(x_1, y_1) \dots (x_n, y_n)\}$ where $x_i \in \mathbb{R}^d$
  - Learn $h_n(x) : \mathbb{R}^d \rightarrow \mathbb{R}$ where $h \in \mathcal{H}_N$
    - $\mathcal{H}_N$ capacity in # parameters: N
  - $argmin_h \frac{1}{n} \sum_{i=1}^{n} l(h(x_i), y_i)$ with *0-1* or *squared loss*
  - Evaluate performance of h_n on unseen test data
    - $\mathbb{E}_{(x,y)\sim P}[l(h(x), y)]$
  - No regularization methods
- **Challenge:** Problem mismatch
  - Explicit ERM optimization problem (rigorous definition and solution)
  - Minimizing true/test risk (goal of machine learning)

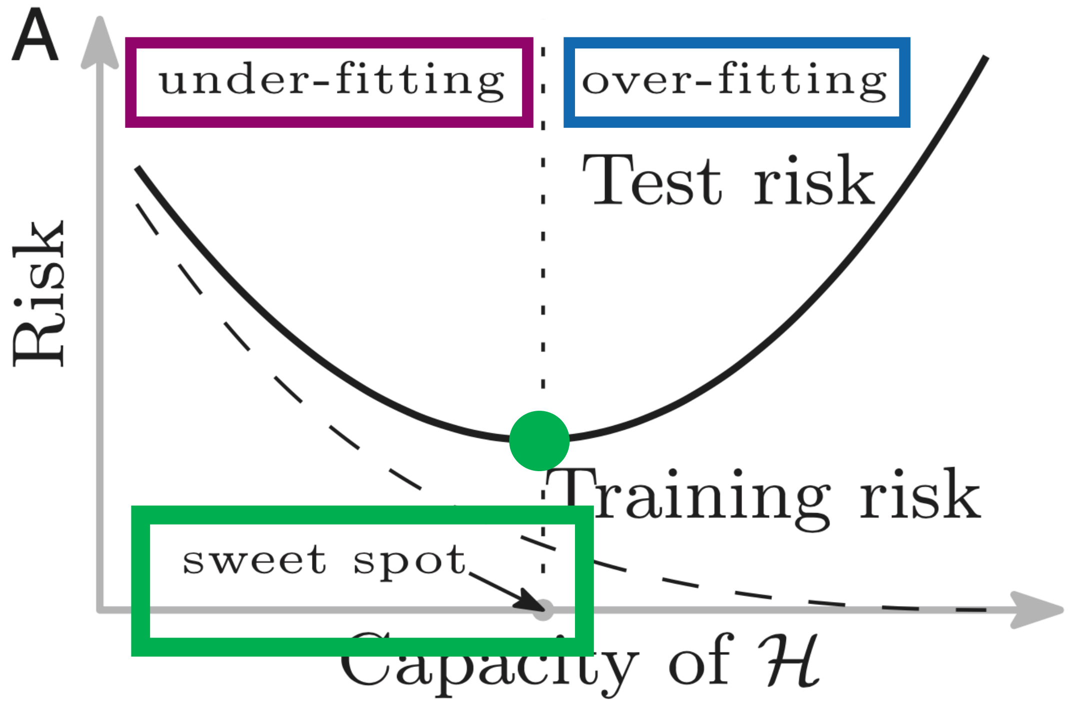
Datasets

- CIFAR-10: object image classification
- MNIST: handwritten digits
- SVHN: house number images
- TIMIT: Speech recognition, dialects
- 20-Newsgroups: News articles and topics



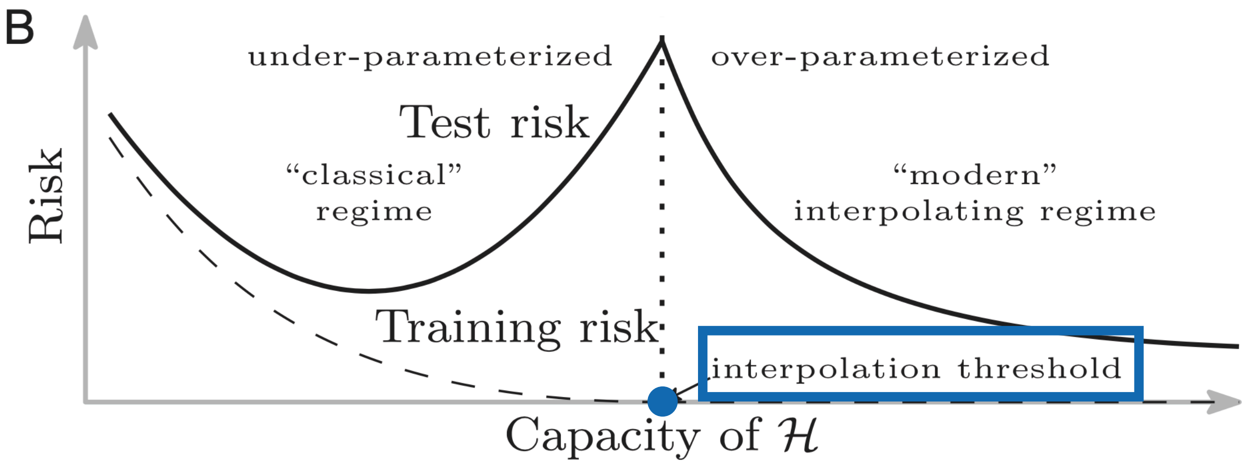
Model Selection - Conventional Approach

- $\mathcal{H}$ too small → underfitting
- $\mathcal{H}$ too large → overfitting
- Find sweet spot
  - Explicit: e.g. choose fixed architecture
  - Implict: Regularization, Early Stopping, …



Model Selection - Modern Approach

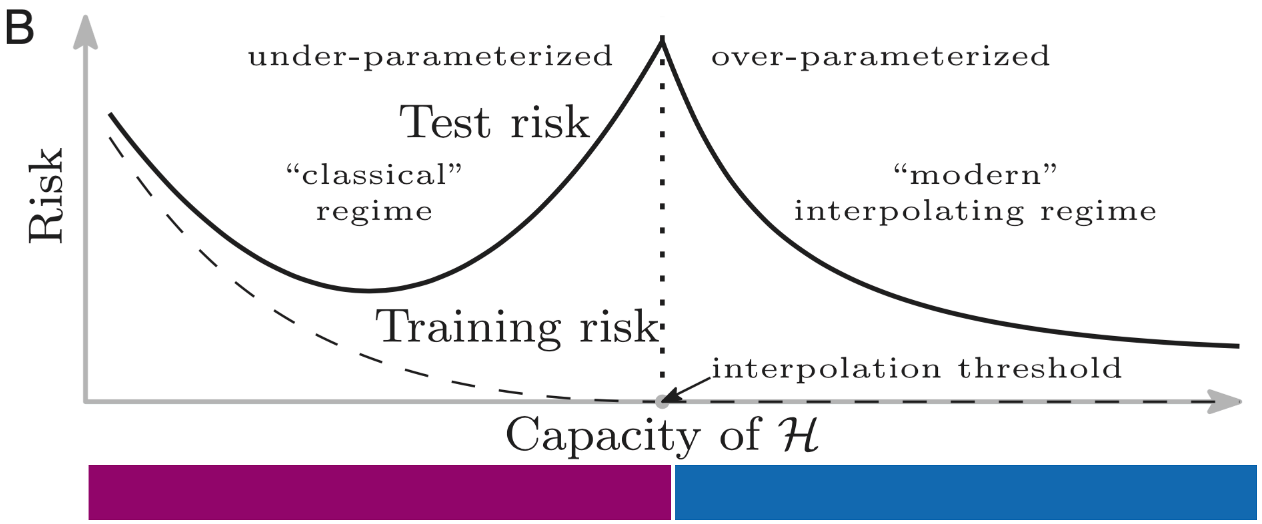
- Select models beyond interpolation threshold
  - 0 training loss
- Use large capacity models
  - Large NNs
  - Other non-linear predictors
- Achieve near-optimal test results
  - Even in high noise settings
  - Better than conventional approach



ETH zürich

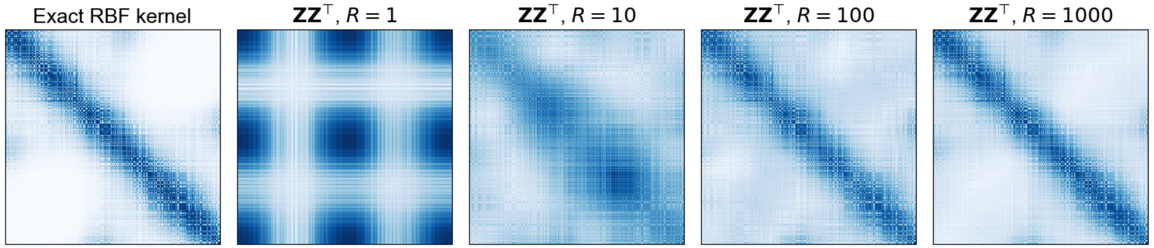
Double-Descent Risk Curve

- $N \leq n$ classical risk behaviour
- $N > n$ double-descent
  - All predictors fit training data perfectly
  - *Capacity of function* vs. *Inductive Bias of problem*
  - *Occam's Razor:* choose simplest explanation possible
    - Find small norm solutions in high capacity space
    - Increased generalization performance



Random Fourier Features

- Class of 2-layer NN with fixed weights in first layer
- v_k sampled from normal distribution in $\mathbb{R}^d$
- $N \rightarrow \infty$ approaches *Gaussian Kernel*
  - Computationally attractive for $N \ll n$
- Optimized with ERM using linear regression
  - $N > n$ choose minimum l_2 norm solution



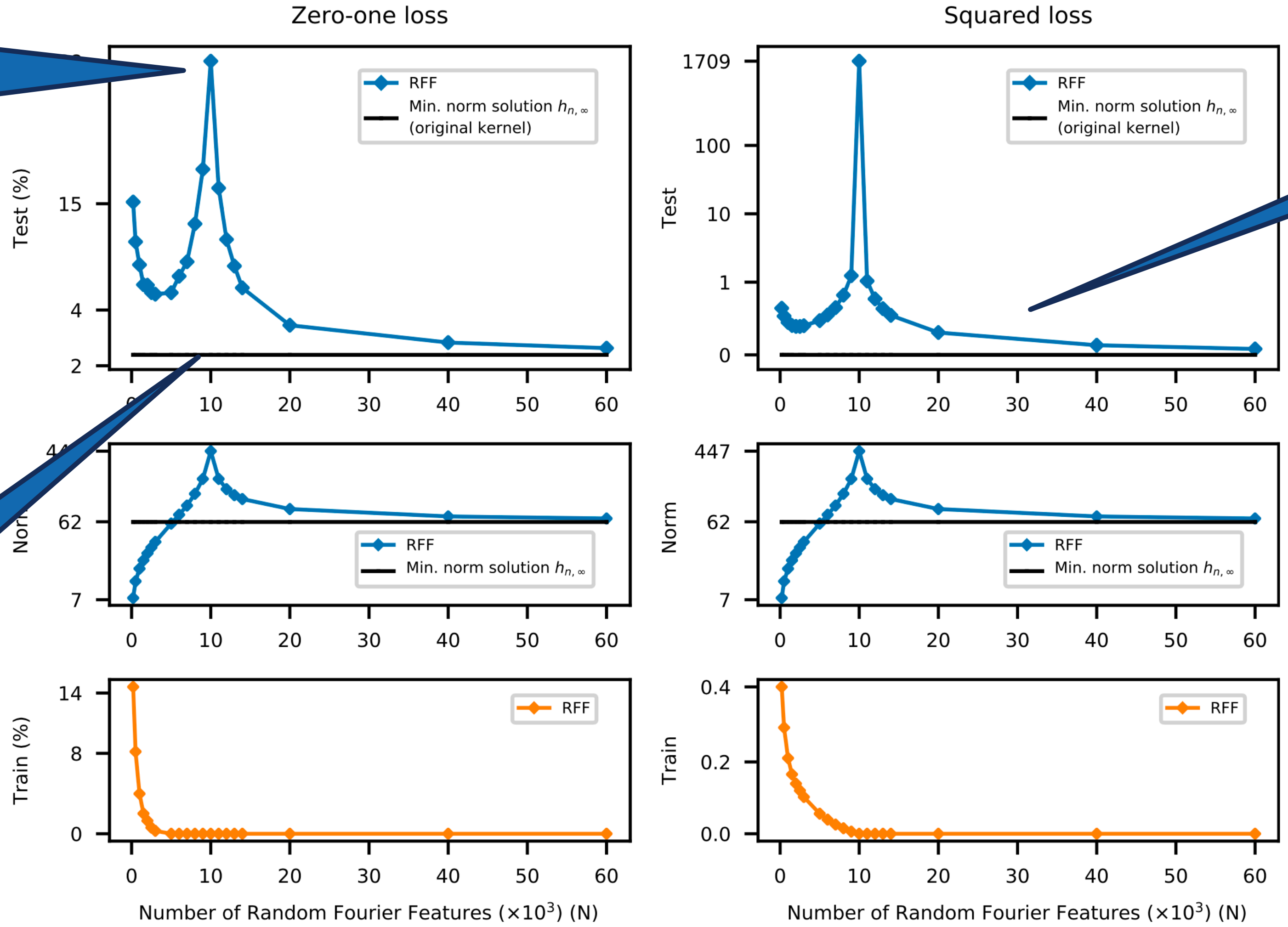
$$h(x) = \sum_{k=1}^{N} a_k \phi(x; v_k)$$

$$\phi(x; v) := e^{\sqrt{-1}\langle v, x\rangle}$$

Image: http://gregorygundersen.com/blog/2019/12/23/random-fourier-features/

Random Fourier Features on MNIST

Peak at interpolation treshold *N*=*n*



Modern model selection regime

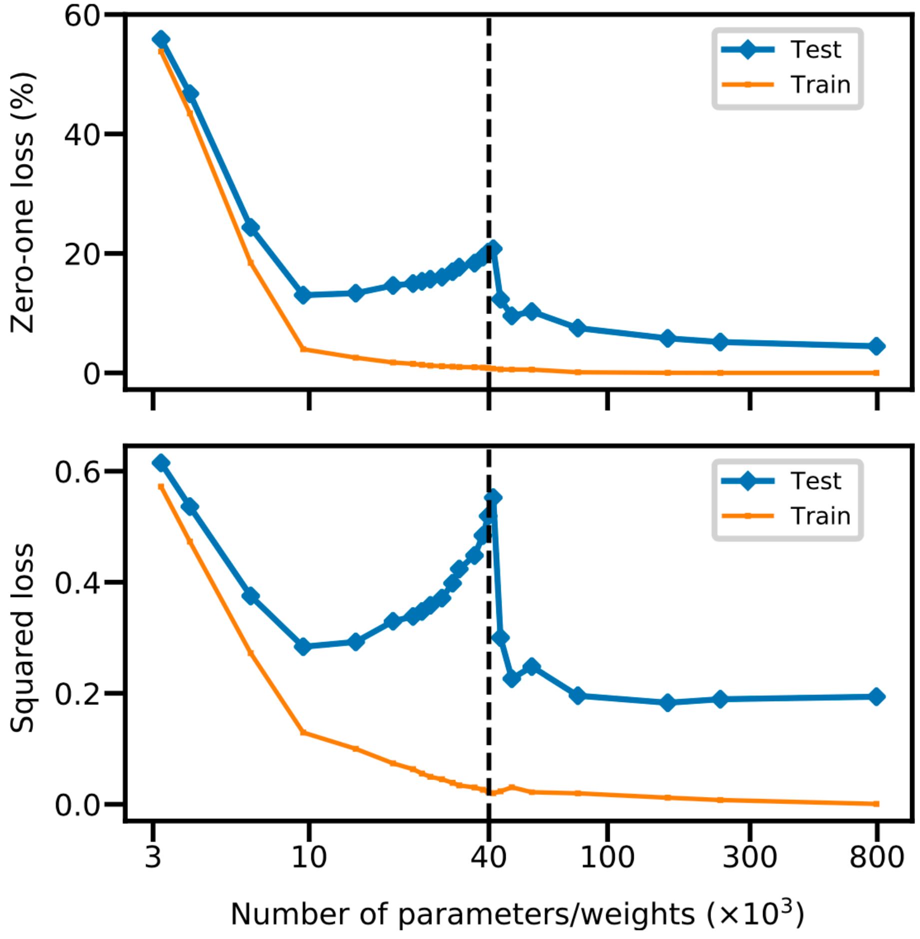
$h_{n,\infty}$ aka Gaussian kernel

Random ReLU Feature models with similar behaviour

Belkin et al. Figure 2 – Double-Descent for RFF on MNIST

General Neural Networks

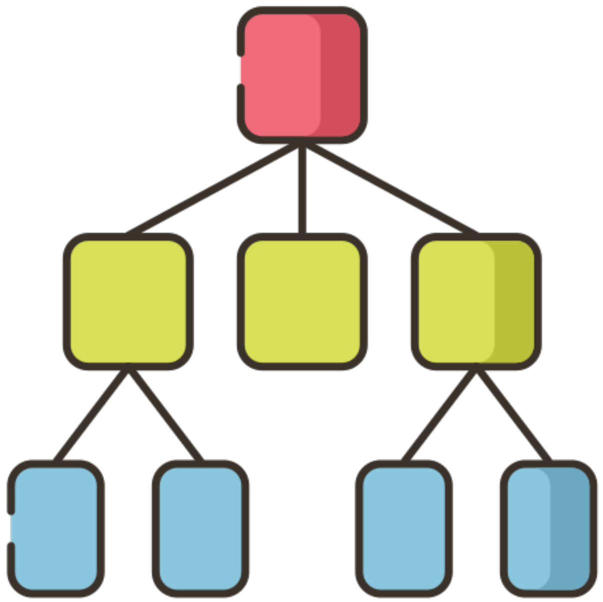
- SGD/Backpropagation
- Observe *double-descent*
- Compatible with previous work suggesting "small norm" *inductive bias* for optim. algo.
  - *Inductive Bias* in architectures
- Interpolation treshold at *#samples x #classes*
  - Requires very large networks
  - ImageNet: 10^6 samples and 10^3 classes
- *N* << *n* high sensitivity to initialization
  - Can mask double-descent curve
  - Weight reuse scheme applied



Belkin et al. Figure 3 – Double-Descent for NN on MNIST

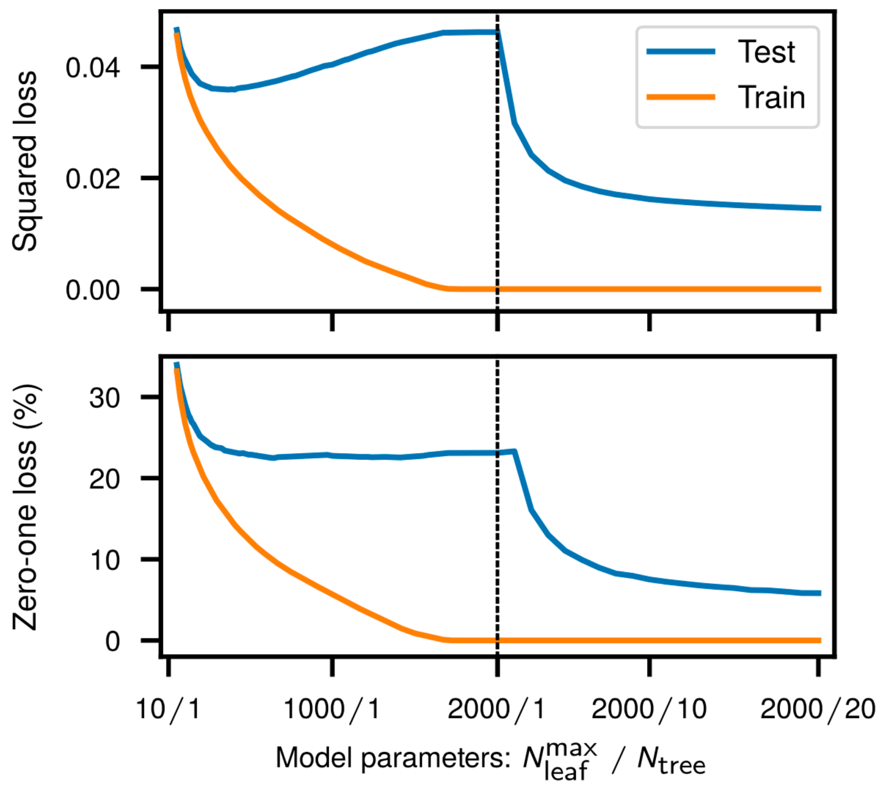
Decision Trees

- Control size of tree by *#leaves*
- Maximally large trees can interpolate data
  - Ensembles achieve smoothness
  - Good *Inductive Bias*
- Beyond interpolation treshold use multiple trees (ensembles)
- Empirical evidence suggests:
  - Adaboost and RF more robust to noise with deep trees than with shallow trees



Random Forests

- Observe *double-descent* risk curve with random forests on MNIST
  - Classical setting for increasing *#leaves*
  - *Double-descent for increasing #deepTrees*
- Similar observation with L_2-boosting



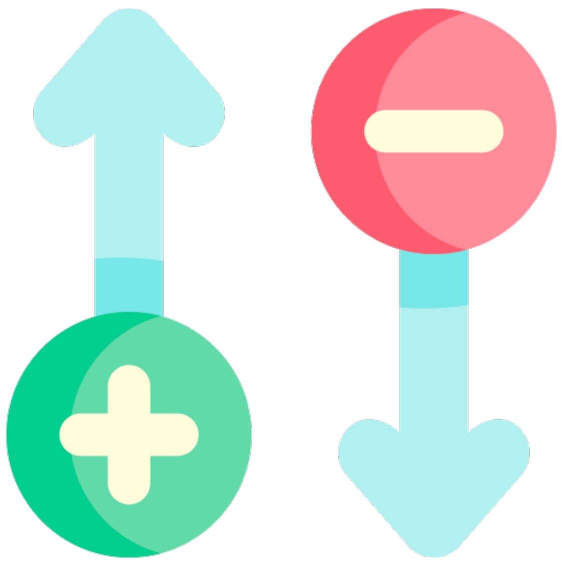
Belkin et al. Figure 4 – Double-Descent for Random Forest on MNIST

Conclusion

- *Double-descent* curve observed
  - Mechanism: *Inductive Bias*
- Historical Absence:
  - Statistical analysis considers small feature space
  - Regularization
  - Smaller models computationally more attractive
  - Observed peak within narrow parameter range
- Inductive Bias
- ”Modern” model selection has better performance and ”easy” to optimize

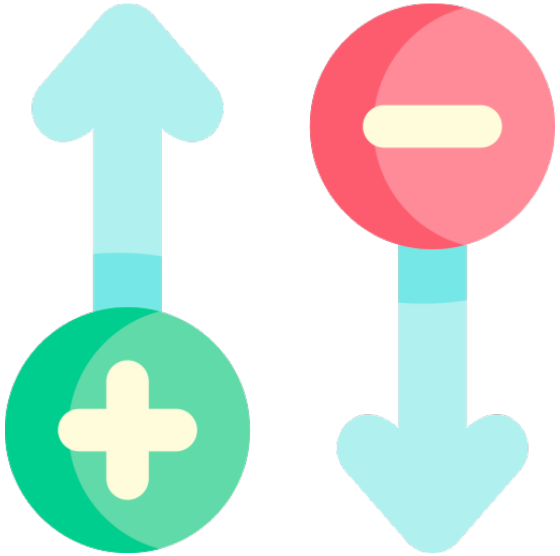
Critique - Strength

- Questioning the *status quo*
  - Encouraging new ways of thinking about model selection
  - Better generalization
  - Suggests “easier” to train models
- Empirical evidence across a range of important predictors
- Considering all major data sources
- High-level analysis widely applicable



Critique - Weaknesses

- Are the examples *designed* to *fit?*
  - Random Features
  - Single hidden layer network
  - Switch from increasing leaves to trees
- *Deep NNs*
  - Difficult to get a *capacity* estimate
- *Modern* Optimizers (Adam, ….)
- *Inductive Bias* vs. regularization?
- Lack of a rigorous explanation
- Increased computational cost



https://www.freepik.com

Outlook

- Investigate optimization properties of solutions
- Find rigorous explanation for the found evidence
- Verify for other common models
  - *Deep NNs*
  - *Modern* optimizers

https://www.freepik.com

Related Work

UNDERSTANDING DEEP LEARNING REQUIRES RE-THINKING GENERALIZATION

Chiyuan Zhang*
Massachusetts Institute of Technology
chiyuan@mit.edu

Samy Bengio
Google Brain
bengio@google.com

Moritz Hardt
Google Brain
mrtz@google.com

Benjamin Recht†
University of California, Berkeley
brecht@berkeley.edu

Oriol Vinyals
Google DeepMind
vinyals@google.com

Related Work

To Understand Deep Learning We Need to Understand Kernel Learning

Mikhail Belkin, Siyuan Ma, Soumik Mandal
Department of Computer Science and Engineering
Ohio State University
{mbelkin, masi}@cse.ohio-state.edu, mandal.32@osu.edu

Related Work

Overfitting or perfect fitting? Risk bounds for classification and regression rules that interpolate

Mikhail Belkin[1], Daniel Hsu[2], and Partha P. Mitra[3]

[1]*The Ohio State University, Columbus, OH*
[2]*Columbia University, New York, NY*
[3]*Cold Spring Harbor Laboratory, Cold Spring Harbor, NY*

Related Work

A Modern Take on the Bias-Variance Tradeoff in Neural Networks

Brady Neal Sarthak Mittal Aristide Baratin Vinayak Tantia Matthew Scicluna
Simon Lacoste-Julien[†,‡] Ioannis Mitliagkas[†]
Mila, Université de Montréal
[†]Canada CIFAR AI Chair [‡]CIFAR Fellow

Cited By

High-dimensional dynamics of generalization error in neural networks

Madhu S. Advani [a,1], Andrew M. Saxe [a,2,*,1], Haim Sompolinsky [a,b]

[a] *Center for Brain Science, Harvard University, Cambridge, MA 02138, United States of America*
[b] *Edmond and Lily Safra Center for Brain Sciences, Hebrew University, Jerusalem 91904, Israel*

Cited By

Deep Double Descent: Where Bigger Models and More Data Hurt

Preetum Nakkiran* Harvard University
Gal Kaplun† Harvard University
Yamini Bansal† Harvard University
Tristan Yang Harvard University
Boaz Barak Harvard University
Ilya Sutskever OpenAI

Thank you for your attention

Discussion

- Have you seen the *double-descent* in practice?
- Knowing about this, will you approach model selection differently?
  - Pro's / Con's
- Do you have any concerns on when this could not work?

https://www.freepik.com