Candidacy Exam: Benign Overfitting in ML Models

Clayton Sanford

November 16, 2021

Committee: Daniel Hsu, Rocco Servedio, Richard Zemel

Candidacy Exam:
“Models with too many parameters unexpectedly work”

Clayton Sanford

November 16, 2021

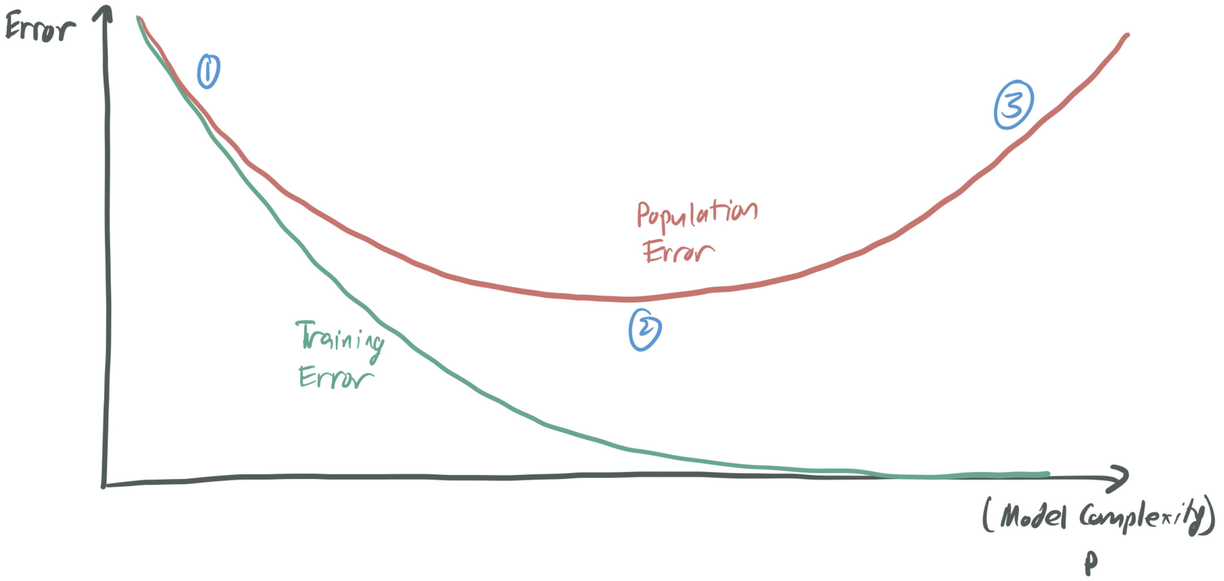
Committee: Daniel Hsu, Rocco Servedio, Richard Zemel

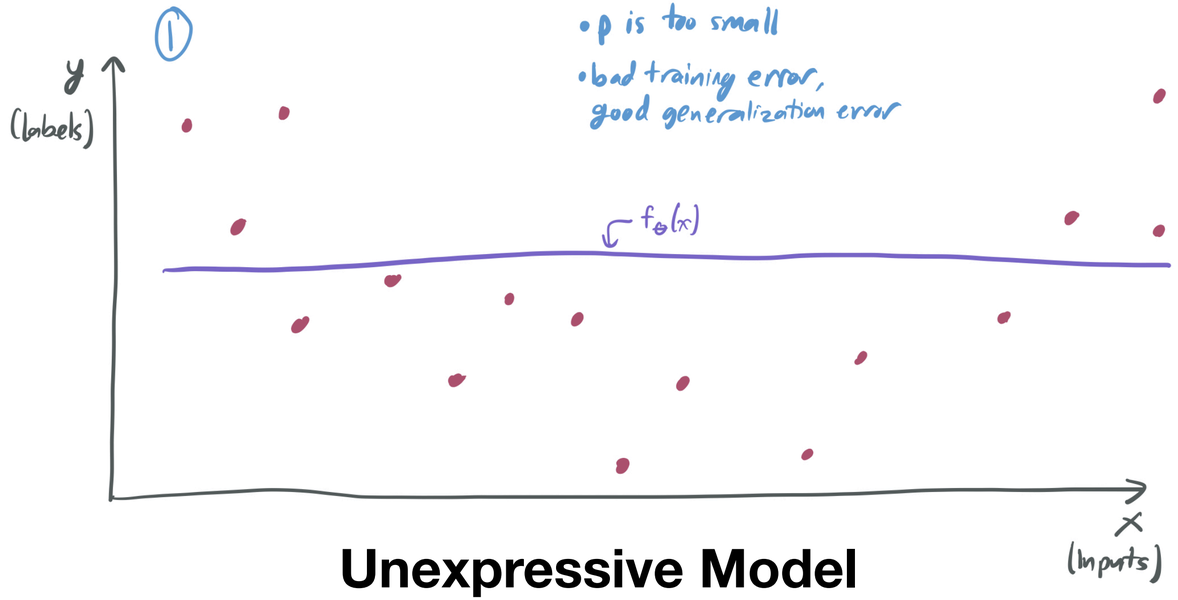
Supervised machine learning

- Given samples $(x_1, y_1), \ldots, (x_n, y_n) \sim \mathscr{D}$.
- Want to learn $h : \mathscr{X} \to \mathscr{Y}$ such that $h(x) \approx y$ for new $(x, y) \sim \mathscr{D}$.
  - $R(h) = \mathbb{E}[\ell(h(x), y)]$ is small.
- How? Find $h \in \mathscr{H}$ minimizing training error: $\hat{R}(h) = \frac{1}{n} \sum_{i=1}^{n} \ell(h(x_i), y_i)$.

$$\underbrace{R(h)}_{\text{population error}} = \underbrace{\hat{R}(h)}_{\text{training error}} + \underbrace{R(h) - \hat{R}(h)}_{\text{generalization error}}$$

Capacity-based narrative

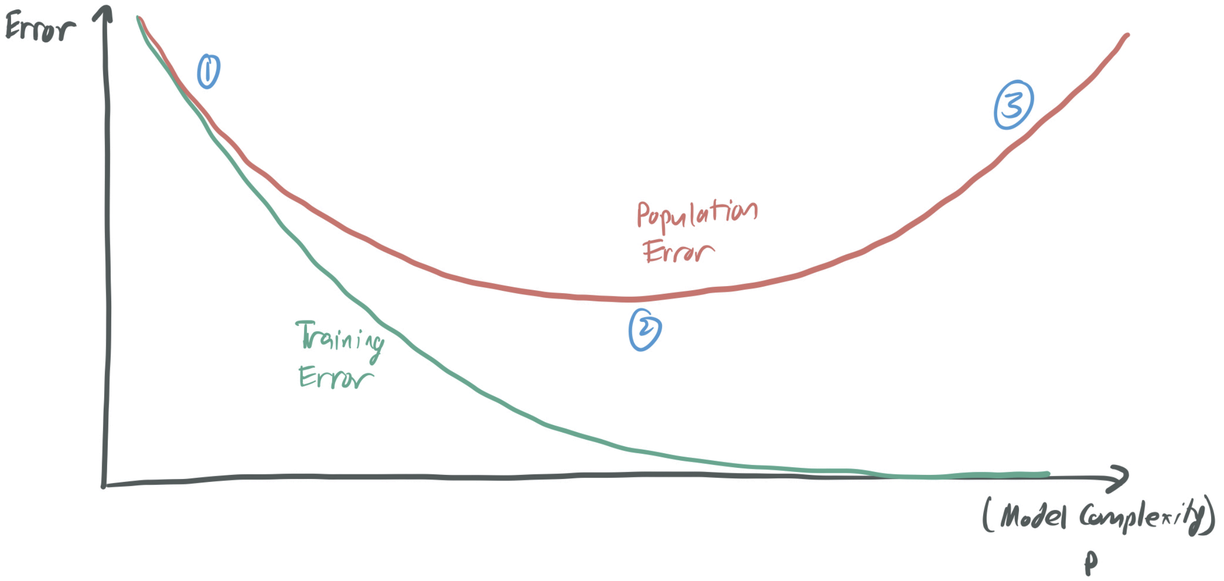


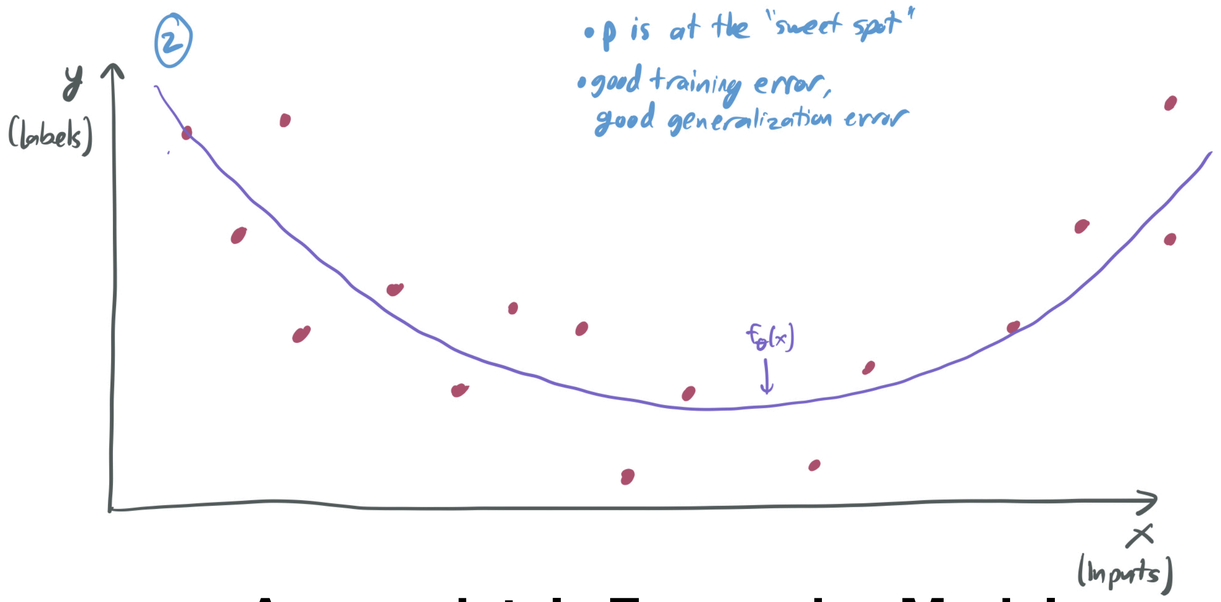


Unexpressive Model

- Capacity-based generalization bounds: VC-dimension, Rademacher, Fat-shattering dimension.
  - e.g. VC-dimension: $R(h) - \hat{R}(h) \leq \tilde{O}\left(\sqrt{\frac{VC(\mathscr{H})}{n}}\right)$

Capacity-based narrative

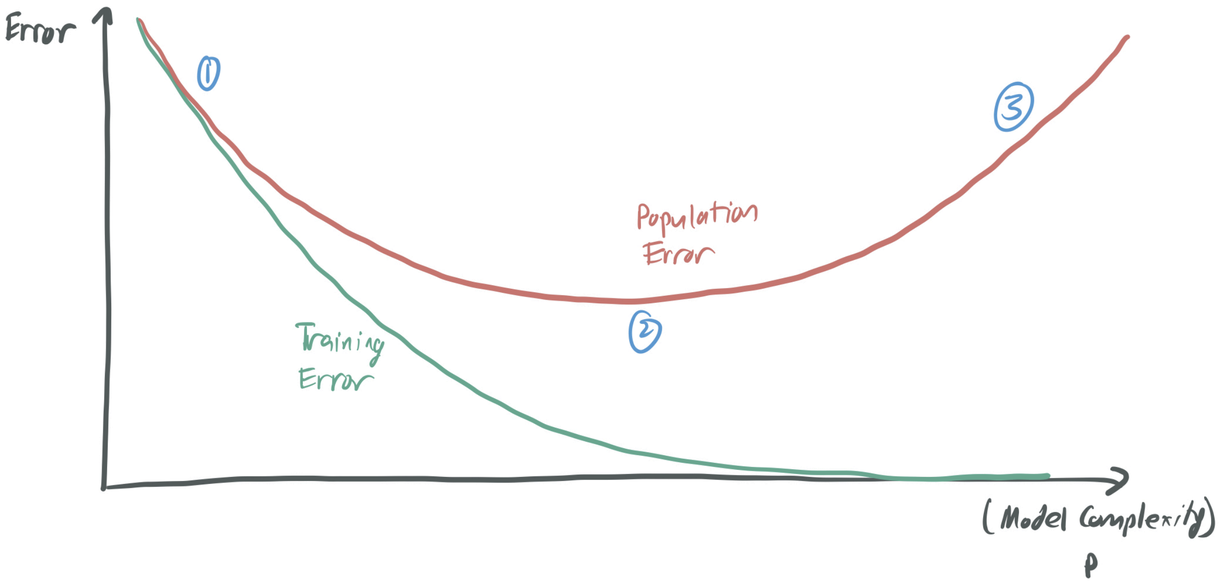


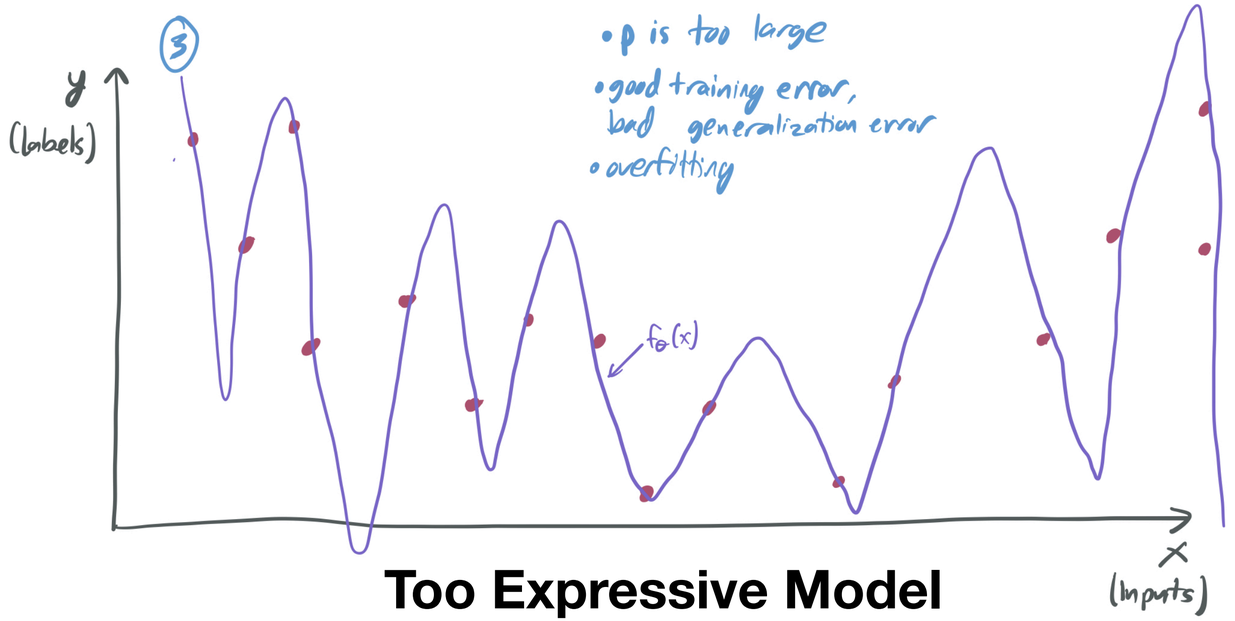


Appropriately Expressive Model

- Capacity-based generalization bounds: VC-dimension, Rademacher, Fat-shattering dimension.
  - e.g. VC-dimension: $R(h) - \hat{R}(h) \leq \tilde{O}\left(\sqrt{\frac{VC(\mathscr{H})}{n}}\right)$

Capacity-based narrative





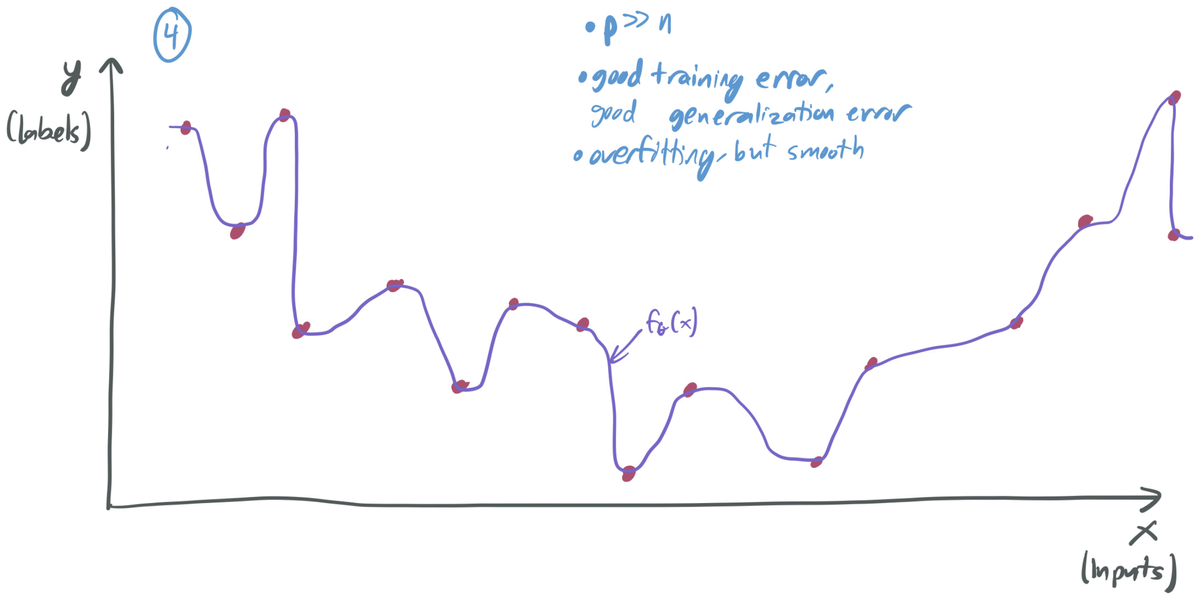
Too Expressive Model

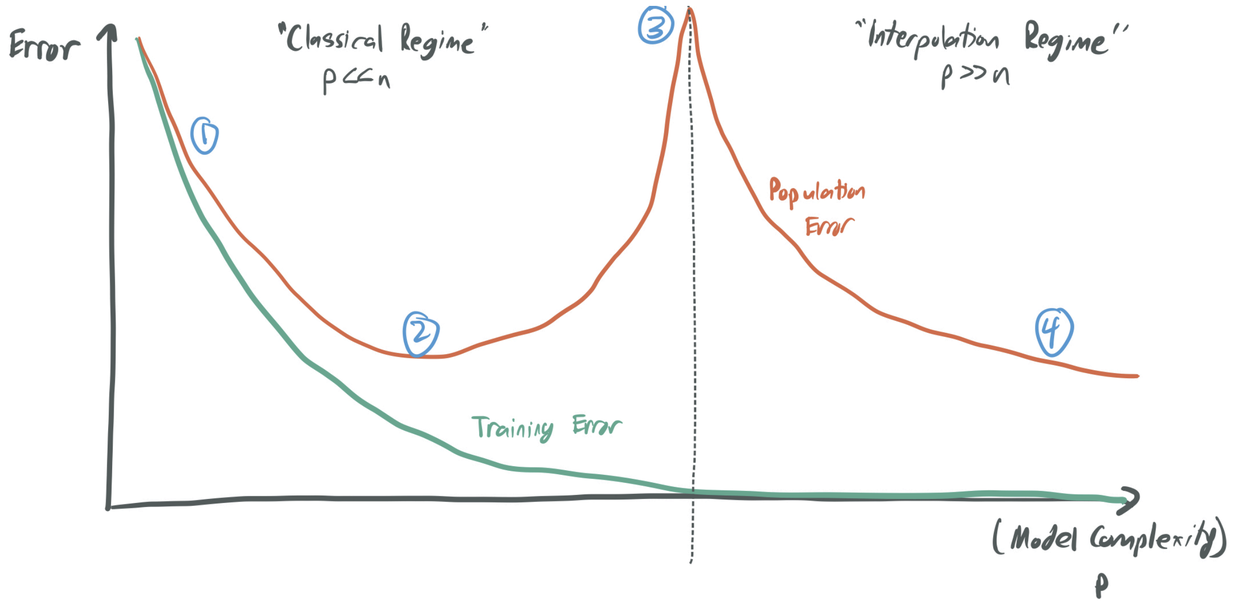
- Capacity-based generalization bounds: VC-dimension, Rademacher, Fat-shattering dimension.
  - e.g. VC-dimension: $R(h) - \hat{R}(h) \leq \tilde{O}\left(\sqrt{\frac{VC(\mathscr{H})}{n}}\right)$

Classical ML theory vs deep learning practice

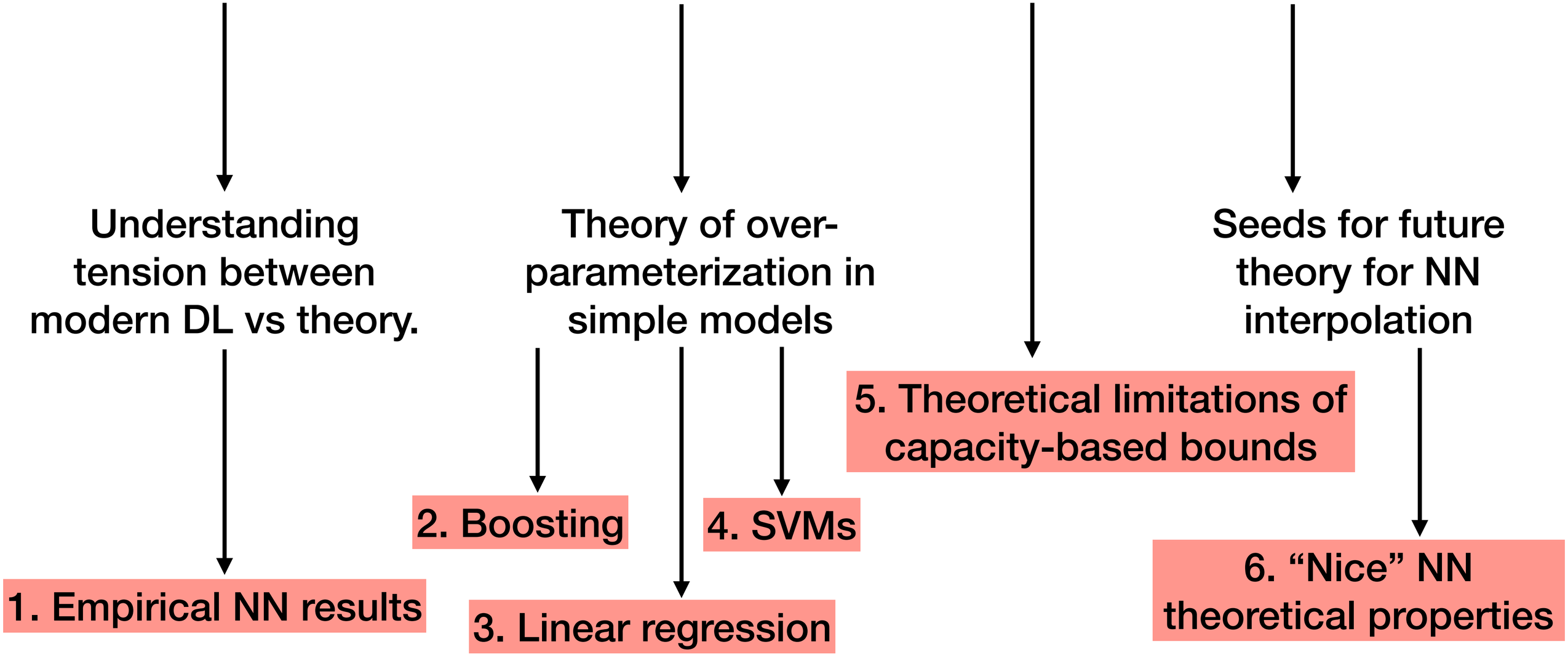
- **Classical ML theory**
  - Choose ML model to balance generalization-vs-approximation tradeoff.
  - Achieve small (but nonzero) training data with ERM.
  - Regularization: sacrifice bias, improve variance
- **Modern deep learning**
  - Design a deep neural network architecture with more parameters than samples.
  - Train to zero training error with SGD or Adam.
  - Great generalization error!?

Benign overfitting and double-descent





How can we align theory with practice?



How can we align theory with practice?

Understanding tension between modern DL vs theory.

Theory of over-parameterization in simple models

Seeds for future theory for NN interpolation

1. Empirical NN results

2. Boosting

3. Linear regression

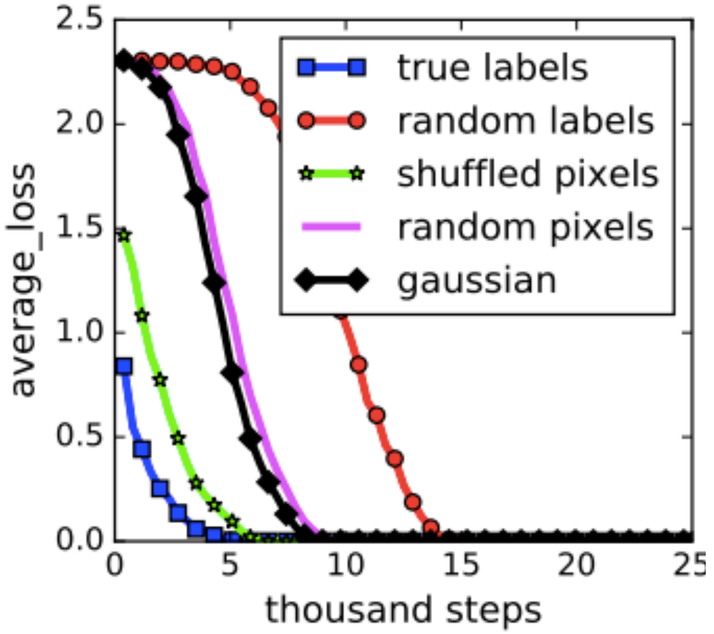
4. SVMs

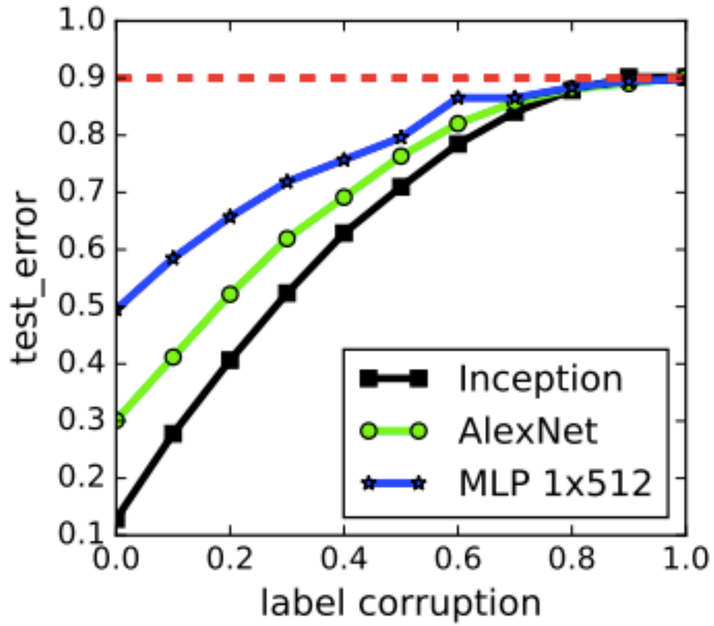
5. Theoretical limitations of capacity-based bounds

6. “Nice” NN theoretical properties

“Rethinking generalization” [ZBHRV17]

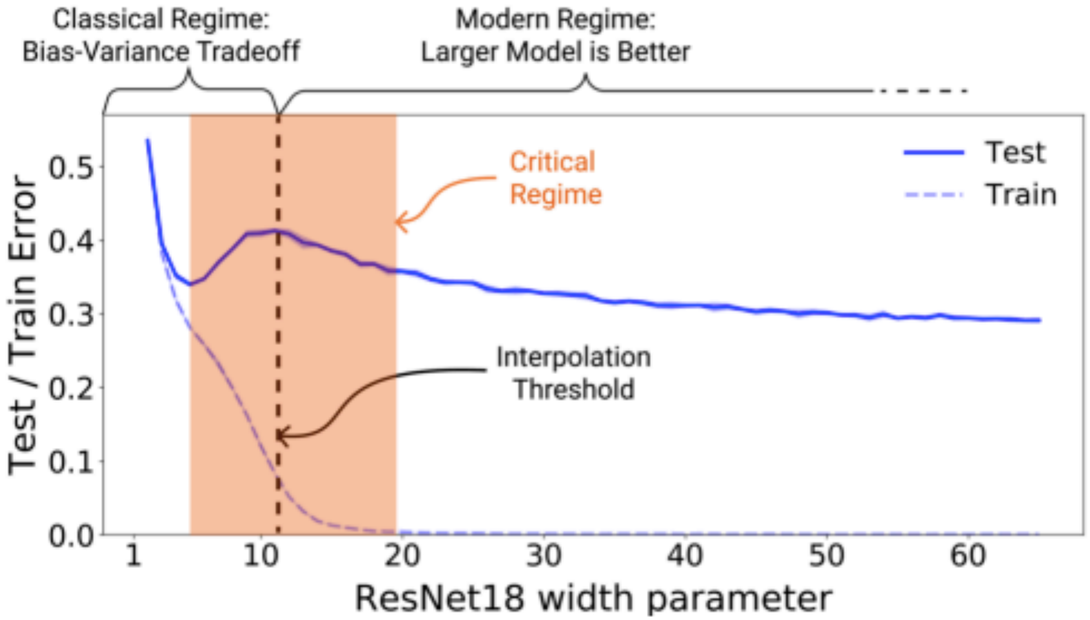
- Could capacity-based bounds describe NN generalization if NNs are biased in favor of “real world” data?
- No… NNs can fit random labels!
- Generalization still possible with corrupted labels.

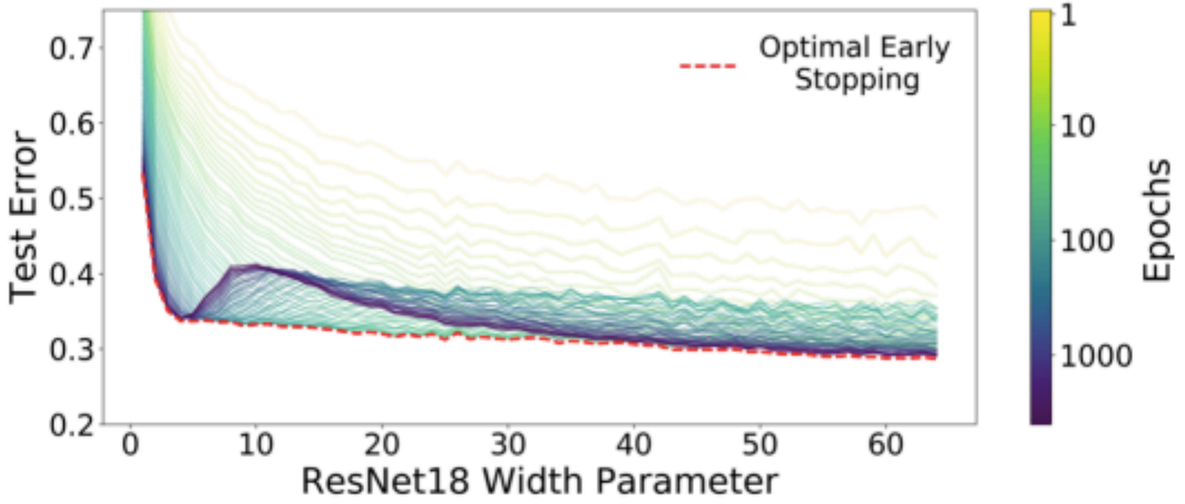




Double-descent in NNs

[NKBYBS19, SGDSBW19, BHMM19]

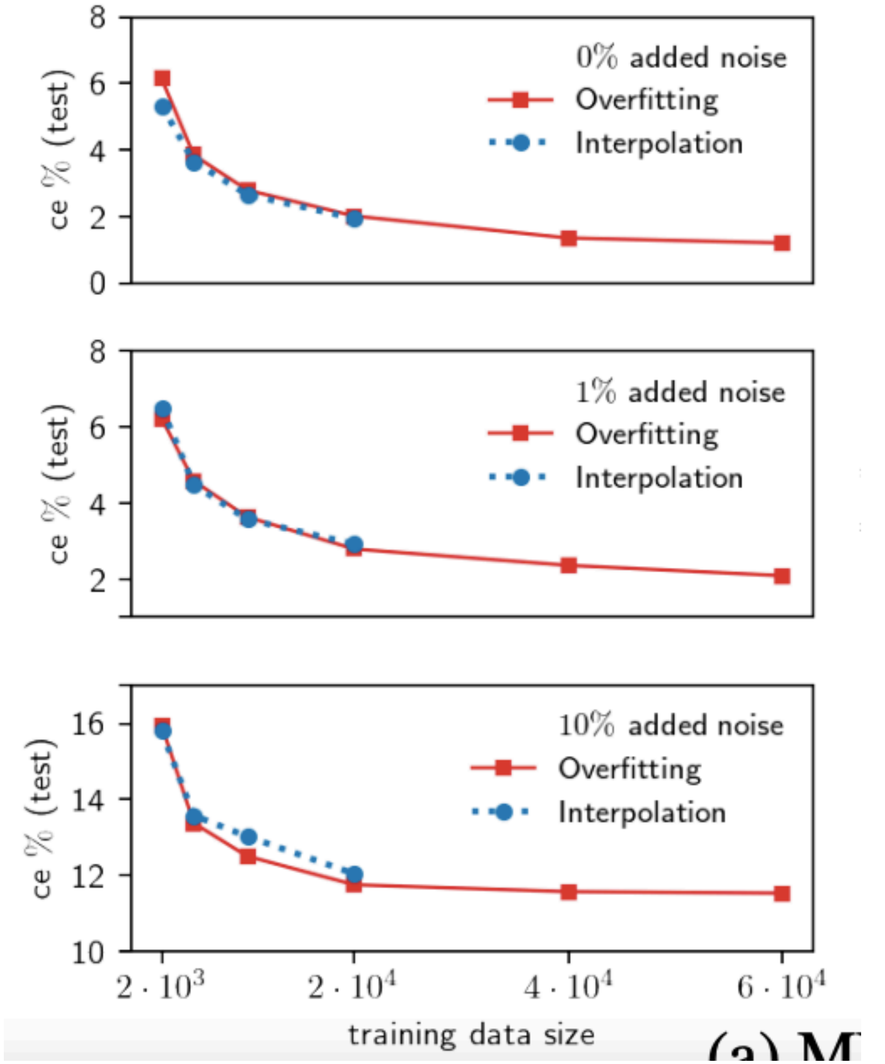




Similar phenomena in other models

[BMM18]

- Kernel methods can have zero training error and small generalization error, like neural networks.
- Laplacian kernels fit random labels.
- Generalization can still occur with corrupted labels
- Capacity-based generalization approaches also don't explain generalization performance of kernel classifiers.



How can we align theory with practice?

Understanding tension between modern DL vs theory.

Theory of over-parameterization in simple models

Seeds for future theory for NN interpolation

1. Empirical NN results

2. Boosting

3. Linear regression

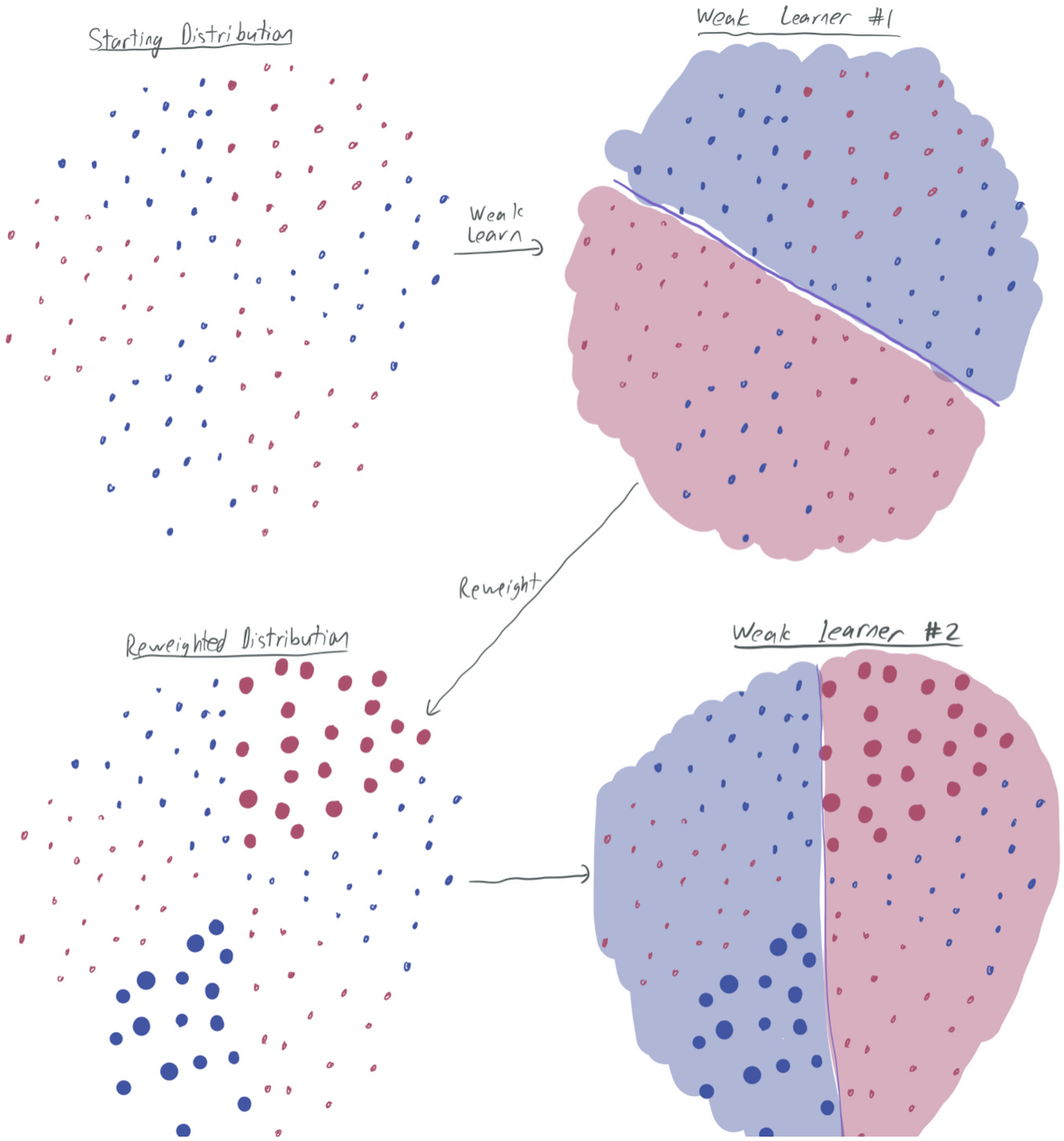
4. SVMs

5. Theoretical limitations of capacity-based bounds

6. “Nice” NN theoretical properties

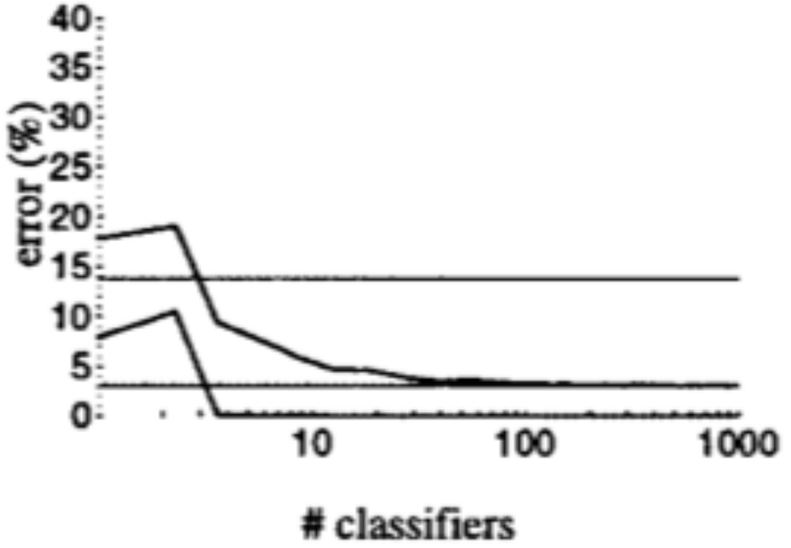
AdaBoost
[FS97]

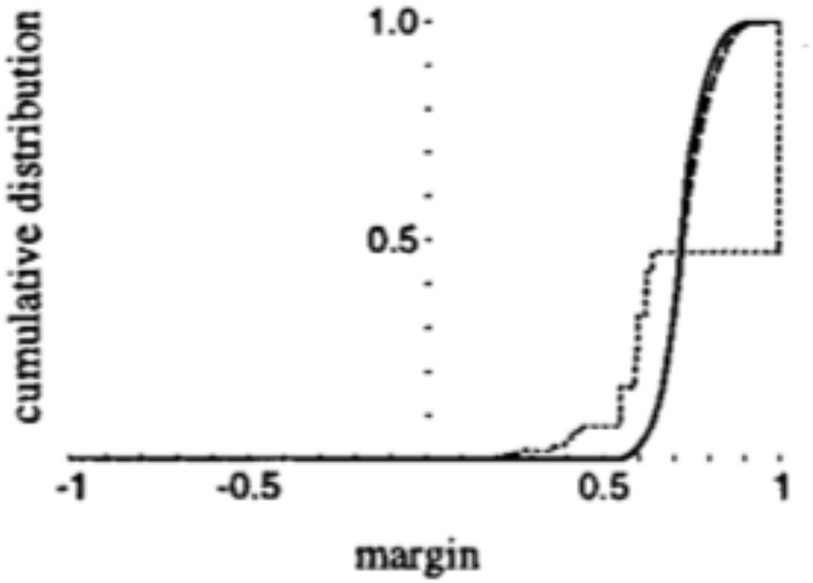
- Idea: aggregate weak learners together into strong learner.
- Guaranteed to fit training data with sufficiently many weak learners.
- $$\epsilon \leq 2^T \prod_{t=1}^{T} \sqrt{\epsilon_t(1-\epsilon_t)}$$



AdaBoost
[FS97]

- Idea: aggregate weak learners together into strong learner.
- Guaranteed to fit training data with sufficiently many weak learners.
- $$\epsilon \leq 2^T \prod_{t=1}^{T} \sqrt{\epsilon_t(1-\epsilon_t)}$$





Boosting the Margin
[BFLS98]

- **Margin bounds**: generalization error is small when training data are decisively classified.

1. Voting classifiers that correctly classify training samples *with large margin* θ have generalization bounds that do not depend on the number of constituent classifiers.
    - Approximates voting classifier with a vote by a random sample of constituents.
    - Decomposes test error with conditional probability and bounds each term with concentration bounds.
2. AdaBoost run for sufficiently many rounds T correctly classifies all training samples with margin θ.
    - Proof similar to AdaBoost convergence result from FS97.

How can we align theory with practice?

Understanding tension between modern DL vs theory.

Theory of over-parameterization in simple models

Seeds for future theory for NN interpolation

1. Empirical NN results

2. Boosting

3. Linear regression

4. SVMs

5. Theoretical limitations of capacity-based bounds

6. “Nice” NN theoretical properties

Linear regression

- Sample $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$. $(X, Y) \in \mathbb{R}^{n \times d} \times \mathbb{R}^n$.
- Learn $x \mapsto \hat{\theta}^T x$.
- **Ordinary least-squares (OLS)** (classical, $n \gg d$):
  - $\hat{\theta} \in \mathbb{R}^d$ minimizes $\sum_{i=1}^{n} (\hat{\theta}^T x_i - y_i)^2$, or $\hat{\theta} = (X^T X)^{-1} X^T Y$.
- **Minimum-norm interpolation** (interpolation, $d \gg n$):
  - $\hat{\theta} \in \mathbb{R}^d$ minimizes $\|\hat{\theta}\|$ such that $\hat{\theta}^T x_i = y_i$, or $\hat{\theta} = X(XX^T)^{-1} Y$.
- **Ridge regression**:
  - $\hat{\theta} \in \mathbb{R}^d$ minimizes $\sum_{i=1}^{n} (\hat{\theta}^T x_i - y_i)^2 + \lambda \|\hat{\theta}\|^2$, or $\hat{\theta} = (X^T X + \lambda I)^{-1} X^T Y$.

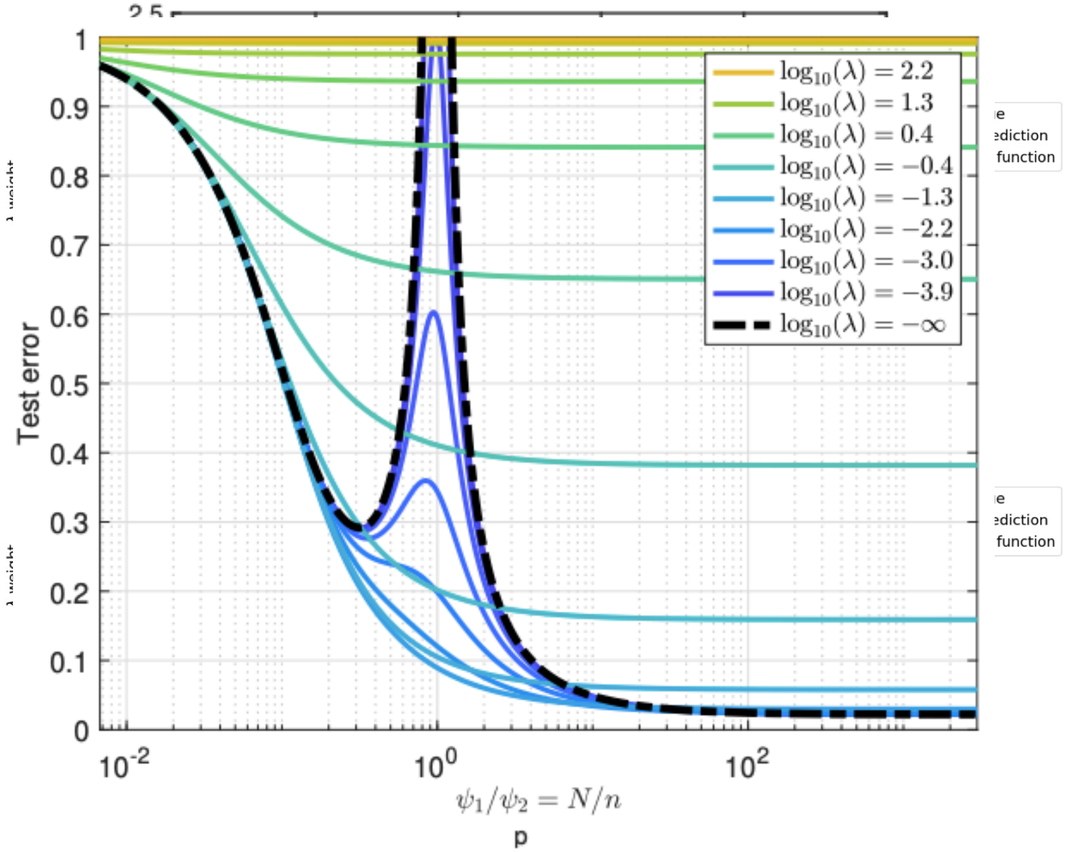
Classical generalization
[Zha05, CD07, AC10]

- OLS generalization bounds are roughly $O(d/n)$ (AC10).
- Can handle potentially infinite-dimensional kernel spaces with notion of effective dimension: $D_\lambda = \operatorname{tr}((\mathbb{E}[xx^T] + \lambda I)^{-1}\mathbb{E}[xx^T])$ (Zha05, CD07).
  - $D_\lambda \to \operatorname{rank}(\mathbb{E}[xx^T])$ as $\lambda \to 0$.
  - $D_\lambda \to 0$ as $\lambda \to \infty$.
  - $D_\lambda = O(\sqrt{n})$ in the worst case.
- Zha05 proves generalization bound on λ-regularized empirical risk-minimizing classifier. Excess error is approximately $\min_{\lambda>0} \lambda + O(D_\lambda/n)$.

Benign overfitting in MNI

[BLLT19, BHX19, HMRT19, MVSS19, Mit19, MN19, MM19]

- Double-descent in **misspecified model** where over-parameterized setting has more information [BHX19, HMRT19, Mit19].
  - Similar phenomenon for random features model [MM19].
- Benign overfitting in full information model when signal concentrated in a small number of important features, surrounded by many unimportant features [BLLT19, HMRT19, MVSS19, MN19].
- Similar results for ridge regression [DW15, TB20].



Benign overfitting: feature importance
[BLLT19]

- MNI for subgaussian inputs with covariance Σ (with eigenvalues $\lambda_1 > \lambda_2 > \ldots$), optimal weights θ^*, and subgaussian noise σ.
- Depends on **effective ranks** of Σ: $r_k(\Sigma) = \sum_{i>k} \lambda_i / \lambda_{k+1}$ and $R_k(\Sigma) = (\sum_{i>k} \lambda_i)^2 / \sum_{i>k} \lambda_i^2$.
- **Theorem**: With probability 0.99 for $k^* = \min\{k \geq 0 : r_k(\Sigma) \geq bn\}$, the excess risk is at most:

$$O\left(\|\theta^*\|^2 \lambda_1 \left(\sqrt{\frac{r_0(\Sigma)}{n}} + \frac{r_0(\Sigma)}{n} \right) + \sigma^2 \left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)} \right) \right)$$

- Bound by **bias-variance** decomposition, concentration bounds based on spectrum, analysis of projection operator onto row space of X.

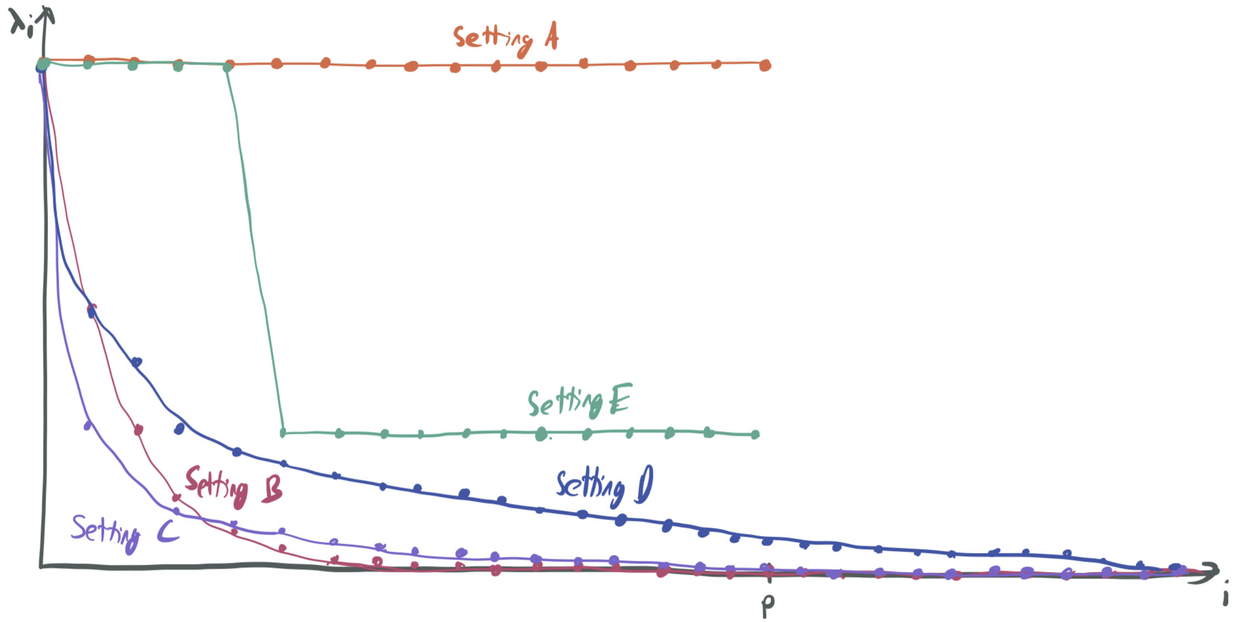
Benign overfitting: feature importance

[BLLT19]

- **Theorem**: With probability 0.99 for $k^* = \min\{k \geq 0 : r_k(\Sigma) \geq bn\}$, the excess risk is at most:

$$O\left(\|\theta^*\|^2 \lambda_1 \left(\sqrt{\frac{r_0(\Sigma)}{n}} + \frac{r_0(\Sigma)}{n} \right) + \sigma^2 \left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)} \right) \right).$$

- **Setting A: Isotropic features**
  - $\Sigma = I_d$.
  - $r_0(\Sigma) = d$.
  - $k^* = 0, R_{k^*}(\Sigma) = d$.
  - $O\left(\|\theta^*\|^2 \frac{d}{n} + \frac{\sigma^2 n}{d} \right)$



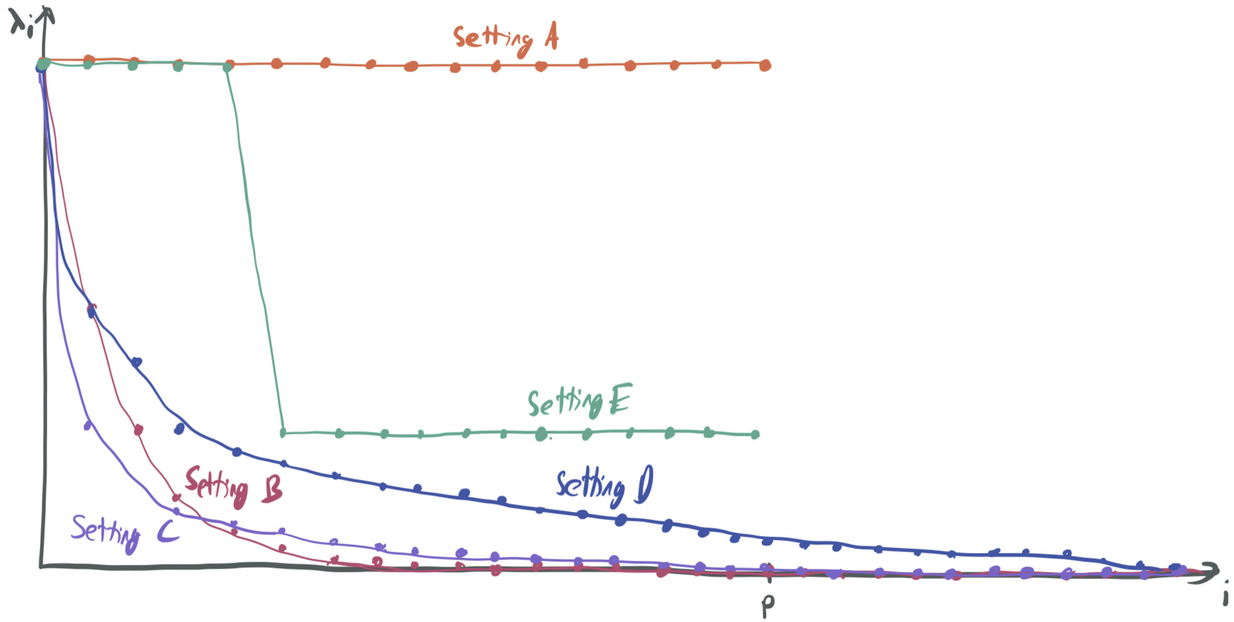
Benign overfitting: feature importance

[BLLT19]

- **Theorem**: With probability 0.99 for $k^* = \min\{k \geq 0 : r_k(\Sigma) \geq bn\}$, the excess risk is at most:

$$O\left(\|\theta^*\|^2 \lambda_1 \left(\sqrt{\frac{r_0(\Sigma)}{n}} + \frac{r_0(\Sigma)}{n} \right) + \sigma^2 \left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)} \right) \right).$$

- **Setting B: Exponential decay**
  - $\lambda_i = 2^{-i}$.
  - $r_0(\Sigma) = 2$.
  - $k^* = \infty \Longrightarrow$ unbounded risk



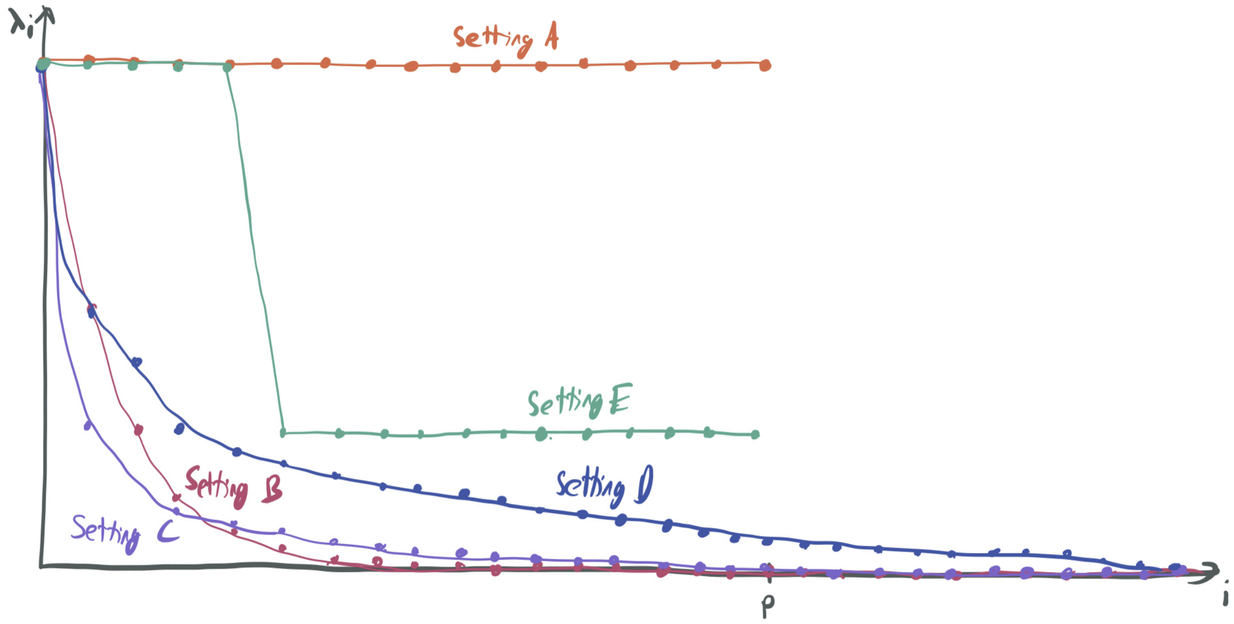
Benign overfitting: feature importance

[BLLT19]

- **Theorem**: With probability 0.99 for $k^* = \min\{k \geq 0 : r_k(\Sigma) \geq bn\}$, the excess risk is at most:

$$O\left(\|\theta^*\|^2 \lambda_1 \left(\sqrt{\frac{r_0(\Sigma)}{n}} + \frac{r_0(\Sigma)}{n}\right) + \sigma^2 \left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}\right)\right).$$

- **Setting C: Quadratic decay**
  - $\lambda_i = 1/i^2$.
  - $r_0(\Sigma) = \Theta(1)$.
  - $k^* = \Theta(n)$, $R_{k^*}(\Sigma) = \Theta(n)$.
  - $O(\|\theta^*\|^2 + \sigma^2)$.



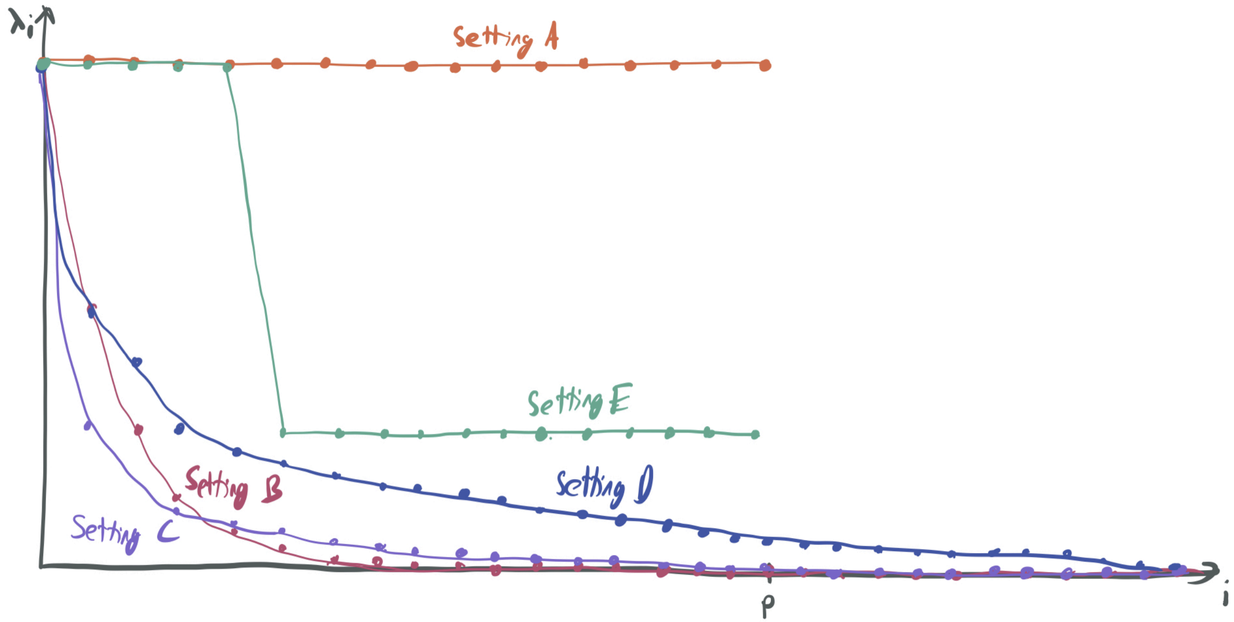
Benign overfitting: feature importance

[BLLT19]

- **Theorem**: With probability 0.99 for $k^* = \min\{k \geq 0 : r_k(\Sigma) \geq bn\}$, the excess risk is at most:

$$O\left(\|\theta^*\|^2 \lambda_1 \left(\sqrt{\frac{r_0(\Sigma)}{n}} + \frac{r_0(\Sigma)}{n}\right) + \sigma^2 \left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}\right)\right).$$

- **Setting D: Slow decay**
  - $\lambda_i = 1/(i \log^2(i+1))$.
  - $r_0(\Sigma) = \Theta(1)$.
  - $k^* = \Theta(n/\log n)$, $R_{k^*}(\Sigma) = \Theta(n \log n)$.
  - $O\left(\frac{\|\theta^*\|^2}{\sqrt{n}} + \frac{\sigma^2}{\log n}\right) \to 0$ as $n \to \infty$.



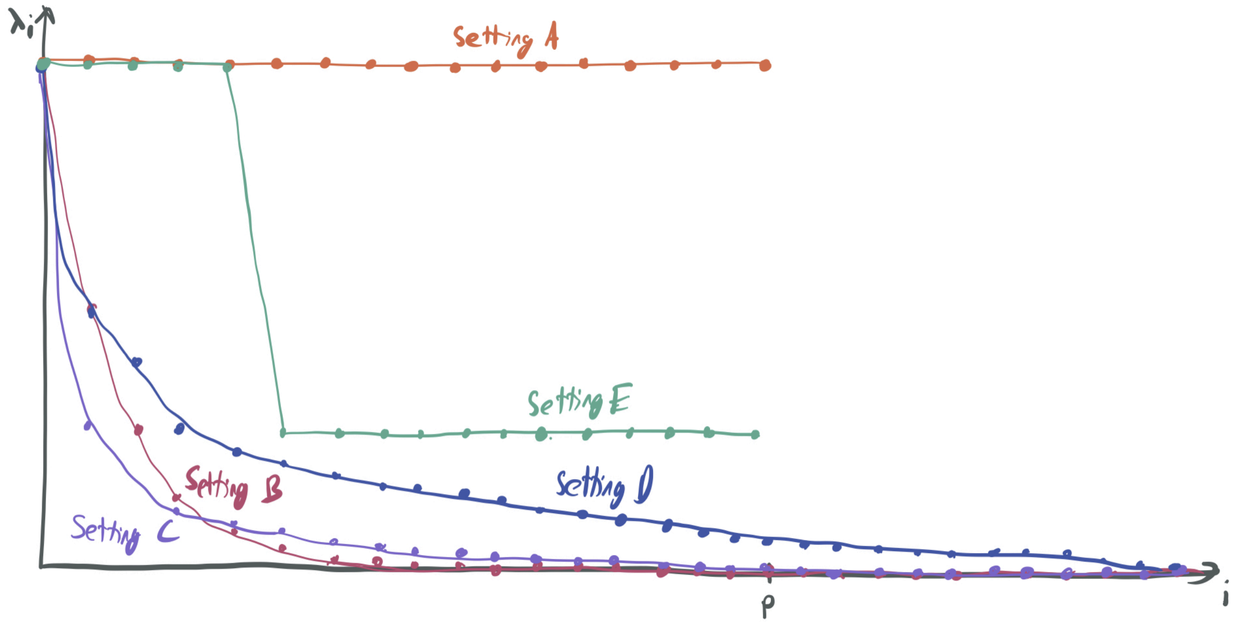
Benign overfitting: feature importance

[BLLT19]

- **Theorem**: With probability 0.99 for $k^* = \min\{k \geq 0 : r_k(\Sigma) \geq bn\}$, the excess risk is at most:

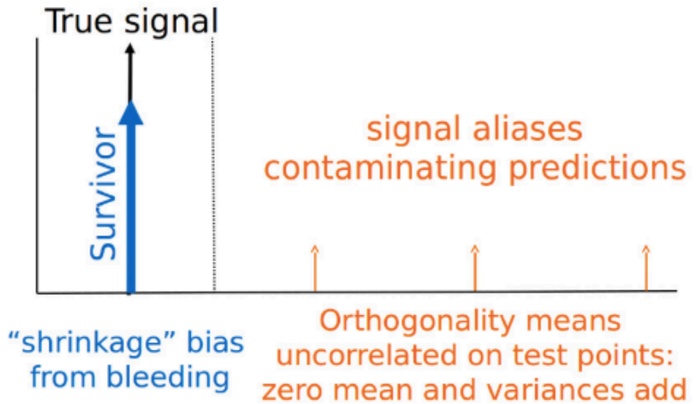
$$O\left(\|\theta^*\|^2 \lambda_1 \left(\sqrt{\frac{r_0(\Sigma)}{n}} + \frac{r_0(\Sigma)}{n} \right) + \sigma^2 \left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)} \right) \right).$$

- **Setting E: Bi-level**
  - $d = n \log n$
  - $\lambda_i = 1$ for $i \leq n/\log n$, $\lambda_i = 1/\log^2 n$ otherwise
  - $r_0(\Sigma) = \Theta(n/\log n)$.
  - $k^* = n/\log n$, $R_{k^*}(\Sigma) = \Theta(n \log n)$.
  - $O\left(\frac{\|\theta^*\|^2}{\sqrt{\log n}} + \frac{\sigma^2}{\log n} \right) \to 0$ as $n \to \infty$.



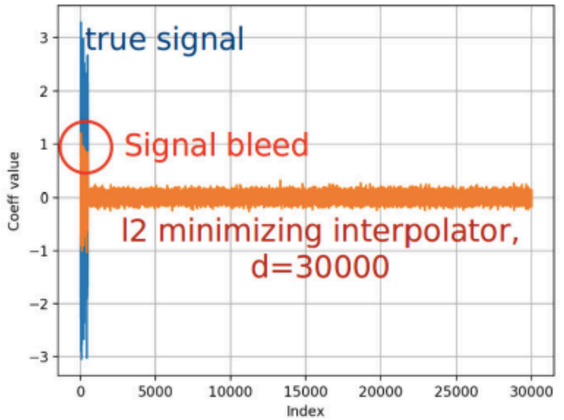
Benign overfitting: signal bleed and contamination [MVSS19]

- Interpolation involves choosing among many **aliases**, or different solutions with zero training error.
- **Signal bleed**: true signal dissipates into different aliases and chosen alias has little signal.
  - Avoided with small number of high-importance features.
- **Signal contamination**: chosen alias incorporates too much noise and info from irrelevant features.
  - Avoided with sufficiently slow decay of feature importance.



(a) Illustration of the "bleed".

Over-parameterized/



(b) Plot of estimated signal components of minimum-ℓ_2-interpolator for iid Gaussian features. Here, $n = 5000$, $d = 30000$ and the true signal $\boldsymbol{\alpha}^*$ has non-zero entries only in the first 500 features.

Model complexity

How can we align theory with practice?

Understanding tension between modern DL vs theory.

Theory of over-parameterization in simple models

Seeds for future theory for NN interpolation

1. Empirical NN results

2. Boosting

3. Linear regression

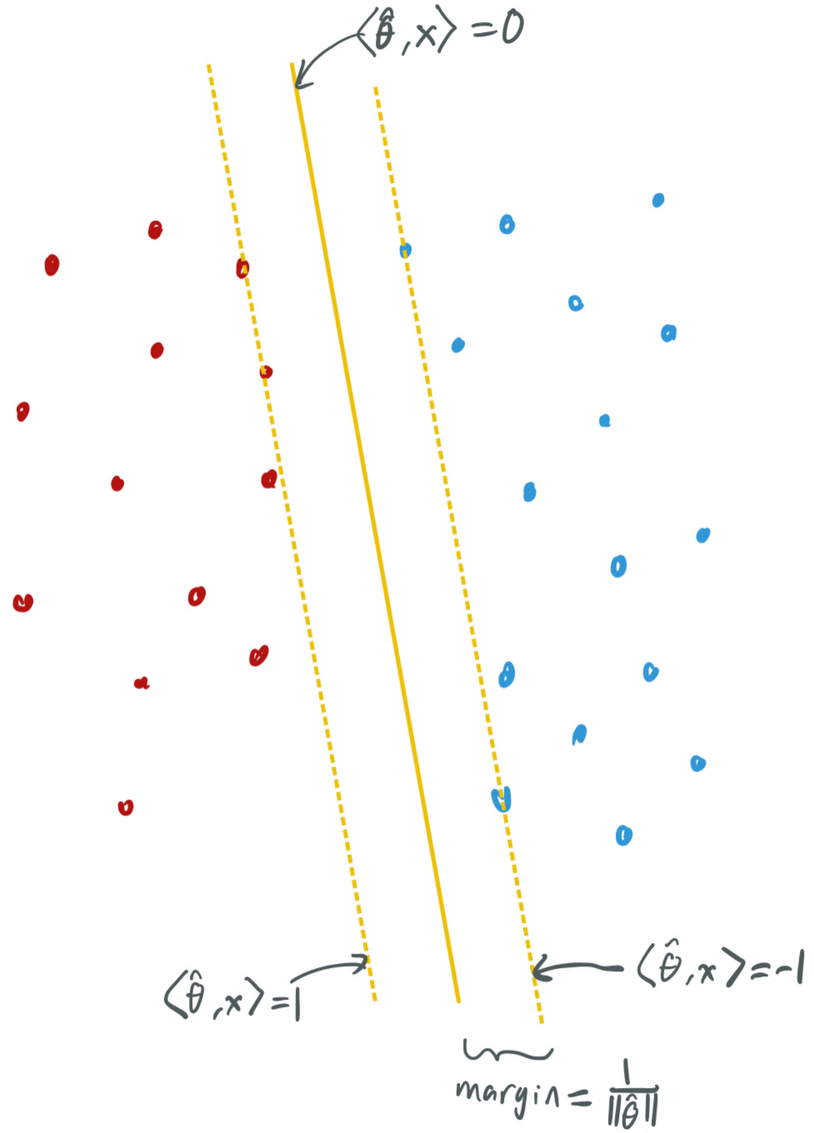
4. SVMs

5. Theoretical limitations of capacity-based bounds

6. “Nice” NN theoretical properties

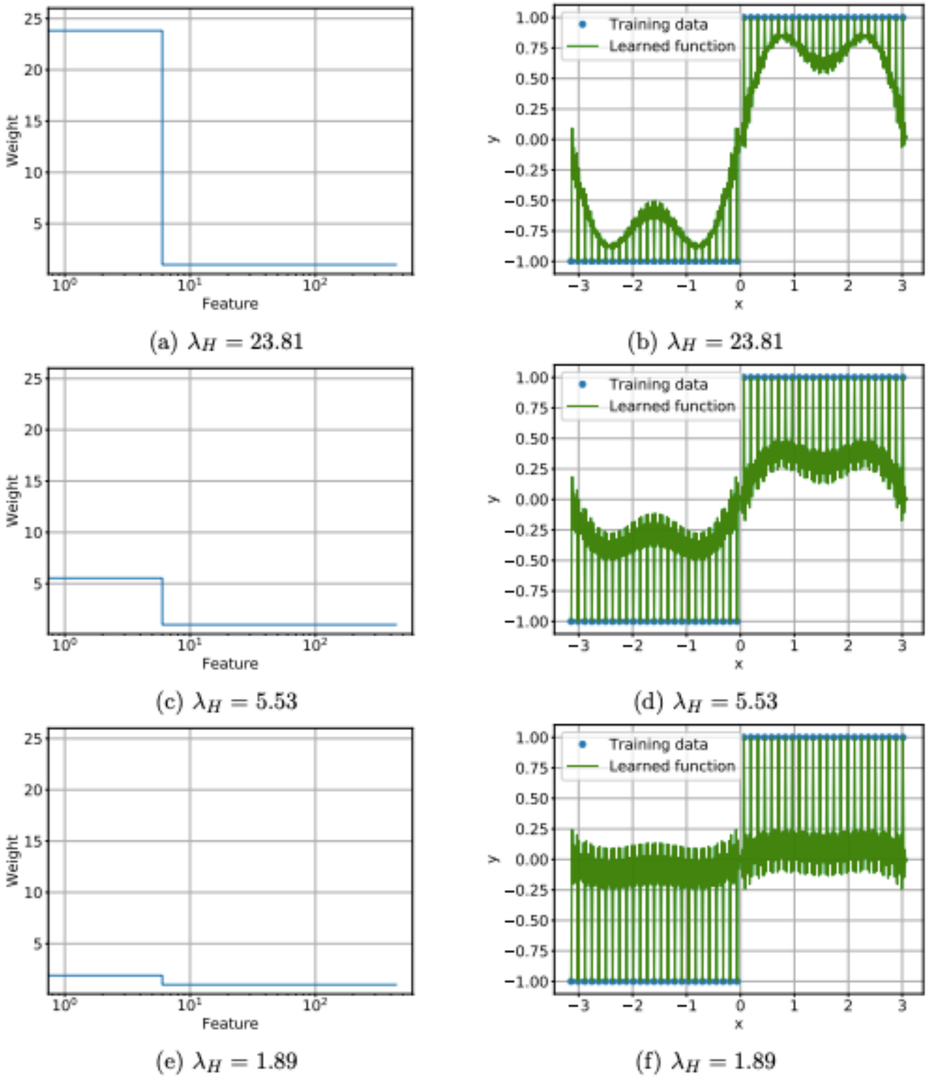
Hard SVM or maximum-margin classification

- Linearly separable
  $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^d \times \{-1, 1\}$.
- Learn $x \mapsto \text{sign}(\hat{\theta}^T x)$.
- $\hat{\theta} \in \mathbb{R}^d$ minimizes $\|\hat{\theta}\|$ such that
  $\hat{\theta}^T x_i \geq y_i$.
- x_i is a **support vector** if $\hat{\theta}^T x_i = y_i$.
- Classical generalization bounds rely on bounding number of support vectors.



SVM benign overfitting by connection to OLS [MNSBHS20]

- When $d = \Omega(n^{3/2} \log n)$, **support vector proliferation** occurs. (Every sample is a support vector **and** MNI = SVM).
  - By HMX21 and **ASH21**, SVP threshold at $\Theta(n \log n)$.
- Relates binary- and real-valued OLS generalization via survival and contamination analysis (like MVSS19).
  - Benign overfitting is "easier" in classification than regression.
- Like BLLT19, benign overfitting occurs when bi-level model does not have too slow a drop-off between high-importance and low-importance features.



(a) $\lambda_H = 23.81$ (b) $\lambda_H = 23.81$

(c) $\lambda_H = 5.53$ (d) $\lambda_H = 5.53$

(e) $\lambda_H = 1.89$ (f) $\lambda_H = 1.89$

SVM benign overfitting by gradient descent [CL20]

- By Soudry, et. al. (2018), gradient descent on separable data with logistic loss converges to hard-margin SVM.
- Benign overfitting for over-parameterized $d = \omega(n^2)$ data drawn from two clusters.
- Direct proof by showing that angle between true separator and learned separator is small.
  - Need to show that noisy samples do not have outsize influence on optimization process.

How can we align theory with practice?

Understanding tension between modern DL vs theory.

Theory of over-parameterization in simple models

Seeds for future theory for NN interpolation

1. Empirical NN results

2. Boosting

3. Linear regression

4. SVMs

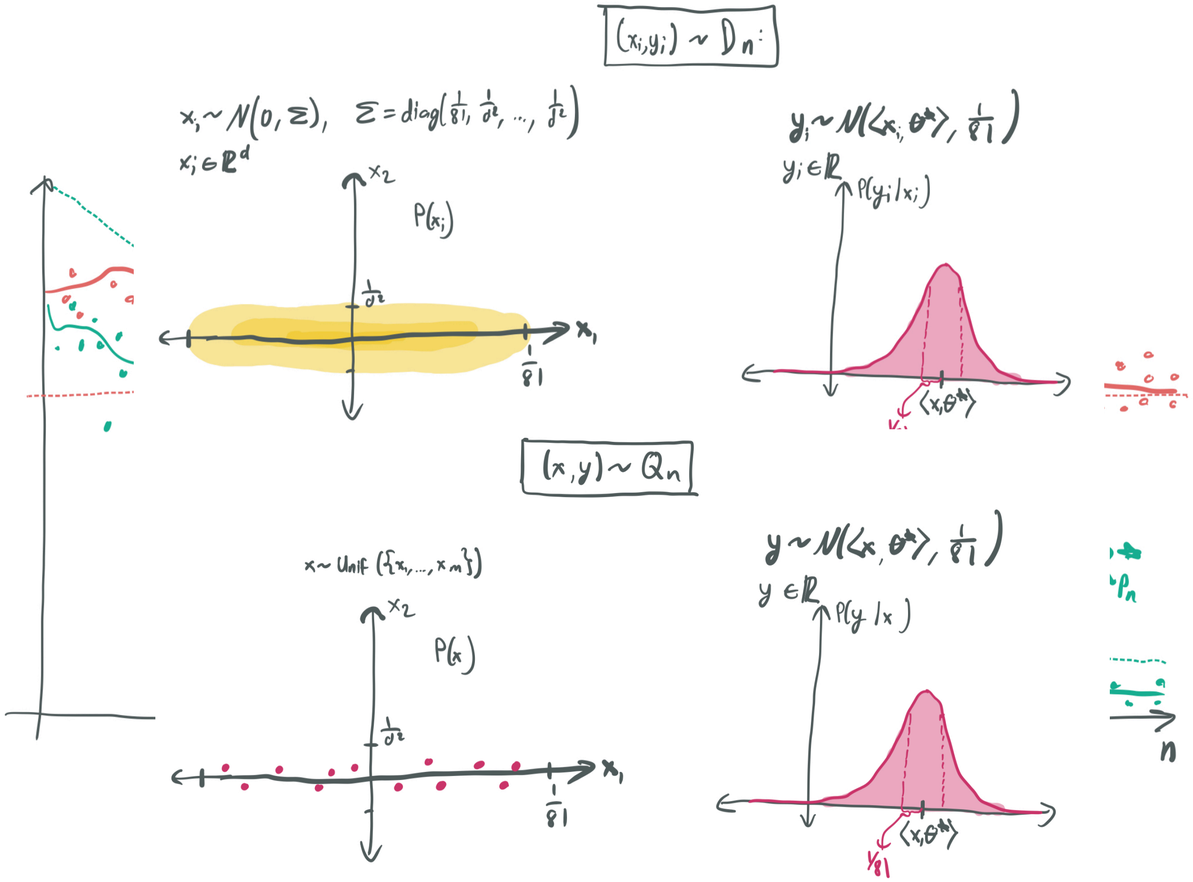
5. Theoretical limitations of capacity-based bounds

6. “Nice” NN theoretical properties

When capacity-based bounds fail

[BL20]

- Considers all valid model-dependent generalization bounds $\epsilon(h, n)$ that bound the excess risk with probability 0.9 for all data distributions P.
- For any bound ϵ and n, exists a distribution P_n over high-dimensional features where the least-norm interpolant h has excess error $O(1/\sqrt{n})$, but $\epsilon(h, n) \geq 1/2$.



How can we align theory with practice?

Understanding tension between modern DL vs theory.

Theory of over-parameterization in simple models

Seeds for future theory for NN interpolation

1. Empirical NN results

2. Boosting

3. Linear regression

4. SVMs

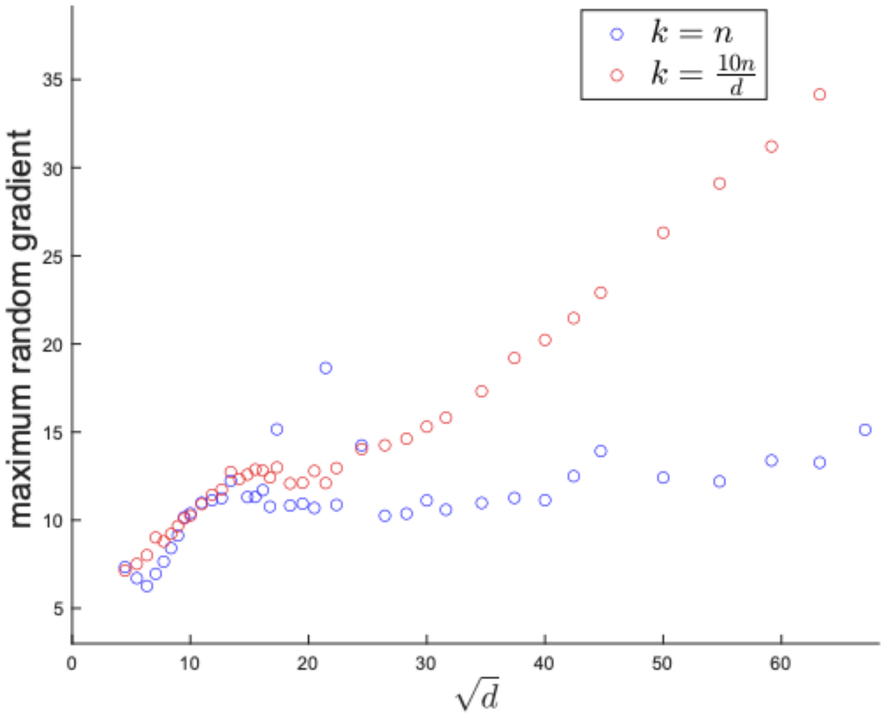
5. Theoretical limitations of capacity-based bounds

6. “Nice” NN theoretical properties

Smoothness and robustness in interpolating NNs

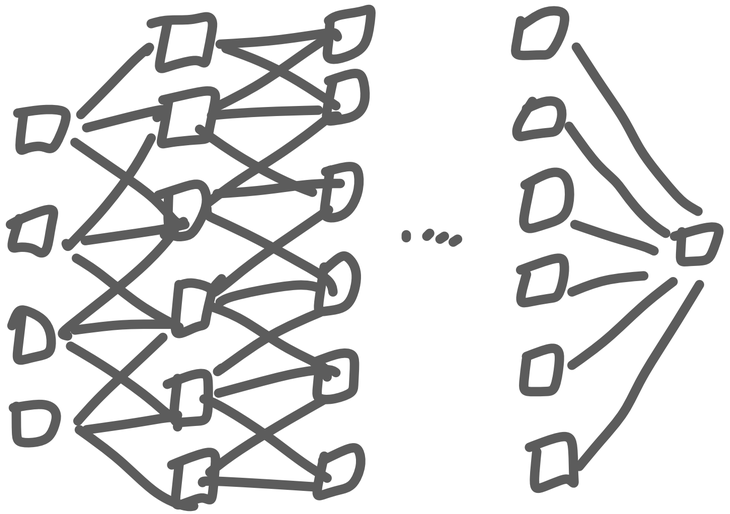
[BLN20, BS21]

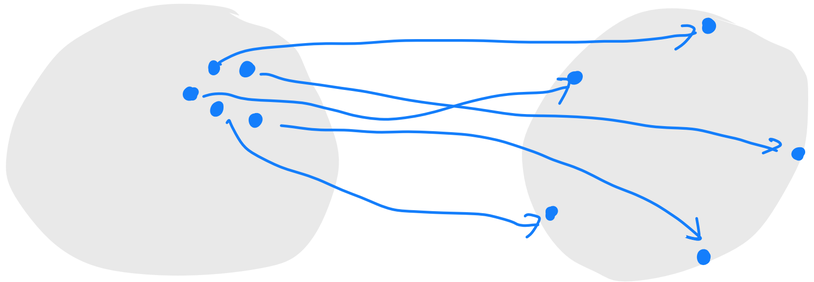
- **Conjecture**: WHP over sample, (1) exists 2-layer NN of width k interpolating n samples of Lipschitz constant $O(\sqrt{n/k})$, and (2) all interpolating NNs are $\Omega(\sqrt{n/k})$-Lipschitz.
- Weaker version of (1) from BLN20.
- Proof of (2) from BS21.



Conditioning from many random layers
[AAK20]

- Random layers orthogonalize inputs
- Implications for SQ learning, optimization.
- Benign overfitting in narrow case for deep features, but too coarse bias bound.
- Idea: Maybe intermediate features have small effective dimensional and favorable conditions for benign overfitting?





Last Thoughts

- **Narrative 1**: benign overfitting “decisive voting,” aggregation of many low-importance signals (boosting, linear regression variances, SVM features)
- **Narrative 2**: double-descent by ill-conditioning and easy approximation/ model simplicity (misspecified model, regularization)
- **Future work**: connections to NN robustness, explorations of connections between algorithms, more formulations of “simplicity.”

Thank you.

Appendix

Appendix

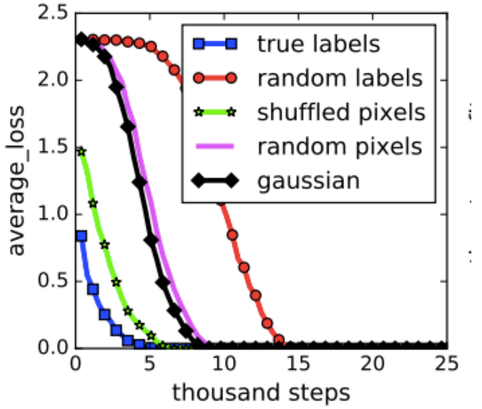
Paper Recaps

1. **Empirical Results for Neural Nets**: ZBHRV17, BMM18, BHMM19, NKBYBS19, SGDSBW19.
2. **Boosting**: FS97, BFLS98.
3. **Linear Regression**:
    A. **Minimum-norm Least-squares Regression**: BHX19, BLLT19, HMRT19, Mit19, MVSS19.
    B. **Spike Covariance and PCA**: WF17, MN19, XH19, HHV20.
    C. **Ridge Regression**: Zha05, CD07, DW15, TB20.
    D. **PAC-Bayesian Linear Regression**: AC10.
    E. **Random Feature Regression**: MM19.
    F. **Kernel Regression**: RZ19, LRZ19.
4. **Support Vector Machines**: MBSBHS20, CL20.
5. **Limitations of Capacity-Based Bounds**: BL20.
6. **Properties of Over-parameterized Neural Networks**: AAK20, BLN20, BS21.

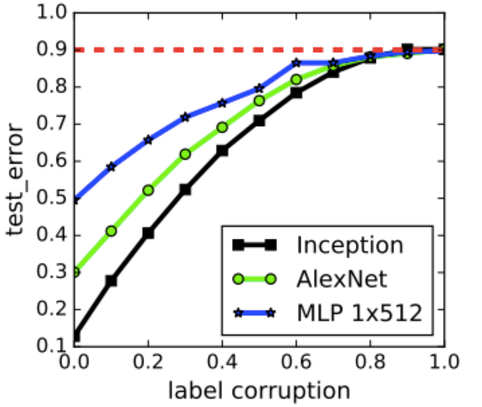
Appendix 1: Empirical Results for Neural Nets

Understanding deep learning requires rethinking generalization.
Zhang, Bengio, Hardt, Recht, and Vinyals (2017).

- **Results**
  - Empirically, NNs trained on random labels can still reach zero training error with more training steps. (a)
  - NNs trained with some fraction of corrupted labels still generalize to new data. (c)
- **Implications**
  - NNs have very high “effective capacity” for fitting arbitrary image datasets, and optimization algorithms continue to work.
  - A working theory of NN generalization will not rely purely on model capacity and explicit regularization.
  - Not all models that fit training data in over-parameterized NNs generalize well; implicit regularization of learning algorithms is key in choosing the right interpolation!



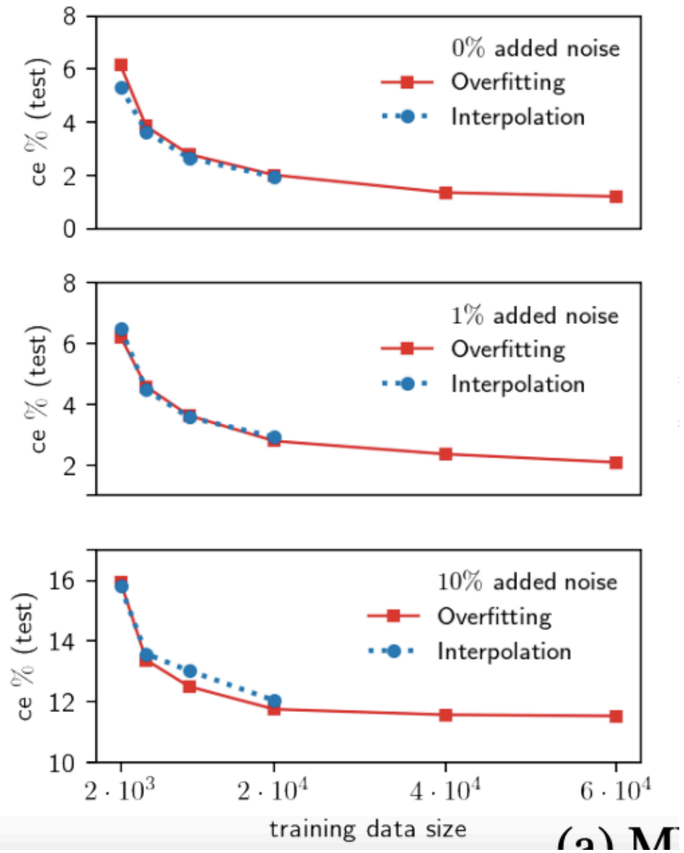
(a) learning curves



(c) generalization error growth

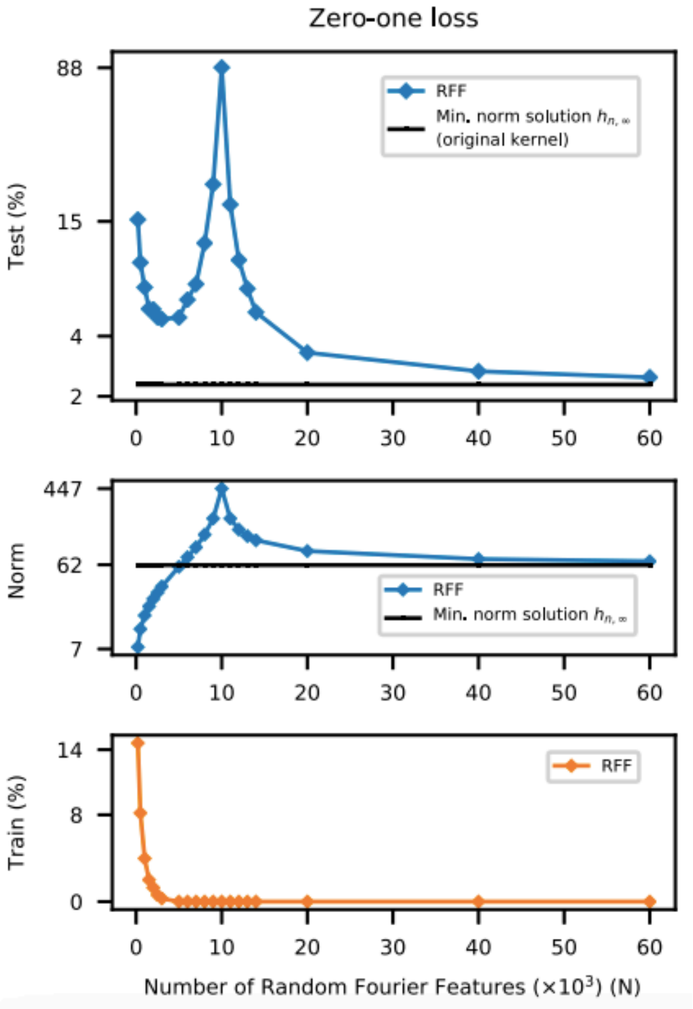
To Understand Deep Learning We Need to Understand Kernel Learning. Belkin, Ma, Mandal (2018).

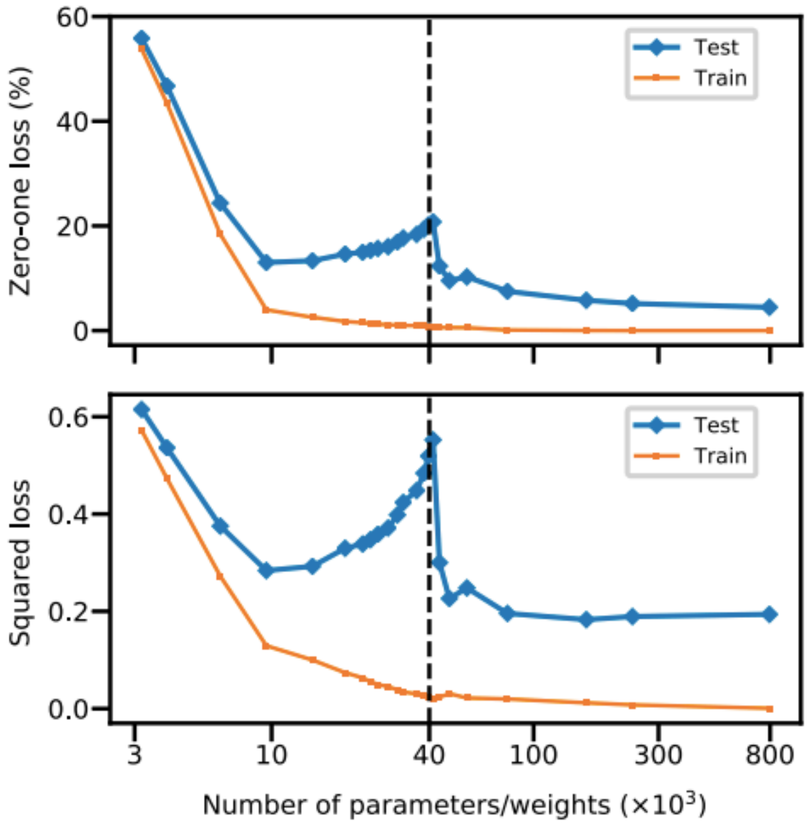
- **Results**
  - Kernel methods can have zero training error and small generalization error, like neural networks. Empirical observation that Laplacian kernels can also easily fit random labels.
  - Theorem 1: Any kernel method overfitting data from some distribution must have a high function norm and hence not be “simple” as required for fat-shattering generalization bounds.
  - Generalization decays gradually as label noise increases.
- **Implications**
  - Many of the issues described by ZBHRV17 for generalization in NNs apply to kernel classifiers too.
  - Capacity-based generalization approaches (e.g. VC-dimension, Rademacher complexity) also don’t explain generalization performance of kernel classifiers.
  - No known generalization bounds can explain noisy generalization, since it would need to be between positive Bayes error and 1 to not be vacuous.



Reconciling modern machine-learning practice and the classical bias–variance trade-off. Belkin, Hsu, Ma, Mandal (2019).

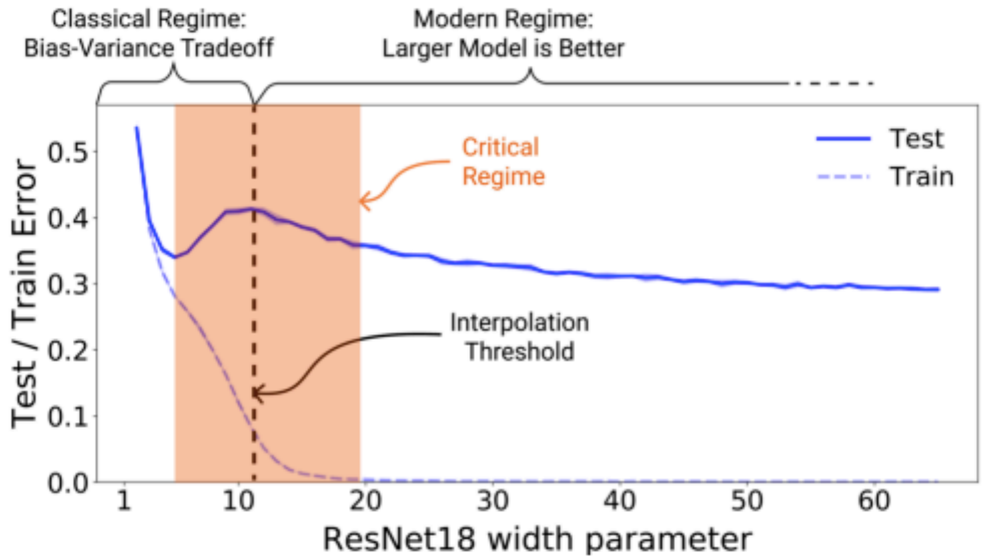
- **Results**
  - Double-descent empirically occurs for random Fourier features, two-layer ReLU networks, and random forests.
- **Implications**
  - Inductive biases of small norms and optimization algorithms encourages good generalization in interpolation regime.

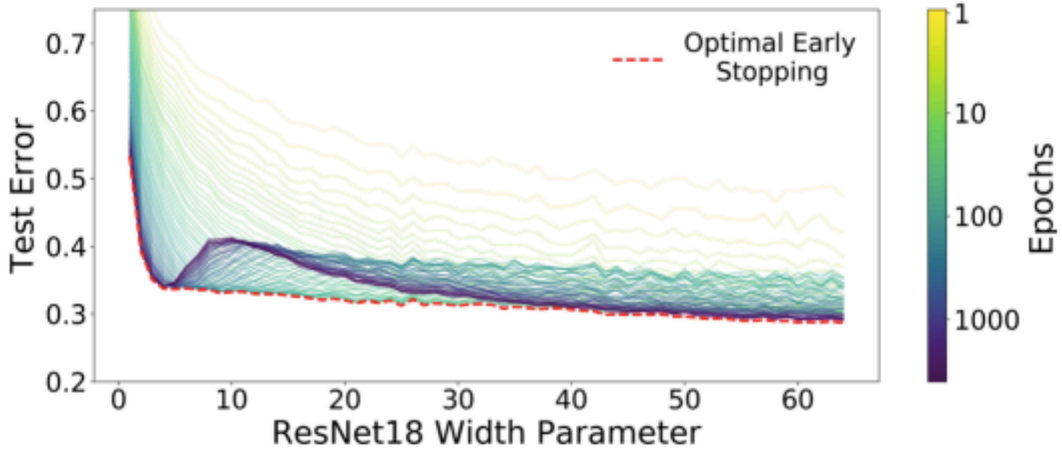




Deep double descent: Where bigger models and more data hurt. Nakkiran, Kaplun, Bansal, Yang, Barak, and Sutskever (2019).

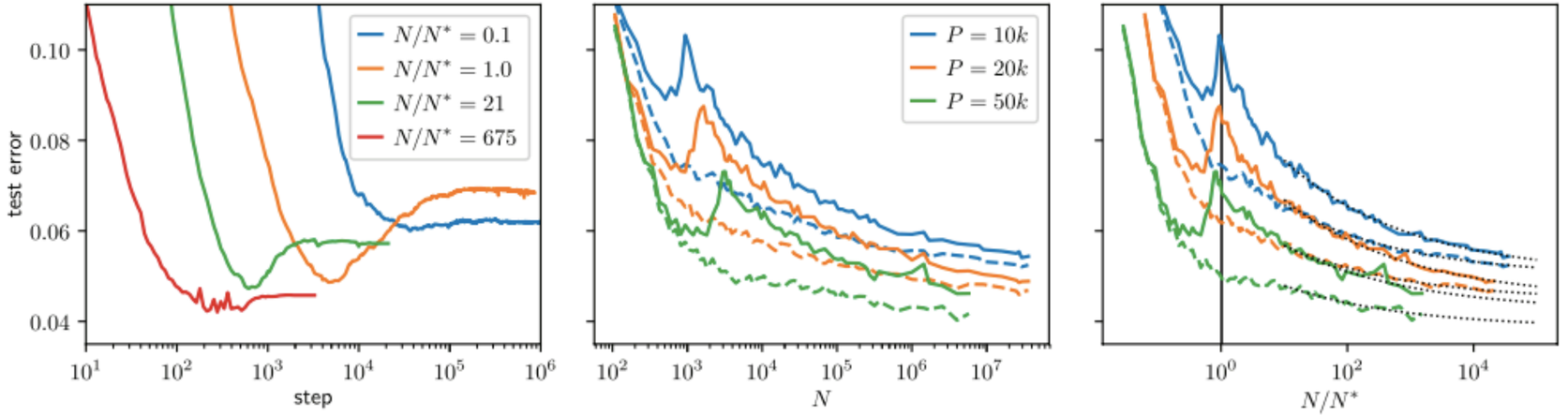
- **Results**
  - Observed double-descent phenomenon for deep neural networks with respect to changes in sample size, model size, and number of epochs.
  - Early stopping can mitigate double-descent in NNs.
  - Double descent curve holds up even in presence of (small amount of) label noise.





A jamming transition from under- to over-parametrization affects generalization in deep learning. Spigler, Geiger, d'Ascoli, Sagun, Biroli, and Wyart (2019).

- **Results**
  - There exists a "jamming transition" in neural networks that causes the generalization error to spike when the number of parameters is approximately the number of samples; similar shape to double-descent.
  - In over-parametrized regime, unlikely to have any bad local minima because constraints make these stable minima unlikely.
  - Early stopping can mitigate the spike of the jamming transition.
- **Techniques**
  - Statistical physics methods, based on the analysis of the dimensionality of manifolds.
  - Analogy between NN and *glasses*, physical systems with exponentially many local minima.



Appendix 2: Boosting

A decision-theoretic generalization of online learning and an application to boosting. Freund and Schapire (1997).

- **Results**
  - Introduces *AdaBoost* algorithm, which combines weak learners of various effectiveness from reweighed data distributions into a single voting-based classifier.
  - Theorem 6: Bound on training error of AdaBoost with T weak learners.
  - Theorems 7 + 8: VC generalization bounds suggest classical capacity and overfitting tradeoffs as T increases.
  - Empirically, however, generalization of AdaBoost improves monotonically with T, even after training error approaches zero.
- **Implications**
  - Poses question about limitations of capacity-based generalization for analyzing some interpolating classifiers.

Algorithm AdaBoost

Input: sequence of N labeled examples $\langle (x_1, y_1), ..., (x_N, y_N) \rangle$
distribution D over the N examples
weak learning algorithm **WeakLearn**
integer T specifying number of iterations

Initialize the weight vector: $w_i^1 = D(i)$ for $i = 1, ..., N$.

Do for $t = 1, 2, ..., T$

1. Set

$$\mathbf{p}^t = \frac{\mathbf{w}^t}{\sum_{i=1}^N w_i^t}$$

2. Call **WeakLearn**, providing it with the distribution $\mathbf{p}^t$; get back a hypothesis $h_t: X \to [0, 1]$.
3. Calculate the error of h_t: $\varepsilon_t = \sum_{i=1}^N p_i^t |h_t(x_i) - y_i|$.
4. Set $\beta_t = \varepsilon_t / (1 - \varepsilon_t)$.
5. Set the new weights vector to be

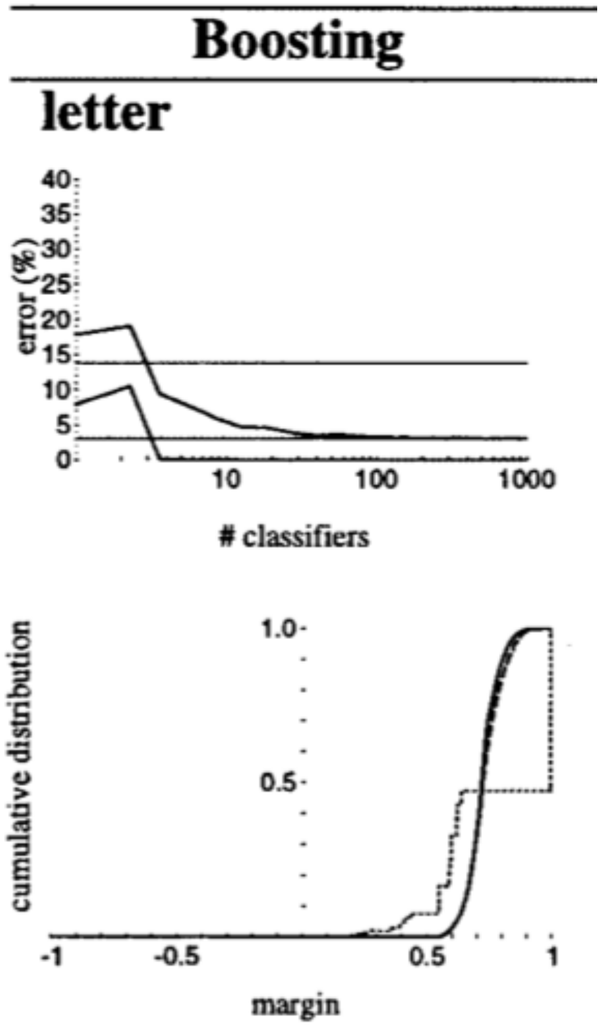
$$w_i^{t+1} = w_i^t \beta_t^{1 - |h_t(x_i) - y_i|}$$

Output the hypothesis

$$h_f(x) = \begin{cases} 1 & \text{if } \sum_{t=1}^T (\log 1/\beta_t) h_t(x) \geqslant \frac{1}{2} \sum_{t=1}^T \log 1/\beta_t \\ 0 & \text{otherwise.} \end{cases}$$

Boosting the margin: a new explanation for the effectiveness of voting methods. Bartlett, Freund, Lee, Schapire (1998).

- **Results**
  - Theorem 2: Voting classifiers that correctly classify training samples *with large margin* θ have generalization bounds that do not depend on the number of constituent classifiers.
    - Approximates voting classifier with a vote by a random sample of constituents.
    - Decomposes test error with conditional probability and bounds each term with concentration bounds.
  - Theorem 4: AdaBoost run for sufficiently many rounds T correctly classifies all training samples with margin θ.
    - Proof similar to AdaBoost convergence result from FS97 by bounding
- **Implications**
  - Explains gap between empirical generalization performance and VC generalization bounds for AdaBoost (FS97).
  - Margin θ continues to grow with T, even as training error equals zeros; hence, continued generalization improvements are explained.



Appendix 3: Linear Regression

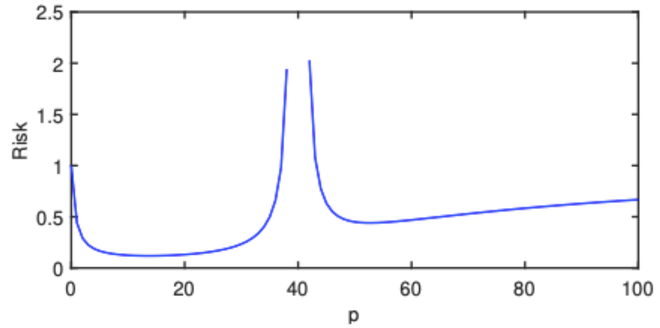
Appendix 3A: Minimum-norm Least-squares Regression

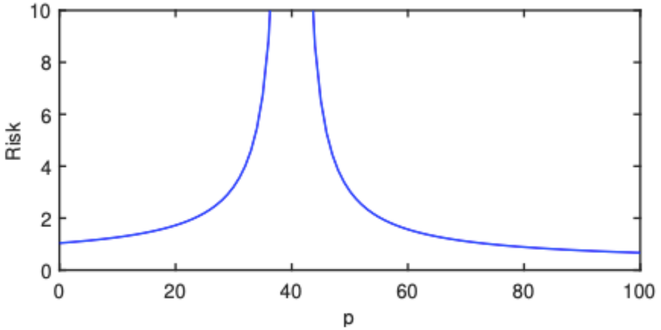
Two models of double descent for weak features. Belkin, Hsu, Xu (2019).

- **Results**
  - Considers the *misspecified model* for minimum-norm interpolation (and least-squares regression) Gaussian and Fourier features where there are N total features, of which the d most significant features are provided to the learner.
  - Generalization bounds depend on bias-variance decomposition.
    - Bias increases with d, but is bounded because true weights β remain close to row space of samples X.
    - Variance decreases with d beyond over-parameterization, $d > n$.
- **Implications**
  - Double descent can occur as the number of features d grows when each new improves the model's abilities to approximate the true function.
  - Does not rely on bounds on effective dimension of feature distribution.
  - In the “prescient feature model” where d features have highest variance, optimal behavior occurs in classical regime, despite second descent.

Theorem 1. *Assume the distribution of $\boldsymbol{x}$ is the standard normal in $\mathbb{R}^D$, ϵ is a standard normal random variable independent of $\boldsymbol{x}$, and $y = \boldsymbol{x}^*\boldsymbol{\beta} + \sigma\epsilon$ for some $\boldsymbol{\beta} \in \mathbb{R}^D$ and $\sigma > 0$. Pick any $p \in \{0, \dots, D\}$ and $T \subseteq [D]$ of cardinality p. The risk of $\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}_T = \boldsymbol{X}_T^\dagger \boldsymbol{y}$ and $\hat{\boldsymbol{\beta}}_{T^c} = \mathbf{0}$, is*

$$\mathbb{E}[(y - \boldsymbol{x}^*\hat{\boldsymbol{\beta}})^2] = \begin{cases} (\|\boldsymbol{\beta}_{T^c}\|^2 + \sigma^2) \cdot \left(1 + \frac{p}{n-p-1}\right) & \text{if } p \le n-2; \\ +\infty & \text{if } n-1 \le p \le n+1; \\ \|\boldsymbol{\beta}_T\|^2 \cdot \left(1 - \frac{n}{p}\right) + (\|\boldsymbol{\beta}_{T^c}\|^2 + \sigma^2) \cdot \left(1 + \frac{n}{p-n-1}\right) & \text{if } p \ge n+2. \end{cases}$$





Benign overfitting in linear regression. Bartlett, Long, Lugosi, Tsigler (2019).

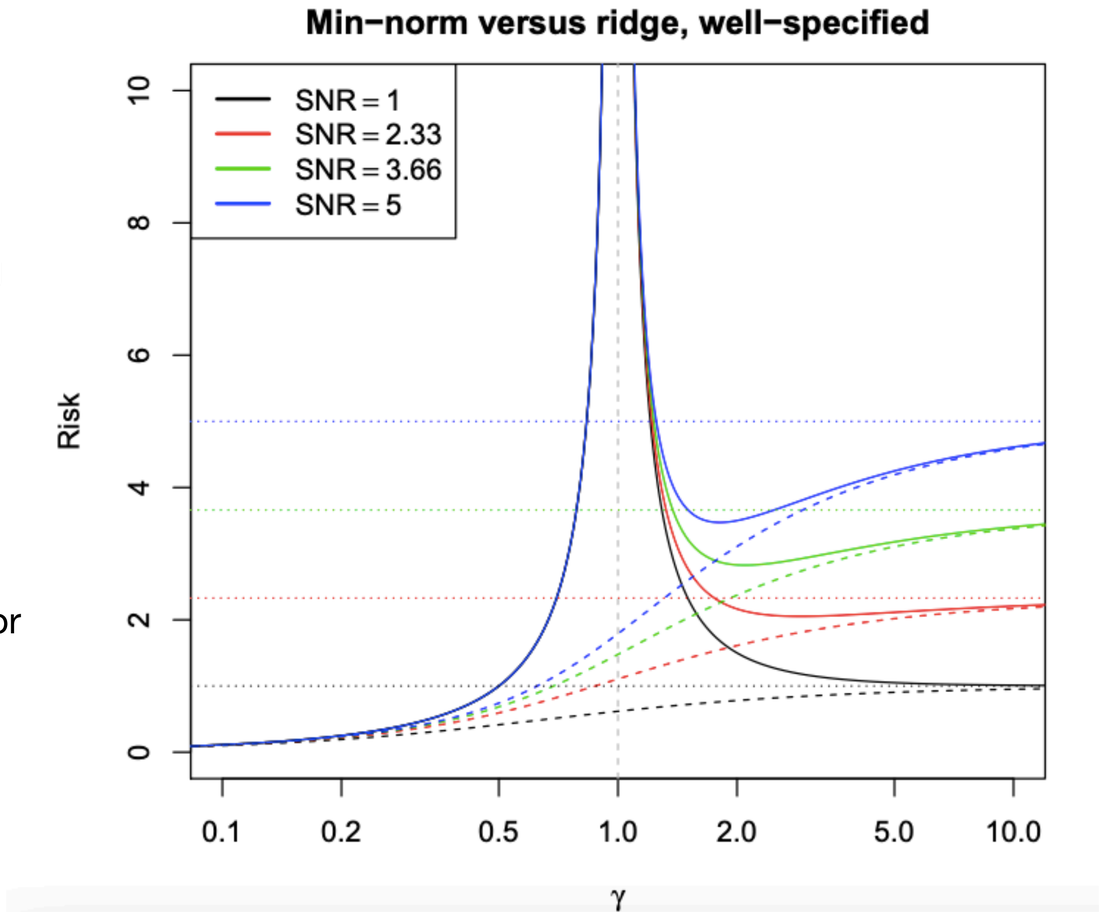
- **Results**
  - Minimum-norm interpolation for subgaussian inputs with covariance Σ (with eigenvalues $\lambda_1 > \lambda_2 > \ldots$), optimal weights θ^*, and subgaussian noise σ.
  - Depends on *effective ranks* of Σ: $r_k(\Sigma) = \sum_{i>k} \lambda_i / \lambda_{k+1}$ and $R_k(\Sigma) = (\sum_{i>k} \lambda_i)^2 / \sum_{i>k} \lambda_i^2$
  - With probability 0.99 for $k^* = \min\{k \geq 0 : r_k(\Sigma) \geq bn\}$, the excess risk is at most:

$$O\left(\|\theta^*\|^2 \lambda_1 \left(\sqrt{\frac{r_0(\Sigma)}{n}} + \frac{r_0(\Sigma)}{n}\right) + \sigma^2 \left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}\right)\right)$$

    - If λ_i's decay rapidly, then effective rank is small and $n \gg R_{k^*}(\Sigma) \implies$ vacuous bound.
    - If λ_i's decay slowly, then effective rank is large and $r_0(\Sigma) \gg n \implies$ vacuous bound.
    - Risk approaches 0 with increased n when there are $o(n)$ “high-importance" features with large eigenvalues (which θ^* depnds on) and $\omega(n)$ “lower-importance" features. (Bias in favor of predictions that pay attention to high-importance features.)
    - Bound by bias-variance decomposition and concentration bounds based on spectrum and analysis of projection operator onto row space of training data.
  - Includes a matching lower bounds with the RHS term.
- **Implications**
  - “Goldilocks” phenomenon for properly specified over-parameterized models; spectrum of features must decay neither too rapidly nor too gradually.
  - Over-parameterization is essential for benign overfitting; cannot occur if number of significant directions isn't much larger than number training samples.

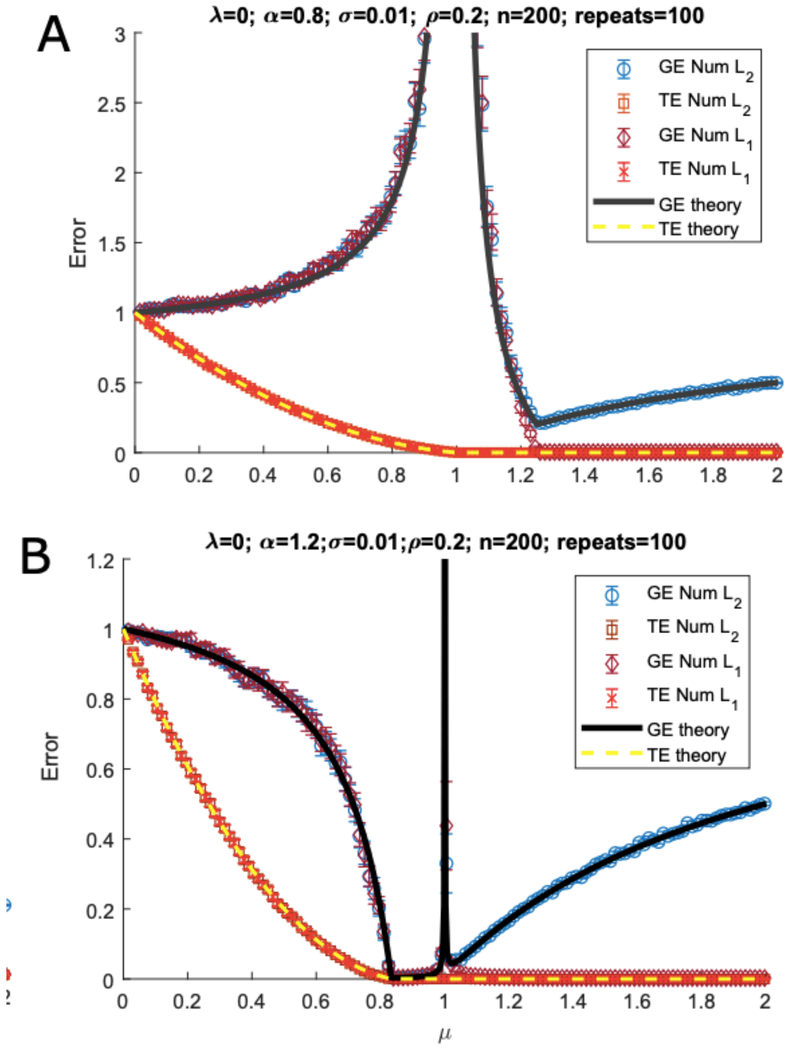
Surprises in high-dimensional ridgeless least squares interpolation. Hastie, Montanari, Rosset, Tibshirani (2019).

- **Results**
  - Considers infinite limit setting for number of samples n and features d; $n \to \infty$ and $d = \gamma n$, where γ quantifies degree of over-parameterization. Instead of analyzing fixed eigenvalues, examines limiting distribution of eigenvalues.
  - Provides very general version of BLLT19 generalization bounds by showing convergence of bias and variance to fixed quantities that depend on Marchenko-Pastur distribution of eigenvalues.
  - Results rely on showing similarity to limiting ridge regression for small regularization parameter λ.
- **Implications**
  - Similar to BLLT19: need to have many unimportant low-variance features for benign overfitting to occur.
  - Similar to HMX19 for misspecified model double descent with isotropic features.
  - Properly-tuned ridge regression dominates over-parameterized minimum-norm regression.



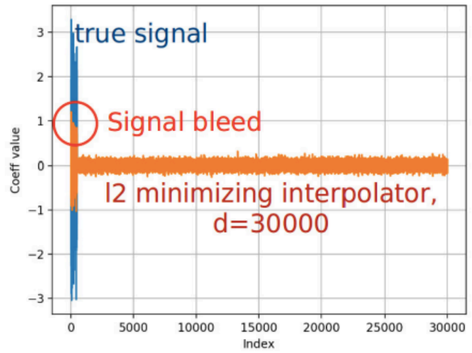
Understanding overfitting peaks in generalization error: Analytical risk curves for ℓ_2 and ℓ_1 penalized interpolation. Mitra (2019).

- **Results**
  - Considers misspecified model of HMX19 and HMRT19 for both ridge (ℓ_2-penalized) and lasso (ℓ_1-penalized) regression in the asymptotic regime where number of samples n, available features d, and total features N go to infinity proportional to one another. Parameter vectors are randomly drawn and sparse.
  - Obtain analytical expression for generalization under simple data model with Gaussian features.
  - Increasing regularization λ eliminates the overfitting peak (and double-descent) in both settings.
  - There is a large interval beyond the interpolation threshold where ℓ_1-penalized regression generalizes and ℓ_2 does not.
- **Implications**
  - Reinforces notion that ℓ_1 inductive biases reward sparsity, which helps in this case with sparse weights.



Harmless interpolation of noisy data in regression. Muthukumar, Vodrahalli, Subramanian, Sahai (2019).

- **Results**
  - Theorem 1: Like lower bound in BLLT, shows that benign overfitting can only occur when a model is sufficiently over-parameterized by lower-bounding the generalization abilities of interpolating solutions as roughly $\Omega(n/d)$.
  - Frames benign overfitting in terms of signal processing and characterizes failure modes.
    - Interpolation involves choosing among many *aliases*, solutions with zero training error that are only distinguished by inductive bias of algorithm (here, minimum norm).
    - If just barely over-parameterized, exist few interpolating aliases; unlikely they'll be any good.
    - Signal Bleed: True signal dissipates into orthogonal aliases… no alias will preserve enough signal! Occurs when insufficient bias for important features. (Need small number of high-importance features from BLLT)
    - Signal Contamination: Too much noise is incorporated into the chosen alias. Can prevent by dissipating noise, which is ensured by having eigenvalues not decrease too rapidly.
- **Implications**
  - Connects double-descent to signal processing and proposes a relatively simple framework for why interpolation succeeds only sometimes.



(b) Plot of estimated signal components of minimum-ℓ_2-interpolator for iid Gaussian features. Here, $n = 5000$, $d = 30000$ and the true signal $\boldsymbol{\alpha}^*$ has non-zero entries only in the first 500 features.

Appendix 3B: Spike Covariance and PCA

Asymptotics of empirical eigenstructure for high dimensional spiked covariance. Wang and Fan (2017).

- **Results**
  - Introduces the *spiked covariance* model, whose covariance matrix Σ has a constant number of large (and decaying) eigenvalues and all other eigenvalues are much smaller.
  - Demonstrates that empirical estimates $\hat{\Sigma}$ have biased predictions of largest eigenvalues and eigenvectors and gives a new algorithm to counter this bias and provide better estimates.
- **Implications**
  - Sets the stage for MN19 work on benign overfitting in the spiked covariance model.
  - Influences XH19 and HHV20 by suggesting the limitations of PCA when covariance spectrum has a sharp drop-off.

Risk of least-squares minimum-norm estimator under the spike covariance model. Mahdaviyeh and Naulet (2019).

- **Results**
  - Considers over-parameterized regime for infinite limit of n and d with spike covariance, where constant number of features have increasing eigenvalues and others are smaller. Eigenvalues do not necessarily decay to zero.
  - BLLT19 bounds hold in this setting, but provide tighter analysis of bias due to known properties of spike covariance matrix from past work. Leverages separation between eigenvalues to obtain much tighter bounds by separating the eigenspaces of each high-importance feature direction.
  - Bounds hold if first eigenvalue is large enough and $d \gg n$. Asymptotic bounds require only bounds on moments, not necessarily
- **Implications**
  - Demonstrates clear family of covariances with much clearer benign overfitting behavior than BLLT19.

On the number of variables to use in principal component regression. Xu and Hsu (2019).

- **Results**
  - Exhibits double-descent for *Principal Component Regression*, where PCA is used to reduce the dimensionality of the inputs and minimum-norm OLS is applied to the low-dimensional inputs. Uses atypical version of PCA where access to covariance matrix Σ is assumed without needing to approximate it.
    - Proofs rely on similar decomposition to HMX19, but with incorporation of distribution over spectrum as parameters go to infinity.
  - Assumes Gaussian features with strictly decreasing eigenvalues of Σ, noiseless labels, and randomly chosen weights. Number of total features N, PCA'd features d, and samples n approach infinity at fixed ratios.
  - In noisy setting, generalization in interpolation regime can only outperform classical regime if eigenvalues decay sufficiently slowly.
- **Implications**
  - Extends double-descent to a different model of linear regression.
  - Informs design choices on how to choose d for PCR.

Dimensionality reduction, regularization, and generalization in overparameterized regressions. Huang, Hogg, Villar (2020).

- **Results**
  - Somewhat counter to XH19, presents bounds that PCA-OLS (or PCR) has no interpolation peaks when PCA is performed on the empirical covariance estimate $\hat{\Sigma}$, rather than the true covariance Σ. Follows because variance can be upper-bounded by a monotonic quantity.
  - Empirical PCA avoids peaking by preventing the condition number of $X^T X$ from becoming large at the interpolation threshold $n \approx p$.
- **Implications**
  - PCA is presented as a way to *avoid* double-descent and have no spike at the interpolation threshold.
  - Peaking in misspecified model of BHX19 is caused by large expected condition number of $X^T X$, and regularization can prevent the peak by reducing the condition number.
  - Regularization can be conceived of more broadly than just norm constraints in objective function.

Appendix 3C: Ridge Regression

Learning bounds for kernel regression using effective data dimensionality. Zhang (2005).

- **Results**
  - Introduces effective dimension quantity $D_\lambda = \operatorname{tr}((\mathbb{E}[\psi\psi^T] + \lambda I)^{-1}\mathbb{E}[\psi\psi^T])$ for (possibly infinite-dimensional) features ψ and regularization parameter λ.
    - $D_\lambda \to \operatorname{rank}(\mathbb{E}[\psi\psi^T])$ as $\lambda \to 0$.
    - $D_\lambda \to 0$ as $\lambda \to \infty$.
    - $D_\lambda = O(\sqrt{n})$ in the worst case.
  - Proves generalization bound on λ-regularized empirical risk-minimizing classifier. Excess error is approximately $\min_{\lambda>0} O(\lambda + D_\lambda/n)$.
- **Implications**
  - The fact that some kernel methods map to an infinite-dimensional vector space and generalize does *not* mean that they conquer the curse of dimensionality; rather, they have a small effective dimension.
  - Sometimes, benign overfitting can be explained by the data being intrinsically low-dimensional.
  - Unlike future benign-overfitting bounds, generalization error bound applies to all classifiers in some convex space with bounded norm.
  - Kernel classifiers can be simplified computationally by only studying the directions of the largest eigenvalues.

Optimal rates for the regularized least-squares algorithm. Caponnetto and De Vito (2007).

- **Results**
  - For ridge regression with expected squared loss and regularization parameterized by λ and bounded feature distributions with subgaussian noise, near-matching upper and lower bounds on population error.
  - Bounds identify an optimal setting for λ as a function of number of training samples n and hold in the limit where $n \to \infty$.
  - Uses same formulation of effective dimension as Zha05 in proof of bounds; higher effective dimension $\Longrightarrow$ slower rate of convergence with n.
- **Implications**
  - Illustration of strong bounds for minimum-norm regression in the classical regime.
  - Limiting behavior of n and fixed (effective) dimension excludes benign overfitting bounds and rather demonstrates optimal behavior in classical regime.

Theorem 1. *Given* $1 < b \leq +\infty$ *and* $1 \leq c \leq 2$, *let*

$$\lambda_\ell = \begin{cases} (1/\ell)^{b/(bc+1)}, & b < +\infty,\ c > 1, \\ (\log \ell/\ell)^{b/(b+1)}, & b < +\infty,\ c = 1, \\ (1/\ell)^{1/2}, & b = +\infty, \end{cases}$$

and

$$a_\ell = \begin{cases} (1/\ell)^{bc/(bc+1)}, & b < +\infty,\ c > 1, \\ (\log \ell/\ell)^{b/(b+1)}, & b < +\infty,\ c = 1, \\ 1/\ell, & b = +\infty, \end{cases}$$

then

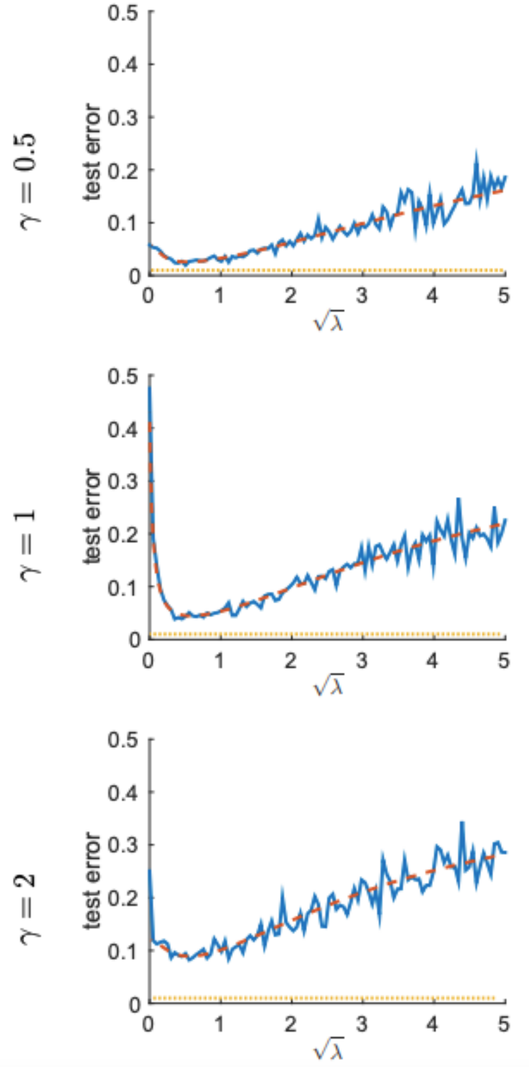
$$\lim_{\tau \to \infty} \limsup_{\ell \to \infty} \sup_{\rho \in \mathcal{P}(b,c)} \mathbb{P}_{\mathbf{z} \sim \rho^\ell}[\mathcal{E}[f_{\mathbf{z}}^{\lambda_\ell}] - \mathcal{E}[f_{\mathcal{H}}] > \tau a_\ell] = 0.$$

Theorem 2. *Assume that* $\dim Y = d < +\infty$, $1 < b < +\infty$ *and* $1 \leq c \leq 2$, *then*

$$\lim_{\tau \to 0} \liminf_{\ell \to +\infty} \inf_{f_\ell} \sup_{\rho \in \mathcal{P}(b,c)} \mathbb{P}_{\mathbf{z} \sim \rho^\ell}[\mathcal{E}[f_{\mathbf{z}}^{\ell}] - \mathcal{E}[f_{\mathcal{H}}] > \tau \ell^{-bc/(bc+1)}] = 1.$$

High-dimensional asymptotics of prediction: ridge regression and classification. Dobriban and Wagner (2015).

- **Results**
  - Asymptotic ridge regression results for n and d approaching infinity at a fixed rate $\gamma \in (0,\infty)$ for features from distribution with bounded moments.
  - Relies on limiting distribution of eigenvalues as summarized by the *Stieltjes transform*, which captures a notion of effective dimension. Provides the optimal regularization parameter λ^* based on the degree of over-parameterization and the signal-to-noise ratio.
- **Implications**
  - Similar types of results as HMRT19, but with a simpler covariance structure and without considering minimum-norm interpolation with $\lambda \to 0$.
  - λ^* increases as γ increases, which means it says little about over-parameterized minimum-norm interpolation, since the functions are distinctly different.



Benign overfitting in ridge regression. Tsigler and Bartlett (2020).

- **Results**
  - Similar story as BLLT19, except considers ridge regression rather than minimum-norm interpolation.
  - Very similar bounds, except that effective ranks include the regularization term λ in the numerator: $r_k(\Sigma) = (\lambda + \sum_{i>k} \lambda_i)/\lambda_{k+1}$ and $R_k(\Sigma) = (\lambda + \sum_{i>k} \lambda_i)^2 / \sum_{i>k} \lambda_i^2$.
  - Increasing λ improves the variance bound and worsens the bias bound of BLLT19.
- **Implications**
  - Very mild regularization λ can mitigate the problems minimum-norm interpolation faces when features decay too rapidly
  - Generalization of BLLT19 bounds.

Appendix 3D: PAC-Bayesian Linear Regression

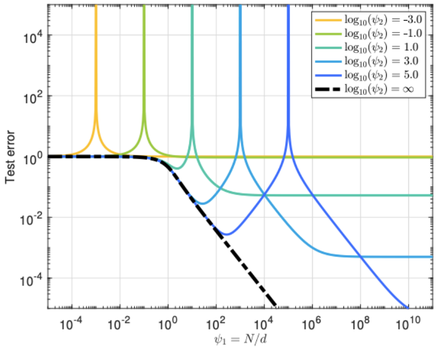
Linear regression through PAC-Bayesian truncation. Audibert and Catoni (2010).

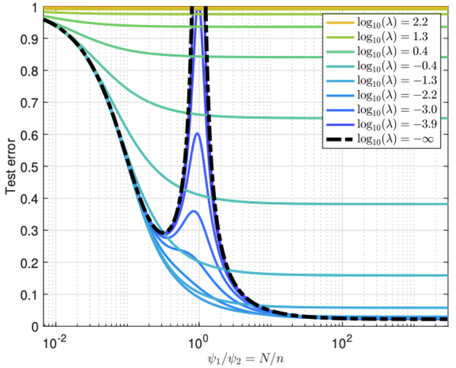
- **Results**
  - Surveys generalization bounds for least-squares linear regression in classical regime of roughly $O(d/n)$ excess error. All fall short in various ways: $\log n$ terms in numerator, strong assumptions about tail behavior of features/noise, dependence on condition number of random matrix that is not know a priori, boundedness of true function.
  - Introduces a PAC-Bayes algorithm that obtains a superior risk bound to any of the least-squares regression bounds without requiring sharp tails and well-conditioned inputs and with no log factor.
    - Algorithm updates a distribution over solutions by choosing solutions with higher probability if they have smaller training error and sampling from distribution.
    - Exponentially-weighted distribution introduces exponential tails for proof, even if they are not assumed to exist.
- **Implications**
  - Demonstrates comprehensiveness of bounds of least-squares regression in classical setting.

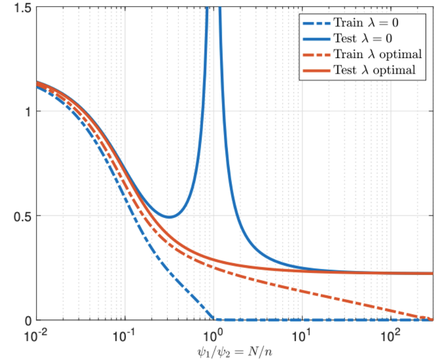
Appendix 3E: Random Feature Regression

The generalization error of random features regression: Precise asymptotics and double descent curve. Mei and Montanari (2019).

- **Results**
  - Considers linear combinations of N random features $\sigma(\langle x, w \rangle)$ for random d-dimensional w. Asymptotic regime where $n, N, d \to \infty$ at fixed ratios. Learn top-level coefficients with ridge regression with parameter λ. Random features and samples are drawn uniformly from a sphere and true function is typically linear.
  - Theoretical results show convergence to fixed "true" bias and variance terms $\mathscr{B}$ and $\mathscr{V}$ that are absurdly complex; similar in flavor to HMRT19 results. Some takeaways:
    - For small λ, peaking still occurs when $n = N$. Optimal choices of λ for each N/n ratio mitigates double-descent and makes error monotonically decrease.
    - For fixed λ, optimal generalization occurs in over-parameterized regime $N \gg n$.
  - Methods draw from random matrix theory and "log-determinants."
- **Implications**
  - Double-descent occurs in a model that is more similar to neural networks without requiring funny data distributions.



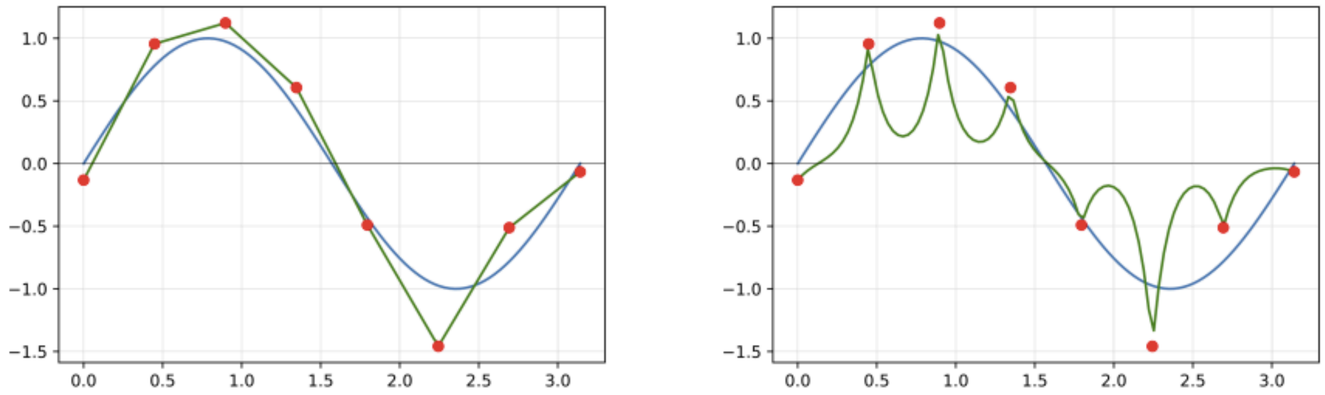




Appendix 3F: Kernel Regression

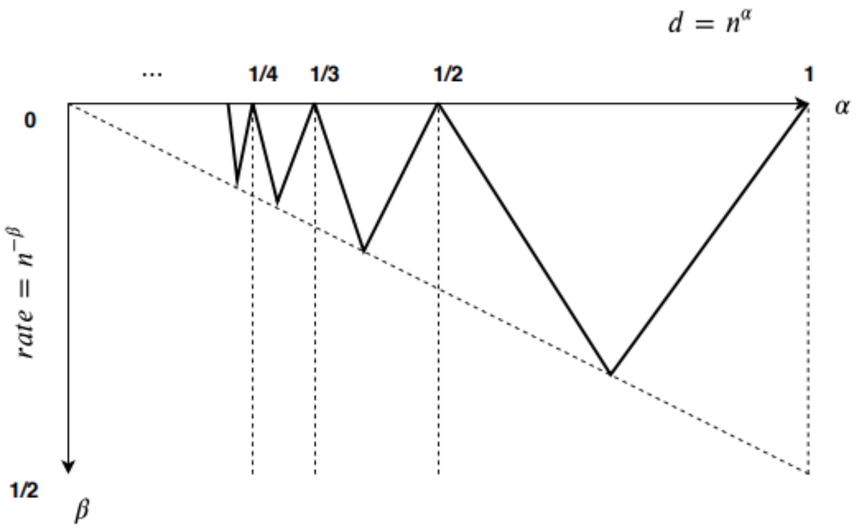
Consistency of Interpolation with Laplace Kernels is a High-Dimensional Phenomenon. Rakhlin and Zhai (2019).

- **Results**
  - Laplacian kernel interpolation with fixed kernel radius has constant lower bound on error in low dimensions.
    - If kernel radius is large, then there are balls around each sample where the classifier has roughly the same output and is hence influenced by the noise of the sample. In low-dimensions, these balls have non-negligible size.
    - If kernel radius is small, then interpolating solution will have much smaller than norm than true solution and cannot be a good approximation.
- **Implications**
  - Some interpolation methods cannot work without high-dimensions, regardless of number of samples.



On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels. Liang, Rakhlin, Zhai (2019).

- **Results**
  - Gives upper bounds that suggest multiple-descent can occur in the under-parameterized regime. Remains to show a matching lower bound to show it must occur.
    - Each multiple descent depends on a term that roughly captures BLLT19 variance.
    - To have a peak at $d = n^{1/\iota}$, the kernel functions must have a non-zero coefficient α_ι. When $d = n^{1/\iota}$, the degree-ι approximation of the kernel matrix may be ill-conditioned, causing a peak.

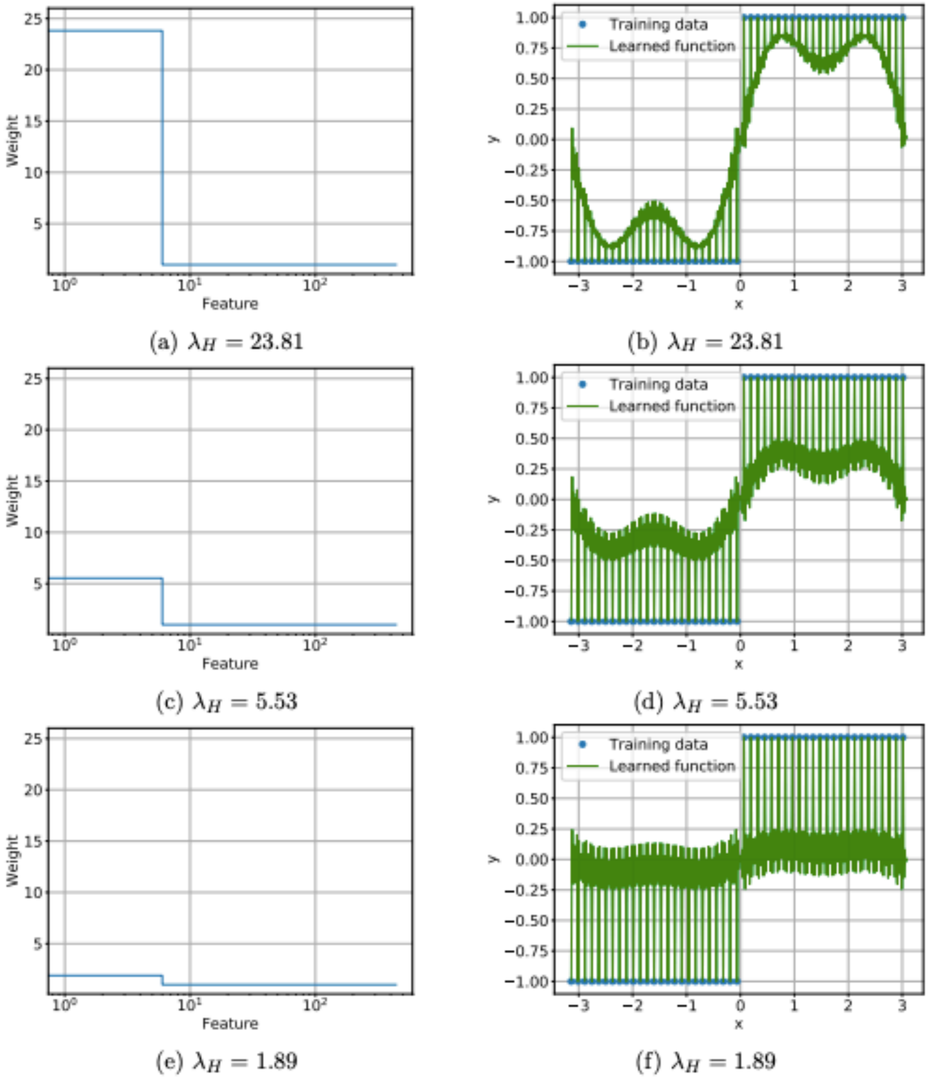


- **Implications**
  - Multiple descent can occur in kernel regression when kernel is the right degree polynomial.

Appendix 4: Support Vector Machines

Classification vs regression in overparameterized regimes: Does the loss function matter? Muthukumar, Narang, Subramanian, Belkin, Hsu, Sahai (2020).

- **Results**
  - Proves that support vector proliferation (or OLS=SVM) occurs for very high-dimensional SVMs.
  - Relates binary- and real-valued OLS benign overfitting by considering simple bi-level model with 1-sparse signal and no noise via survival and contamination analysis (like MVSS19).
    - Like BLLT19, benign overfitting occurs when bi-level model does not have too slow a drop-off between high-importance and low-importance features.
    - Classification generalization requires low contamination; regression requires low contamination *and* high survival.
- **Implications**
  - Benign overfitting can happen for SVMs where every sample is a support vector by transferring results from minimum-norm interpolation.
  - Benign overfitting is "easier" in classification than regression.
  - SVM generalization bounds can be proved outside of the “classical” setting where only a few training samples can be support vectors.



(a) $\lambda_H = 23.81$ (b) $\lambda_H = 23.81$

(c) $\lambda_H = 5.53$ (d) $\lambda_H = 5.53$

(e) $\lambda_H = 1.89$ (f) $\lambda_H = 1.89$

Finite-sample analysis of interpolating linear classifiers in the overparameterized regime. Chatterji, Long (2020).

- **Results**
  - By Soudry, et. al. (2018), gradient descent on separable data with logistic loss converges to maximum-margin classifier (or hard-margin SVM).
  - For over-parameterized $d = \omega(n^2)$ data drawn from two clusters with noisy labels, benign overfitting for classification occurs.
    - Direct proof by showing that angle between true separator and learned separator is small. Follows from convergence of weights of gradient descent on logistic loss and fact that noisy samples are shown to not have out-size influence on optimization process.
    - One instance is the *Boolean noisy rare-weak model*, where small fraction of features weakly indicate the cluster and all others give no information.
- **Implications**
  - SVM benign overfitting can occur in high-dimensional case in more general setting than MNSBHS20 and with a different proof technique.

Appendix 5: Limitations of Capacity-Based Bounds

Failures of model-dependent generalization bounds for least-norm interpolation. Bartlett and Long (2020).

- **Results**
  - Considers all valid model-dependent generalization bounds $\epsilon(h, n)$, which depend only on the hypothesis and the number of training samples that bound the excess risk with probability 0.9 for all data distributions P.
  - For any bound ϵ and sample size n, shows that there exists a probability distribution P_n over high-dimensional features where the least-norm interpolant h has excess error $O(1/\sqrt{n})$, but $\epsilon(h, n) \geq 1/2$.
  - Proof relies on supplying two similar distributions, one where minimum-norm interpolation does very well and one where it does poorly. The two are constructed so that MNI returns the same hypothesis in each. Thus, the generalization bound must be very loose on the “good" distribution.
- **Implications**
  - The success of least-norm interpolation can only be predicted given knowledge of the data distribution; seeing the solution and the sample size alone are insufficient to distinguish it from poor solutions. Establishes limitations of generalization approaches based on model capacity and smoothness of solutions.

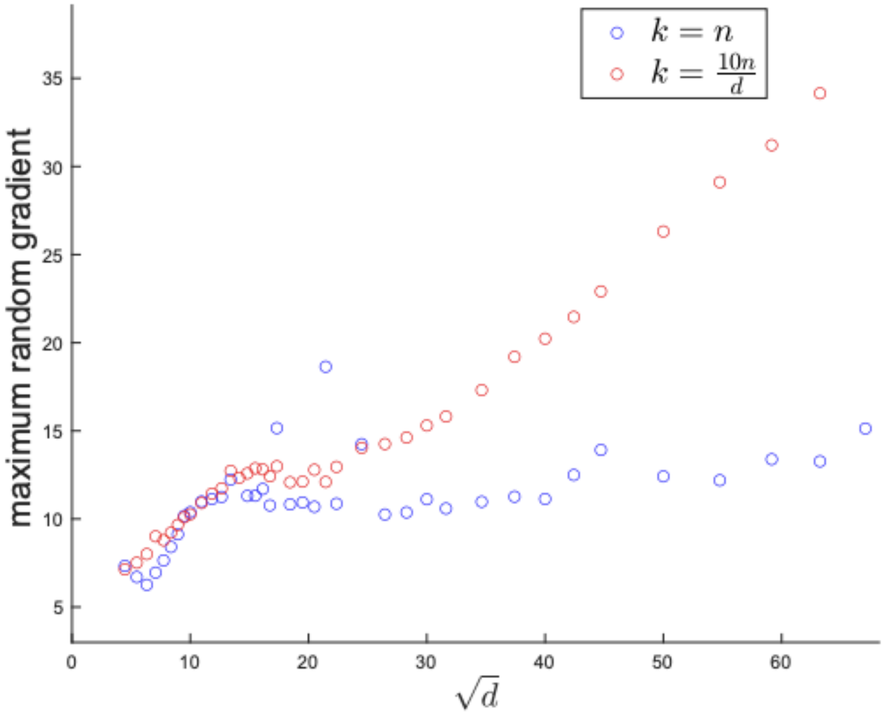
Appendix 6: Properties of Over-parameterized Neural Networks

A deep conditioning treatment of neural networks. Agarwal, Awasthi, Kale (2020).

- **Results**
  - Passing arbitrary inputs with bounded inner products through many randomly-initialized layers produces features that are nearly orthogonal.
  - Suggests bounds on fast training and hardness results for SQ learning.
  - Applies BLLT19 results to second-last layer to give generalization bounds on treating last layer of weights as minimum-norm interpolation.
    - However, bounds are vacuous for most settings because impossible to make many assumptions on data distribution of last layer; some distributions with large BLLT19 bias bounds may be large because over-parameterized models with too slow decay can produce orthogonal features.
- **Implications**
  - Conceivable that depth of neural networks influences high-depth features to be drawn from distribution with favorable conditions for benign overfitting. Possible area of future research?

A law of robustness for two-layer neural networks. Bubeck, Li, Nagaraj (2020).

- **Results**
  - Conjectures that a 2-layer neural nets of width k and Lipschitz constant $O(\sqrt{n/k})$ can perfectly fit n training samples, and that all interpolating NNs have Lipschitz constant $\Omega(\sqrt{n/k})$.
  - Proves several weaker statements:
    - If $d \gg n, k = 1$ neurons can fit the data with Lipschitz constant $O(\sqrt{n})$.
    - If $k = n$, there exists an $O(1)$-Lipschitz network fitting the data.
    - Can fit with $O(n/k)$-Lipschitz network.
    - For very small dimensions, near fit, and polynomial activation, can fit with $O(\sqrt{n/k})$-Lipschitz by tensor interpolation argument.
- **Implications**
  - If true, would suggest favorable robustness and (possibly) generalization properties of very over-parameterized neural networks.
  - Would also suggest that robustness is only possible if a neural network is over-parameterized.



A universal law of robustness via isoperimetry. Bubeck and Sellke (2021).

- **Results**
  - Proves a stronger version of the negative part of the BLN20 conjecture: with high probability over data, all neural networks approximately interpolating n samples with width k and constant depth and bounded weights have a Lipschitz constant of $\Omega(\sqrt{n/k \log(nk)})$.
  - Proof follows by isoperimetry argument.
    - A fixed L-Lipschitz function will not fit n random samples with high probability.
    - For small L, an ϵ-net argument can show that no L-Lipschitz function can fit n random samples WHP.
- **Implications**
  - Neural networks can only be robust (i.e. have a small Lipschitz constant) if they are highly over-parameterized.