Model Selection And Ensembling When There Are More Parameters Than Data

Michael W. Mahoney

ICSI, LBNL, and Department of Statistics, UC Berkeley

(Joint work with L. Hodgkinson, C. van der Heide, R. Salomone, and F. Roosta; R. Theisen, H. Kim, Y. Yang, and L. Hodgkinson; and others.)

Background (1 of 2)

- Belkin et al., “Reconciling modern machine-learning practice and the classical bias–variance trade-off,” PNAS (2019)
- Zhang et al., “Understanding deep learning requires rethinking generalization,” ICLR (2017)
- . . .
- . . .
- Opper and Kinzel, “Statistical Mechanics of Generalization” (1996); and many others (1990s)

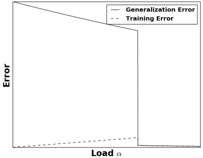
Background (2 of 2)

Martin and Mahoney, Rethinking generalization requires revisiting old ideas, arxiv:1710.09553

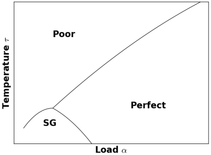
Very Simple Deep Learning (VSDL) model:

- DNN is a black box, load-like parameters α, & temperature-like parameters τ
- Adding noise to training data decreases α
- Early stopping increases τ

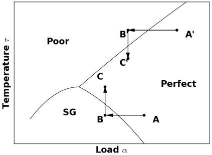
Nearly any non-trivial model[1] exhibits “phase diagrams,” with *qualitatively* different generalization properties, for different parameter values.



(a) Training/generalization error.



(b) Learning phases in τ-α plane.



(c) Noisifying data & adjusting knobs.

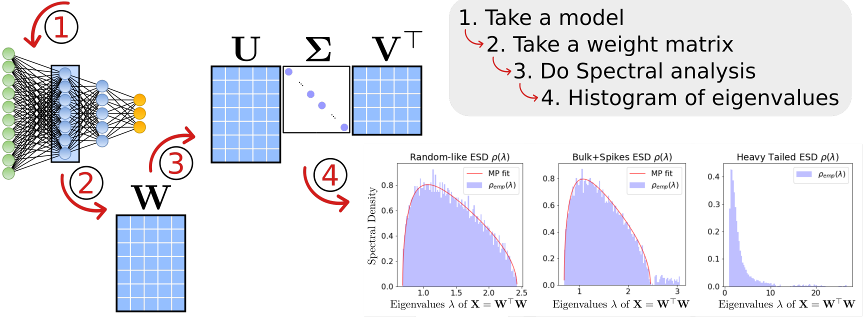
[1] *when analyzed via the Statistical Mechanics Theory of Generalization (SMToG)*

Summary

- Recent “deep learning” boom provides an opportunity: **current theory still does not reflect observations**.
- Current popular theoretical frameworks appear to be *incapable* of bridging this gap.
- First: examine the **data**, meaning the models
- By examining **overparameterized models**, we develop a **new information criterion** for this task.
- We also learn **when ensembles *really* are effective**.

Examine the data (i.e., models)

Analyzing DNN Weight matrices with WeightWatcher



1. Take a model
2. Take a weight matrix
3. Do Spectral analysis
4. Histogram of eigenvalues

You can "pip install weightwatcher"

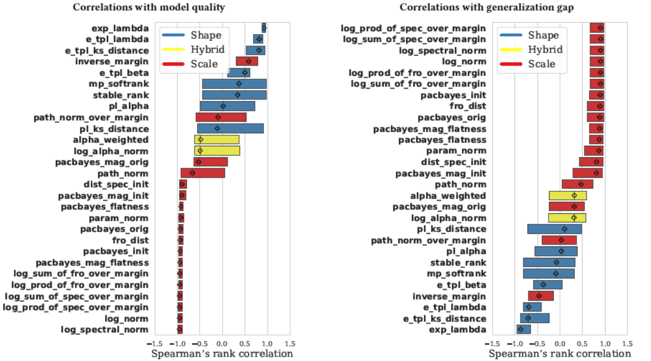
Results of examining the data

Martin and Mahoney, "*Implicit Self-Regularization* in Deep Neural Networks: Evidence from Random Matrix Theory and Implications for Learning," JMLR 2021.

Martin et al., "Predicting trends in the quality of state-of-the-art neural networks *without access to training or testing data*," Nature Communications 2021.

Martin and Mahoney, "Post-mortem on a deep learning contest: a *Simpson's paradox* and the complementary roles of scale metrics versus shape metrics," arXiv:2106.00734.

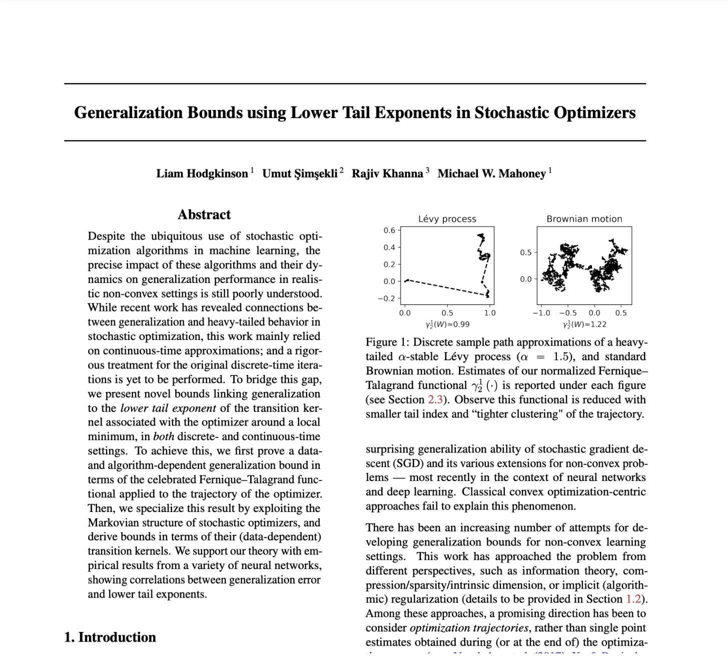
Y. Yang, et al., "Test accuracy vs generalization gap: Model selection in NLP *without accessing any training or testing data*," KDD 2023.



(a) Correlations with model quality. Spearman's rank correlation between various generalization metrics and BLEU.

(b) Correlations with generalization gap. Spearman's rank correlation between various generalization metrics and generalization gap.

Let’s try to develop theory ML people will grok.



Generalization Bounds using Lower Tail Exponents in Stochastic Optimizers

Liam Hodgkinson[1] Umut Şimşekli[2] Rajiv Khanna[3] Michael W. Mahoney[1]

Abstract

Despite the ubiquitous use of stochastic optimization algorithms in machine learning, the precise impact of these algorithms and their dynamics on generalization performance in realistic non-convex settings is still poorly understood. While recent work has revealed connections between generalization and heavy-tailed behavior in stochastic optimization, this work mainly relied on continuous-time approximations; and a rigorous treatment for the original discrete-time iterations is yet to be performed. To bridge this gap, we present novel bounds linking generalization to the *lower tail exponent* of the transition kernel associated with the optimizer around a local minimum, in *both* discrete- and continuous-time settings. To achieve this, we first prove a data- and algorithm-dependent generalization bound in terms of the celebrated Fernique–Talagrand functional applied to the trajectory of the optimizer. Then, we specialize this result by exploiting the Markovian structure of stochastic optimizers, and derive bounds in terms of their (data-dependent) transition kernels. We support our theory with empirical results from a variety of neural networks, showing correlations between generalization error and lower tail exponents.

1. Introduction

Figure 1: Discrete sample path approximations of a heavy-tailed α-stable Lévy process ($\alpha = 1.5$), and standard Brownian motion. Estimates of our normalized Fernique–Talagrand functional $\gamma_2^1(\cdot)$ is reported under each figure (see Section 2.3). Observe this functional is reduced with smaller tail index and “tighter clustering" of the trajectory.

surprising generalization ability of stochastic gradient descent (SGD) and its various extensions for non-convex problems — most recently in the context of neural networks and deep learning. Classical convex optimization-centric approaches fail to explain this phenomenon.

There has been an increasing number of attempts for developing generalization bounds for non-convex learning settings. This work has approached the problem from different perspectives, such as information theory, compression/sparsity/intrinsic dimension, or implicit (algorithmic) regularization (details to be provided in Section 1.2). Among these approaches, a promising direction has been to consider *optimization trajectories*, rather than single point estimates obtained during (or at the end of) the optimiza-

But the bound did not correlate with real-world performance!

But the bound did not correlate with real-world performance!

Options for Theoretical Framework:

- PAC bounds (inadequate)
- Mutual information approaches (unlikely)
- PAC-Bayes framework (good, but hard)

Where does it go wrong?

Parameterized Models

Consider a parameterized model class

$$(\theta \in \mathbb{R}^d, x \in \mathcal{X}) \mapsto f(\theta, x) \in \mathbb{R}^m$$

where

- *n*: number of samples
- *m*: number of scalar outputs
- *d*: number of parameters

$$\begin{array}{rll} \textbf{(under)parameterized} & \text{if} & d \leq mn \\ \textbf{overparameterized} & \text{if} & d > mn \end{array}$$

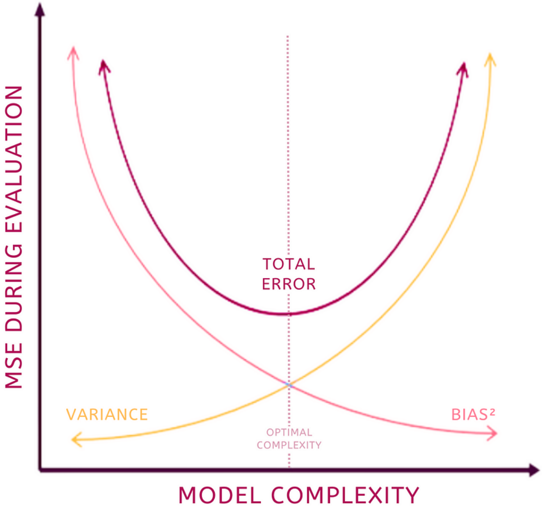
Overparameterized Models

The most performant models are *often* overparameterized.

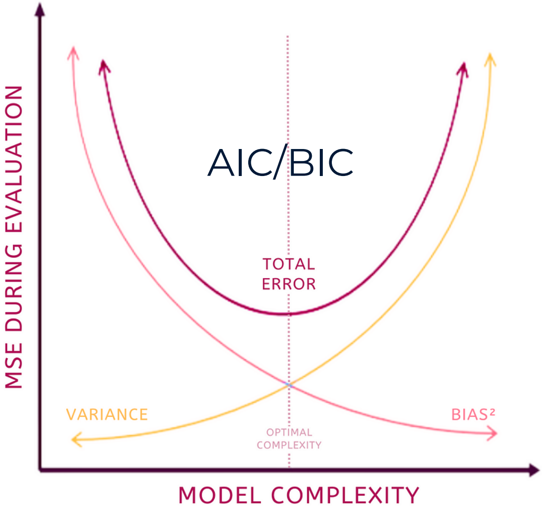
		(log$_{10}$ scale)		
Model	Dataset	nm	d	Test Accuracy
`ResNet18`	`CIFAR-10`	5.7	7.0	93%
`WRN-28-10`	`SVHN`	6.8	7.6	98%
`ViT-E`	`ImageNet-1k`	9.1	9.6	91%
`EFL`	`SNLI`	6.2	8.6	93%
`ResNet34`	`Chaoyang` (noisy)	4.4	7.8	83%
`FiLM`	`ETT` (noisy)	4.0	6.0	—

...and have virtually zero error on training set.

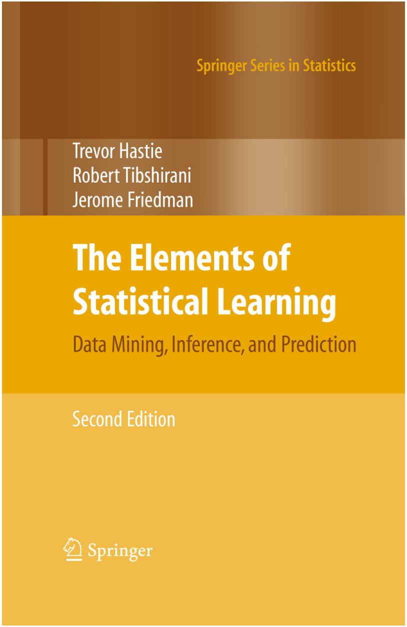
Bias–Variance Tradeoff



Bias–Variance Tradeoff



AIC/BIC



Springer Series in Statistics

Trevor Hastie
Robert Tibshirani
Jerome Friedman

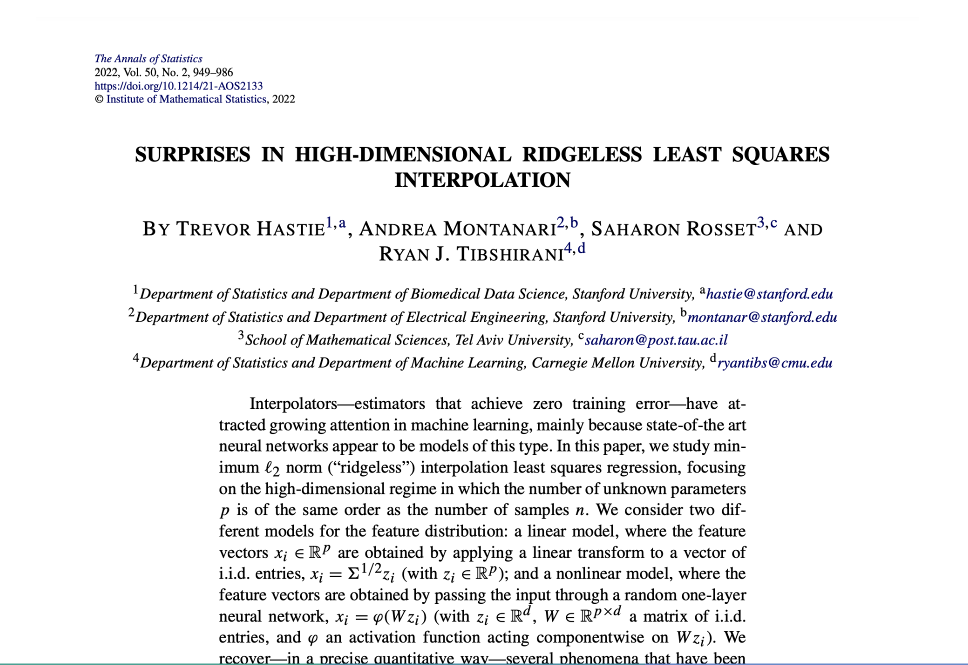
The Elements of Statistical Learning

Data Mining, Inference, and Prediction

Second Edition

Springer

> *“interpolating fits… [are] unlikely to predict future data well at all.”*
>
> pg. 37



The Annals of Statistics
2022, Vol. 50, No. 2, 949–986
https://doi.org/10.1214/21-AOS2133
© Institute of Mathematical Statistics, 2022

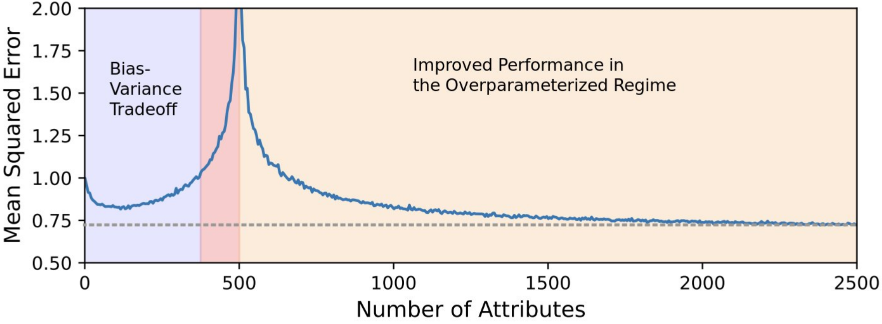
SURPRISES IN HIGH-DIMENSIONAL RIDGELESS LEAST SQUARES INTERPOLATION

BY TREVOR HASTIE[1,a], ANDREA MONTANARI[2,b], SAHARON ROSSET[3,c] AND RYAN J. TIBSHIRANI[4,d]

[1]*Department of Statistics and Department of Biomedical Data Science, Stanford University,* [a]*hastie@stanford.edu*
[2]*Department of Statistics and Department of Electrical Engineering, Stanford University,* [b]*montanar@stanford.edu*
[3]*School of Mathematical Sciences, Tel Aviv University,* [c]*saharon@post.tau.ac.il*
[4]*Department of Statistics and Department of Machine Learning, Carnegie Mellon University,* [d]*ryantibs@cmu.edu*

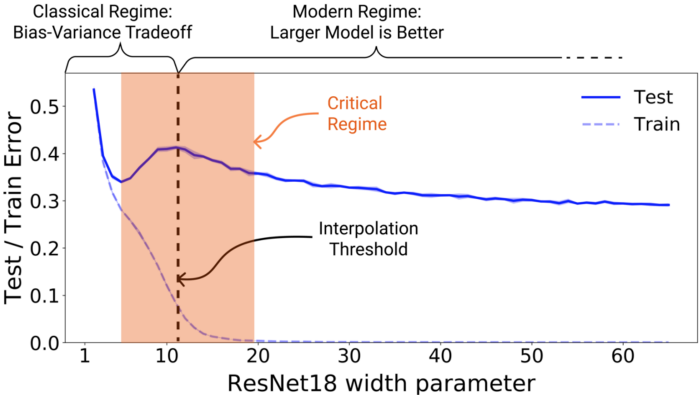
Interpolators—estimators that achieve zero training error—have attracted growing attention in machine learning, mainly because state-of-the art neural networks appear to be models of this type. In this paper, we study minimum ℓ_2 norm ("ridgeless") interpolation least squares regression, focusing on the high-dimensional regime in which the number of unknown parameters p is of the same order as the number of samples n. We consider two different models for the feature distribution: a linear model, where the feature vectors $x_i \in \mathbb{R}^p$ are obtained by applying a linear transform to a vector of i.i.d. entries, $x_i = \Sigma^{1/2} z_i$ (with $z_i \in \mathbb{R}^p$); and a nonlinear model, where the feature vectors are obtained by passing the input through a random one-layer neural network, $x_i = \varphi(W z_i)$ (with $z_i \in \mathbb{R}^d$, $W \in \mathbb{R}^{p \times d}$ a matrix of i.i.d. entries, and φ an activation function acting componentwise on $W z_i$). We recover—in a precise quantitative way—several phenomena that have been

Double Descent ...



Random Fourier Features on MNIST Dataset: $n = 500$

...In Neural Networks



Nakkiran, Preetum, et al. *Deep double descent: Where bigger models and more data hurt*. Journal of Statistical Mechanics: Theory and Experiment 2021.12 (2021): 124003.

Observation

Large class of interpolating solutions

- PAC bounds (**worst case error**) will *often* be vacuous in this regime.
- The $n \to \infty$ limit is not relevant to deep learning

Implicit Regularization

Implicit Regularization?

- *Explicit*: Replace $\min f$ with $\min f + \lambda g$: interpret heuristically or i.t.o. a Bayesian prior.
- *Implicit 1*: $\min f$ is intractable $\rightarrow$ so approximate it:
  Thm 1: $f_{approx} \approx f_{opt}$
  Thm 2: f_{approx} *exactly* solves $\min f + \lambda g$, for *some* λ, g.

  "Approximate Computation and Implicit Regularization ..." Mahoney, PODS 2012.
- *Implicit 2*: Do SGD for NN training and fiddle with knobs: Every training knob *de facto* is a regularization knob.

  "Regularization for Deep Learning: A Taxonomy," Kukacka et al. 2017.
- *Implicit Self*-Regularization: The training process itself regularizes, depending on (correlated) properties of the data.

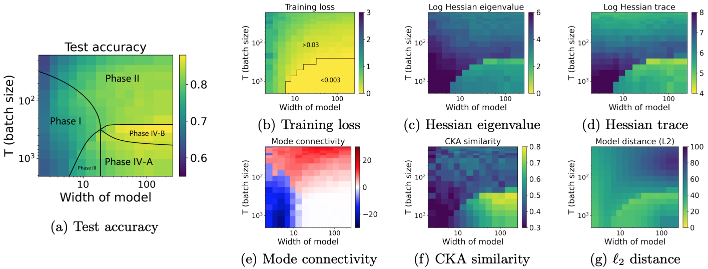
  "Implicit Self-Regularization in Deep Neural Networks ...," M&M, JMLR 2021.

In practice, the error curve can exhibit any form of multiple descent.

But something strange is going on: **no general theory seems to be able to predict this** (relies on MSE calculations).

Taxonomy for Model Quality

Yang et al. “Taxonomizing local versus global structure in neural network loss landscapes,” NeurIPS (2021).



(a) Test accuracy

(b) Training loss

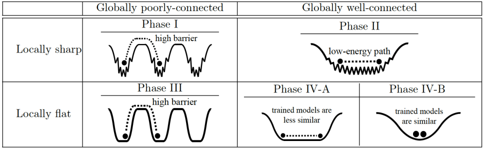
(c) Hessian eigenvalue

(d) Hessian trace

(e) Mode connectivity

(f) CKA similarity

(g) ℓ_2 distance



	Globally poorly-connected	Globally well-connected	
Locally sharp	Phase I high barrier	Phase II low-energy path	
Locally flat	Phase III high barrier	Phase IV-A trained models are less similar	Phase IV-B trained models are similar

Figure 1: **(Caricature of different types of loss landscapes).** Globally well-connected versus globally poorly-connected loss landscapes; and locally sharp versus locally flat loss landscapes. Globally well-connected loss landscapes can be interpreted in terms of a global “rugged convexity”; and globally well-connected and locally flat loss landscapes can be further divided into two sub-cases, based on the similarity of trained models.

Observations

- Double descent cannot arise in the **large data limit** $n \to \infty$ (with other things fixed)!
- Most successful predictions for neural networks arise from **Bayesian approaches**.
- Some form of **duality** taking place.

Setup

Parameterized models are fitted using **empirical risk minimization**:

$$\min_{\theta} \sum_{i=1}^{n} \ell(\underbrace{f(\theta, x_i)}_{\text{predictions}}, \underbrace{y_i}_{\text{label}}) =: L(F(\theta), y)$$

Equivalently, **maximum likelihood estimation** under Gibbs likelihood

$$p(y|x, \theta) \propto \exp\left(-\frac{1}{\gamma} L(F(\theta), y)\right)$$

Interpolators

In the **overparameterized regime**, solutions are **interpolators**:

$$f(\theta^*, x_i) = y_i \text{ for all } i = 1, \ldots, n.$$

How do we uniquely identify θ^*? **Regularizers**.

$$\min_{\theta} R(\theta)$$

$$\text{subject to} \quad f(\theta, x_i) = y_i \text{ for all } i = 1, \ldots, n.$$

Interpolators

Example (Linear Regression)

$$\min_{\theta} ||\theta||^2 \text{ subject to } X\theta = y.$$

Example (Stochastic Gradient Descent)

$$\min_{\theta} \mathbb{E}\|\nabla_x f(\theta, X)\|^2 \text{ subject to } f(\theta, x_i) = y_i.$$

Smith, Samuel L., et al. *On the Origin of Implicit Regularization in Stochastic Gradient Descent*. ICLR 2020.

Note that

$$\theta^\star = \arg\min_\theta R(\theta) \quad \text{subject to} \quad L(F(\theta), y) = 0.$$

is not the same as

$$\theta^\star = \arg\min_\theta L(F(\theta), y) + \gamma R(\theta)$$

Duality

Hard constraints exhibit duality, soft constraints do not

Duality

Hard constraints exhibit duality, soft constraints do not

$$\Lambda(\lambda) = \sup_{\gamma > 0} \inf_{\theta} \left[R(\theta) + \frac{1}{\gamma} L(F(\theta), y) + \sum_{i=1}^{n} \lambda_i \cdot (f(x_i, \theta) - y_i) \right].$$

Theorem (Augmented Lagrange Duality)

$$\overbrace{\inf_{\theta \in \text{zero-loss}} R(\theta)}^{\text{problem in } \mathbb{R}^d} = \overbrace{\sup_{\lambda \in \mathbb{R}^{mn}} \Lambda(\lambda)}^{\text{problem in } \mathbb{R}^{nm}} .$$

Duality

Hard constraints exhibit duality, soft constraints do not

$$\Lambda(\lambda) = \sup_{\gamma>0} \inf_{\theta} \left[R(\theta) + \frac{1}{\gamma} L(F(\theta), y) + \sum_{i=1}^{n} \lambda_i \cdot (f(x_i, \theta) - y_i) \right].$$

Theorem (Augmented Lagrange Duality)

$$\overbrace{\inf_{\theta \in \text{zero-loss}} R(\theta)}^{\text{overparameterized}} = \overbrace{\sup_{\lambda \in \mathbb{R}^{mn}} \Lambda(\lambda)}^{\text{underparameterized}} .$$

Corollary

Any overparameterized model with a regularizer R has a corresponding *dual* underparameterized model with loss Λ.

Going Bayes (1 of 2)

- ERM is **maximum likelihood estimation** under Gibbs likelihood

$$p(y|x, \theta) \propto \exp\left(-\frac{1}{\gamma} L(F(\theta), y)\right).$$

- Encode the **regularizer** as a prior (the regularizer is the log-prior)

$$\pi(\theta) \propto \exp\left(-\frac{1}{\tau} R(\theta)\right).$$

- **Interpolator:** maximize prior over the set of MLEs.

Going Bayes (2 of 2)

- **Two temperatures**: γ (likelihood) and τ (prior)
- The **posterior distribution**

$$\rho_{\gamma,\tau}(\theta|x,y) \propto \exp\left(-\frac{1}{\gamma}L(F(\theta),y) - \frac{1}{\tau}R(\theta)\right)$$

concentrates about the interpolator as

$$\gamma \to 0^+ \qquad \text{and } \textit{then} \qquad \tau \to 0^+.$$

- So we can measure the error by examining the posterior under this limit.

Marginal Likelihood

A powerful measure of model quality:

$$\mathcal{Z}_n = \int_{\mathbb{R}^d} p(y|x,\theta)\pi(\theta)\mathrm{d}\theta$$

- Connections to **cross-validation**

  Fong and Holmes. *On the marginal likelihood and cross-validation.* Biometrika (2020).

- Integrates into the **PAC-Bayes** framework

  Germain et al. *PAC-Bayesian theory meets Bayesian inference.* NIPS (2016).

- The *only* challenging part of the **PAC-Bayes** bound depends on the marginal likelihood.

Bayesian Information Criterion

The marginal likelihood is approximated using **Laplace's method** when $n \to \infty$.

Fails in the overparameterized setting

Bayesian Duality

Theorem

There is an **underparameterized dual model** with the same marginal likelihood:

$$\int_{\mathbb{R}^d} p(y|x,\theta)\pi(\theta)\mathrm{d}\theta = \int_{\mathbb{R}^{mn}} p^*(y|z)\pi^*(z)\mathrm{d}z.$$

Under some regularity conditions,

- p^* is log-concave; and
- π^* is smooth.

The Key Technical Trick

Recall the method of integrating in polar coordinates:

$$\int_{\mathbb{R}^d} f(\theta)\mathrm{d}\theta = \int_0^\infty \left(\int_{\|x\|=r} f(x)\mathrm{d}x \right) \mathrm{d}r.$$

This is an example of the coarea formula

Idea: Integrate over the level sets of the model

Consequences of Bayesian Duality

The roles of sample size and model size alternate in the overparameterized setting!

	$d < mn$	$d > mn$
Sample size ↑	good	penalized
Model size ↑	penalized	good

Interpolating Information Criterion?

What if we apply the same techniques used to derive BIC on the dual model?

Concentrate likelihood first, then concentrate the prior.

Central PAC-Bayes Bound

Theorem

Under mild conditions, if the loss is σ^2-subgaussian, the expected test error in a neighbourhood of the interpolating solution is bounded above by

$$\frac{m}{2}\mathbf{IIC} + \sigma^2 + n^{-1}\log(\delta^{-1}) + \text{const.} + \mathcal{O}(n^{-2}),$$

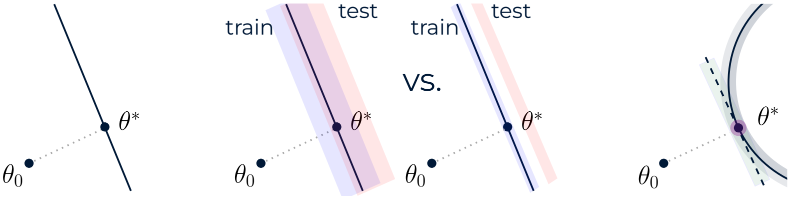
with probability at least $1 - \delta$, where IIC is our *Interpolating Information Criterion*.

Interpolating Information Criterion!

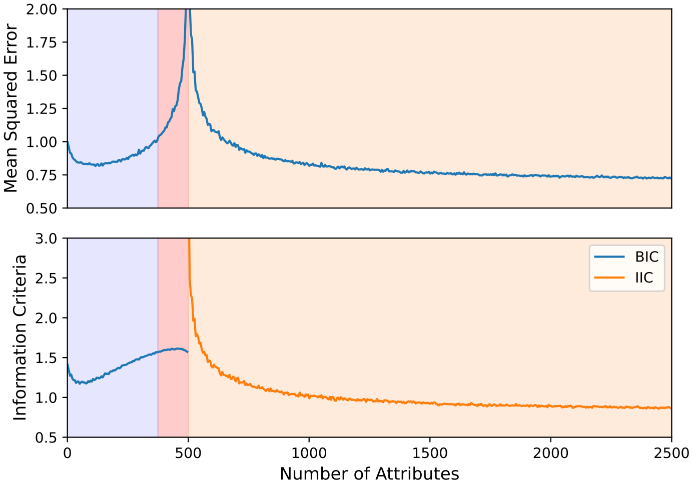
Each of these terms corresponds to and generalized popular heuristics

$$\mathrm{IIC} = -\log n$$

$$\underbrace{+\log[R(\theta^*) - \min_{\theta} R(\theta)]}_{\textbf{Log-regulariser}} \quad \underbrace{+\frac{1}{n}\log\det(\mathrm{NTK})}_{\textbf{Sharpness}} \qquad \underbrace{+\frac{1}{n}\log\mathcal{K}^{\pi}_{\mathcal{M}}(\theta^*)}_{\textbf{Curvature}}$$



MSE, BIC, and IIC on RFF model



Uses of the IIC

Lots!

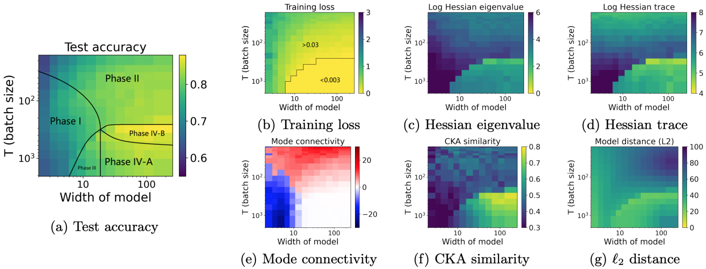
- Model selection: like with AIC, BIC, ...
- PAC-Bayes bounds: on arXiv soon
- Improvements from ensembling
- Basis for HTSR-bsed semi-empirical theory
- UQ: especially in scientific/engineering ML
- Regression diagnostics: on NN models
- ...

Hodgkinson et al., “The Interpolating Information Criterion for Overparameterized Models,” 2023

Hodgkinson et al., “Monotonicity and Double Descent in Uncertainty Estimation with Gaussian Processes,” 2022

Taxonomy for Model Quality

Yang et al. "Taxonomizing local versus global structure in neural network loss landscapes," NeurIPS (2021).



(a) Test accuracy

(b) Training loss

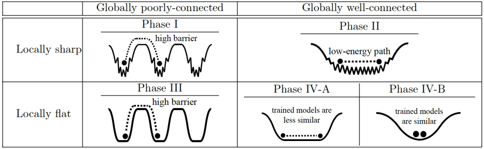
(c) Hessian eigenvalue

(d) Hessian trace

(e) Mode connectivity

(f) CKA similarity

(g) ℓ_2 distance



	Globally poorly-connected	Globally well-connected	
Locally sharp	Phase I	Phase II	
Locally flat	Phase III	Phase IV-A	Phase IV-B

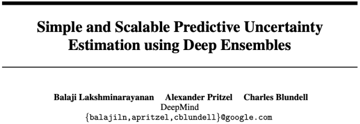
Figure 1: **(Caricature of different types of loss landscapes).** Globally well-connected versus globally poorly-connected loss landscapes; and locally sharp versus locally flat loss landscapes. Globally well-connected loss landscapes can be interpreted in terms of a global "rugged convexity"; and globally well-connected and locally flat loss landscapes can be further divided into two sub-cases, based on the similarity of trained models.

Ensembling: background

- Ensembling has a rich history in ML, with many impactful applications, e.g., random forests, XGBoost

- Many approaches for ensembling NNs have been proposed:
  - Deep ensembles (ensembling many NNs trained from independent initialization), Bayesian NNs, etc.

- Recently, additional use cases of ensembling:
  - Robustness, uncertainty quantification



Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles

Balaji Lakshminarayanan **Alexander Pritzel** **Charles Blundell**
DeepMind
{balajiln,apritzel,cblundell}@google.com

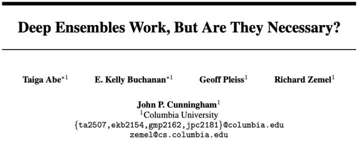
Ensembling: context

In theory:

- large literature on ensembling;
- most is either specialized to particular settings (like random forests),
- or is too weak to even guarantee that ensembling can help at all,
- much less accurately quantify how much it can help

In practice:

- wide variety of (often contradictory) results
- especially for “deep ensembles”
- some work suggests ensembling is highly beneficial,
- other work suggests it is less so,
- and in particular that it is unnecessary for large modern models



Deep Ensembles Work, But Are They Necessary?

Taiga Abe[*1] E. Kelly Buchanan[*1] Geoff Pleiss[1] Richard Zemel[1]

John P. Cunningham[1]
[1]Columbia University
{ta2507,ekb2154,gmp2162,jpc2181}@columbia.edu
zemel@cs.columbia.edu

Theoretical question: Can we characterize when, and by how much, ensembling benefits?

Empirical question: When can we expect ensembling to help significantly in practice?

Broader question: How does this relate to classical statistics versus modern ML/NNs?

When are ensembles *really* effective?

Ensembling: setup

Focus on ensembles of classifiers $h \sim \rho$, where ρ could represent, e.g.:

1. A distribution over parameters obtained from independent runs of SGD, from either dependent (e.g. fine-tuning) or independent initializations
2. A finite set of classifiers $h_1, \ldots, h_M$ with weights $\rho_1, \ldots, \rho_M$
3. A Bayesian posterior distribution over classifiers

We focus on the widely used majority-vote classifier:

$$h_{MV}(x) = \arg\max_y E_{h\sim\rho}[\mathbf{1}(h(x) = y)]$$

We measure performance with the standard misclassification rate: $L(h) = E_{X,Y}[1(h(X) \neq Y)]$

We are interested in characterizing how much ensembling improves performance *relative to the performance of any one classifier, on average:*

Ensemble improvement rate

$$EIR = \frac{\overbrace{E_{h\sim\rho}[L(h)] - L(h_{MV})}^{\text{Improvement vs the average error rate}}}{\underbrace{E_{h\sim\rho}[L(h)]}_{\text{Relative to the average error rate}}}$$

Theory: The Competence Assumption

Intuitively: “it is more likely that *slightly* more classifiers are correct than *slightly* less”

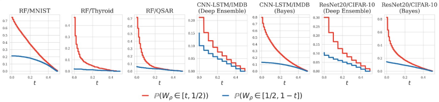
Competent ensembles

Let $W_\rho = E_{h\sim\rho}[1(h(X) \neq Y)]$ be the fraction of classifiers that predict incorrectly on a pair (X, Y). We say the ensemble ρ is *competent* if for all $0 \leq t \leq 1/2$,

$$P_{X,Y}\left(W_\rho \in [t, 0.5\,)\right) \geq P_{X,Y}\left(W_\rho \in [0.5, 1-t]\right)$$

The competence assumption:

- Rules out pathological cases that limit previous theoretical analyses of ensembling
- Easy to check if assumption holds in practice
- Assumption holds broadly for a variety of datasets, ensembling methods



Theory

New bounds on the majority-vote error rate
(assuming competence holds)

Theorem 1 (first-order bound)

$$EIR \geq 0$$

First of its kind to actually *guarantee* ensembling cannot hurt performance ✓

It cannot be used to quantify *how much* ensembling improves performance (since it only uses first-order information) ✗

Comparison to prior results

“Naïve” first-order bound, widely known in the literature, only guarantees that

$$L(h_{MV}) \leq 2E[L(h)]$$

implying the significantly weaker result

$$EIR \geq -1$$

Theory

Characterizing the ensemble improvement rate with the disagreement-error ratio

Def: Model disagreement rate

$$D(h, h') = P_X(h(X) \neq h'(X))$$

Def: Disagreement-error ratio

$$DER = \frac{E_{h,h' \sim \rho}[D(h, h')]}{E_{h \sim \rho}[L(h)]}$$

Theorem 2 (second-order bound)

$$DER \geq EIR \geq \frac{2(K-1)}{K} DER - \frac{3K-4}{K}$$

Two regimes:

1. Ensembles improve performance when DER is large, **disagreement > average error**
2. Ensembles do *not* improve performance by much when DER is small, **disagreement < average error**

Theory

Corollary: new bounds on the majority-vote error rate (assuming competence holds)

Rearranging the previous theorem, new bound on the error rate of the MV classifier

Theorem 3 (Corollary of Theorem 2)

$$L(h_{MV}) \leq \frac{4(K-1)}{K}(E_{h\sim\rho}[L(h)] - \frac{1}{2}E_{h,h'\sim\rho}[D(h,h')])$$

Analytically it generalizes and improves on a prior bound (Masegosa et al 2020) in the special case of binary classification by a factor of 2

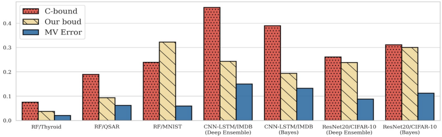
Theory

Corollary: new bounds on the majority-vote error rate (assuming competence holds)

Empirically, Theorem 3 is often significantly sharper than best-known C-bound on the MV classifier, given by

$$L(h_{MV}) \leq 1 - E\left[M_\rho\right]^2 / E[M_\rho^2]$$

where $M_\rho(X,Y) = E[1(h(X) = Y)] - \max_{j \neq Y} E[1(h(X) = j)]$ is the margin.

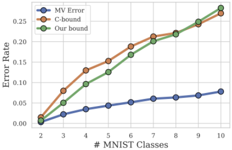


Theory

Corollary: refinement in the case of finite ensembles with many classes

One weakness of Theorem 3 comes in the case of poor scaling with many classes (see figure on MNIST the right)

Forfinite ensembles, where # classifiers < # classes (e.g., ImageNet with 1000 classes), we can prove a refined version of Theorem 3



Theorem 4

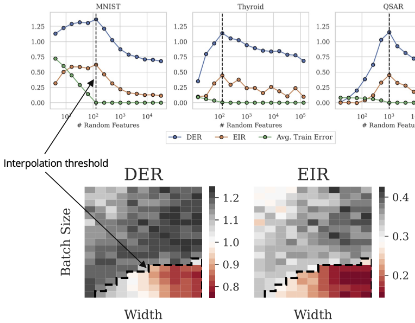
Let $M = \min(K, N)$ $(K = \#classes, N = \#classifiers)$, then

$$L(h_{MV}) \leq \frac{4(M-1)}{M}(E_{h\sim\rho}[L(h)] - \frac{1}{2}E_{h,h'\sim\rho}\,[D(h,h')])$$

Empirical results (1 of 3)

Ensemble improvement, DER become small beyond the interpolation threshold

Bagged Random Feature classifiers



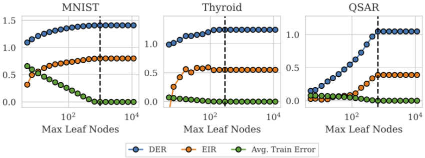
Interpolation threshold

ResNet18/CIFAR-10
Deep Ensembles

- Ensembling becomes less useful for large models which can easily interpolate the training data (i.e., obtain zero training error)
- This corresponds to the fact that the disagreement-error ratio gets small in this regime

Empirical results (2 of 3)

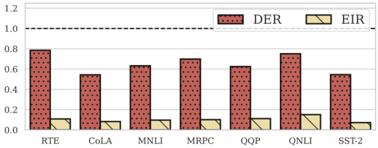
An counterexample: random forests



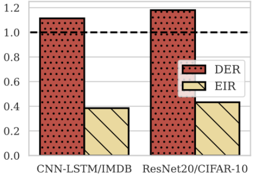
- In contrast to other model classes, random forests ***do not*** undergo a transition at the interpolation threshold
- This is because once zero training error is achieved, trees cannot continue to grow (e.g., Gini impurity = 0)
- Implication: trees are uniquely well-suited to ensembling

Empirical results (3 of 3)

Large language models fine-tuned on small datasets easily interpolate, ensembling doesn't help much



Deep Bayesian ensembles *by design* have high disagreement, expectedly ensembling is very beneficial



Conclusions

Each term in the IIC coincides with a known heuristic for model performance in deep learning.

- **Objective:** Computing the IIC at scale.
- **Prior Choice:** Implicit regularization / incorporating the universal prior (compression).
- Analogue of the **AIC**?

Better diagnostics for SOTA NN models?

Better ensembling / UQ for SOTA NN models?