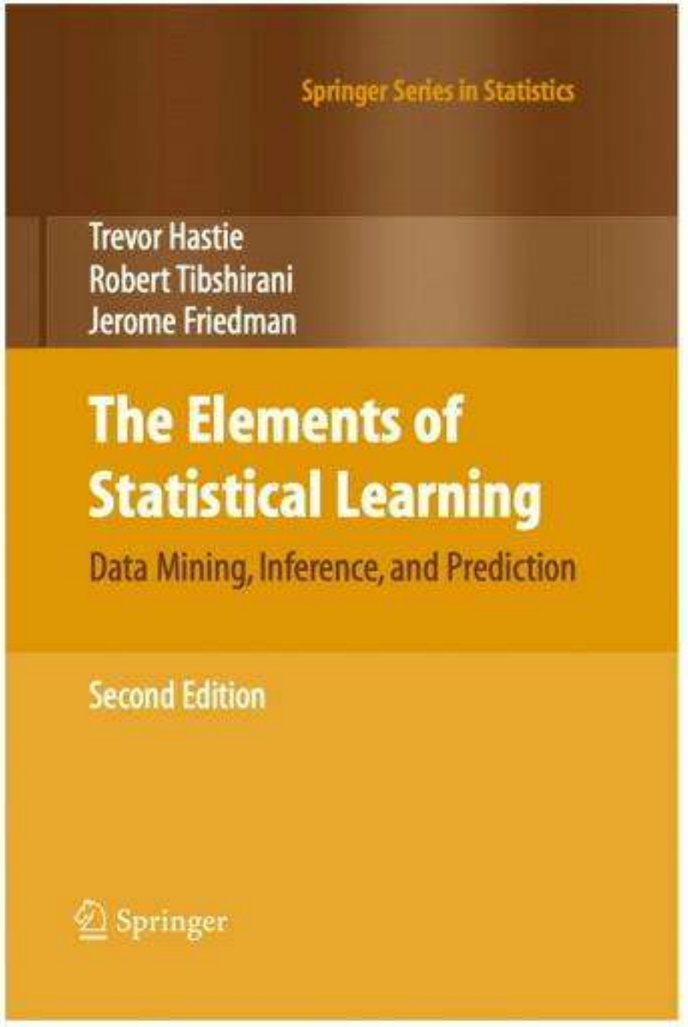
Spoilers

"A model with zero training error is overfit to the training data and will typically generalize poorly."

– Hastie, Tibshirani, & Friedman, *The Elements of Statistical Learning*

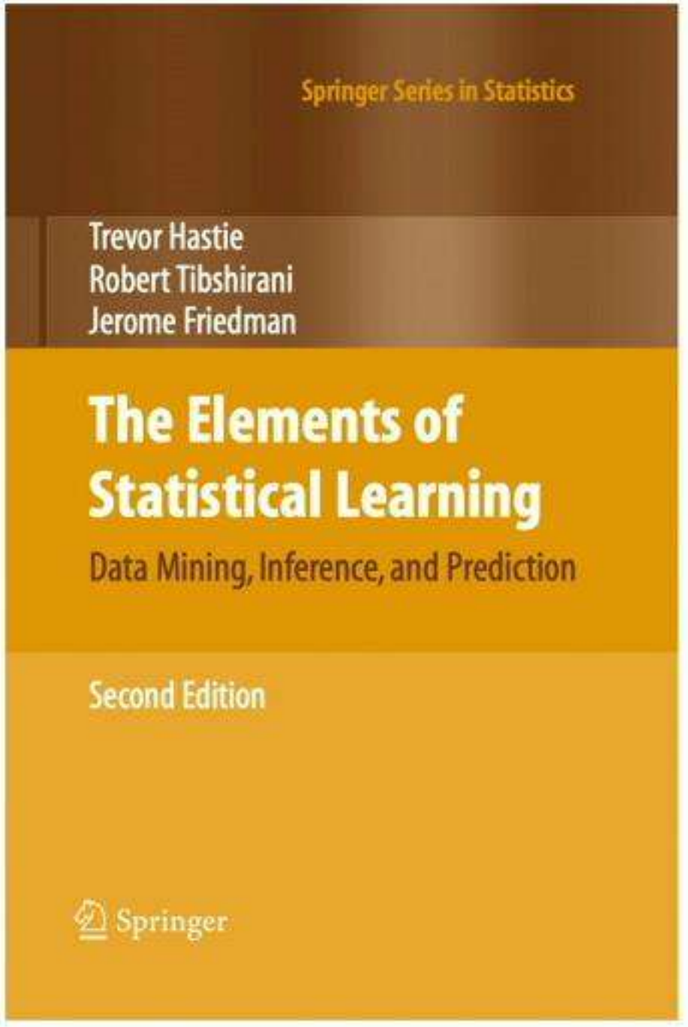


Spoilers

"A model with zero training error is overfit to the training data and will typically generalize poorly."

– Hastie, Tibshirani, & Friedman,
The Elements of Statistical Learning

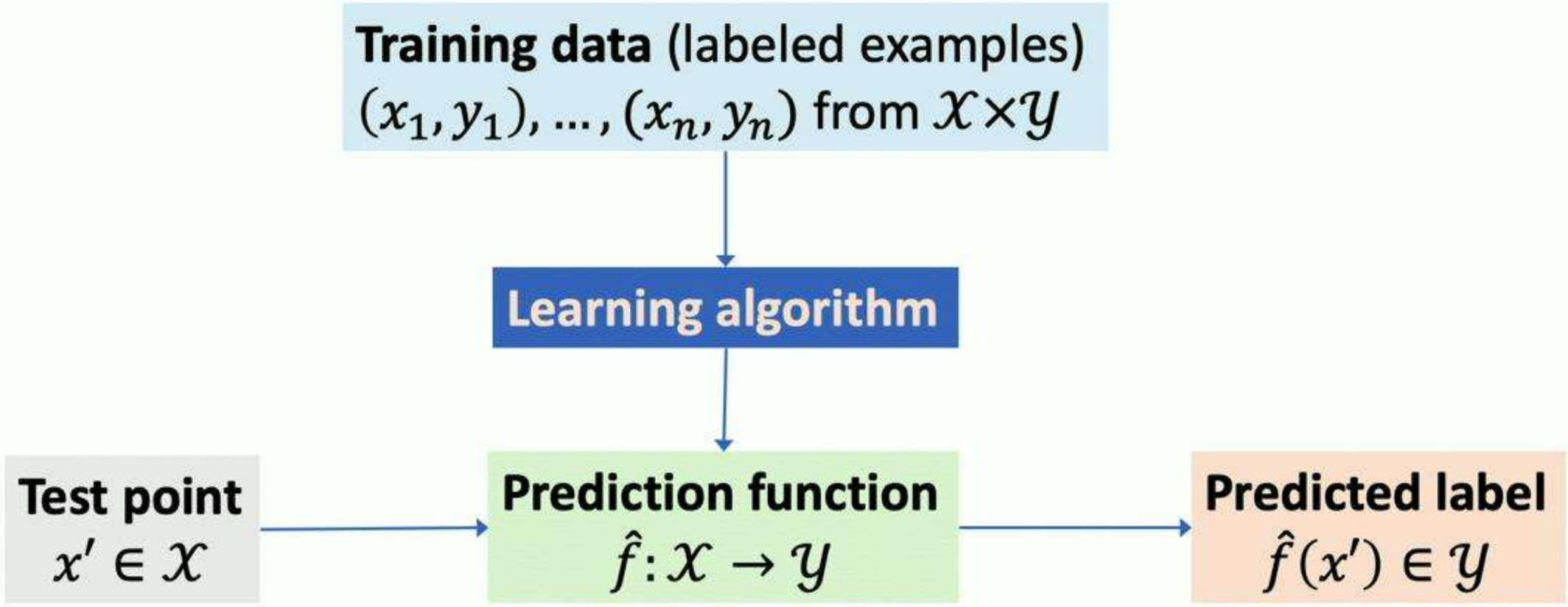
We'll give empirical and theoretical evidence against this conventional wisdom, at least in "modern" settings of machine learning.



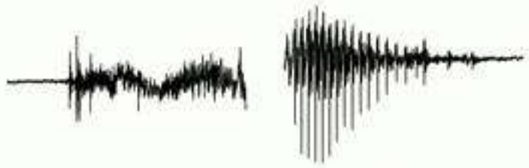
Outline

1. Observations that counter the conventional wisdom
2. Interpolation via local prediction
3. Interpolation via neural nets

Supervised learning



Supervised learning



/k/ /a/ ...

Training data (labeled examples)
$(x_1, y_1), \ldots, (x_n, y_n)$ from $\mathcal{X} \times \mathcal{Y}$

(IID from P)

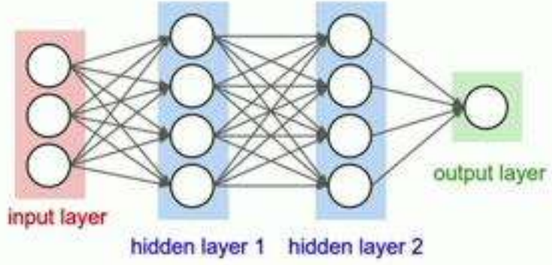
$w \leftarrow w - \eta \nabla \hat{\mathcal{R}}(w)$

Learning algorithm

Risk: $\mathcal{R}(f) := \mathbb{E}[\ell(f(x'), y')]$
where $(x', y') \sim P$

Test point
$x' \in \mathcal{X}$

Prediction function
$\hat{f}: \mathcal{X} \to \mathcal{Y}$



Predicted label
$\hat{f}(x') \in \mathcal{Y}$

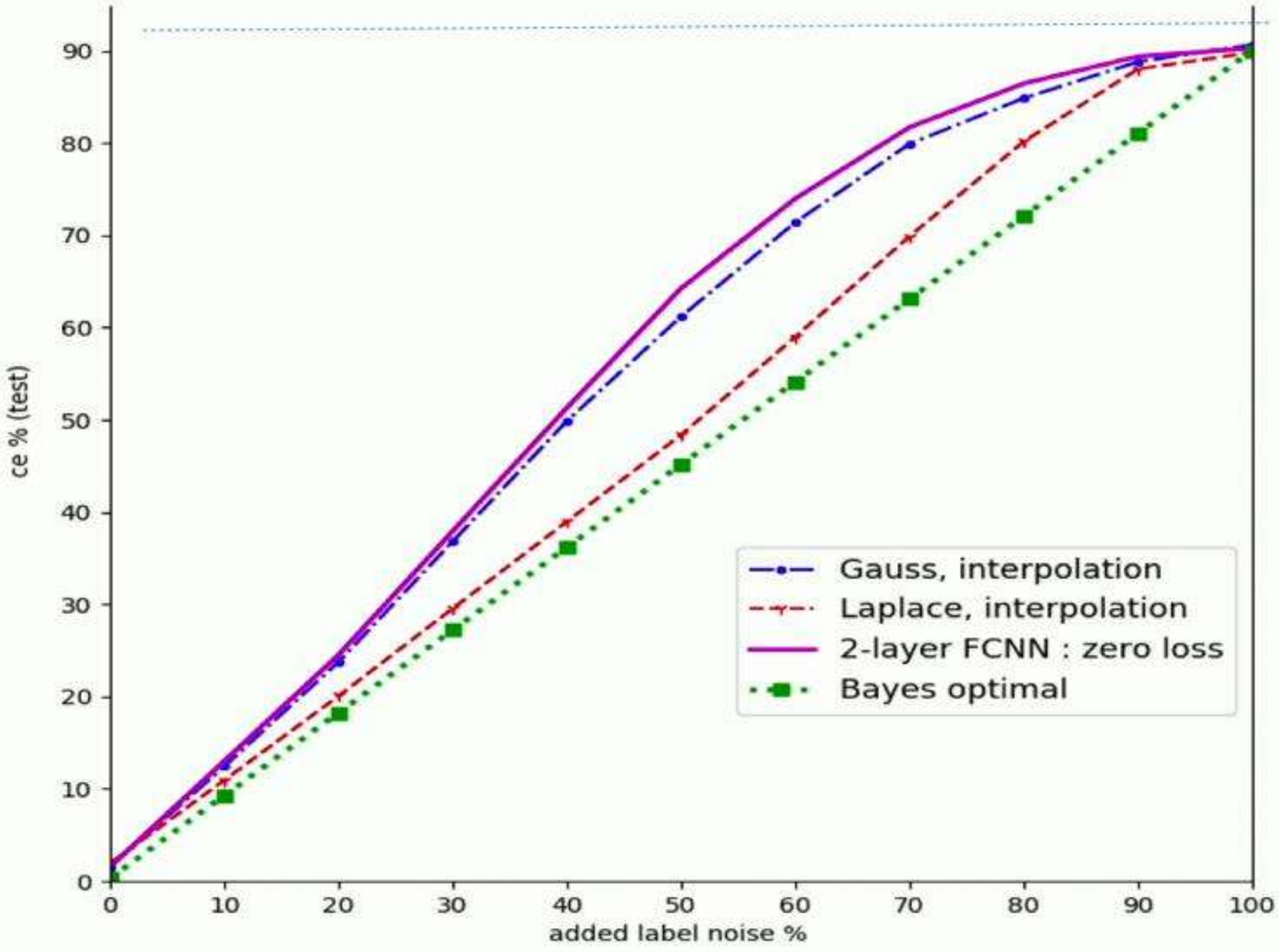
/t/

Observations from the field

(Zhang, Bengio, Hardt, Recht, & Vinyals, 2017; Belkin, Ma, & Mandal, 2018)

Neural nets & kernel machines:

- Can interpolate any training data.
- Can generalize even when training data has substantial amount of label noise.

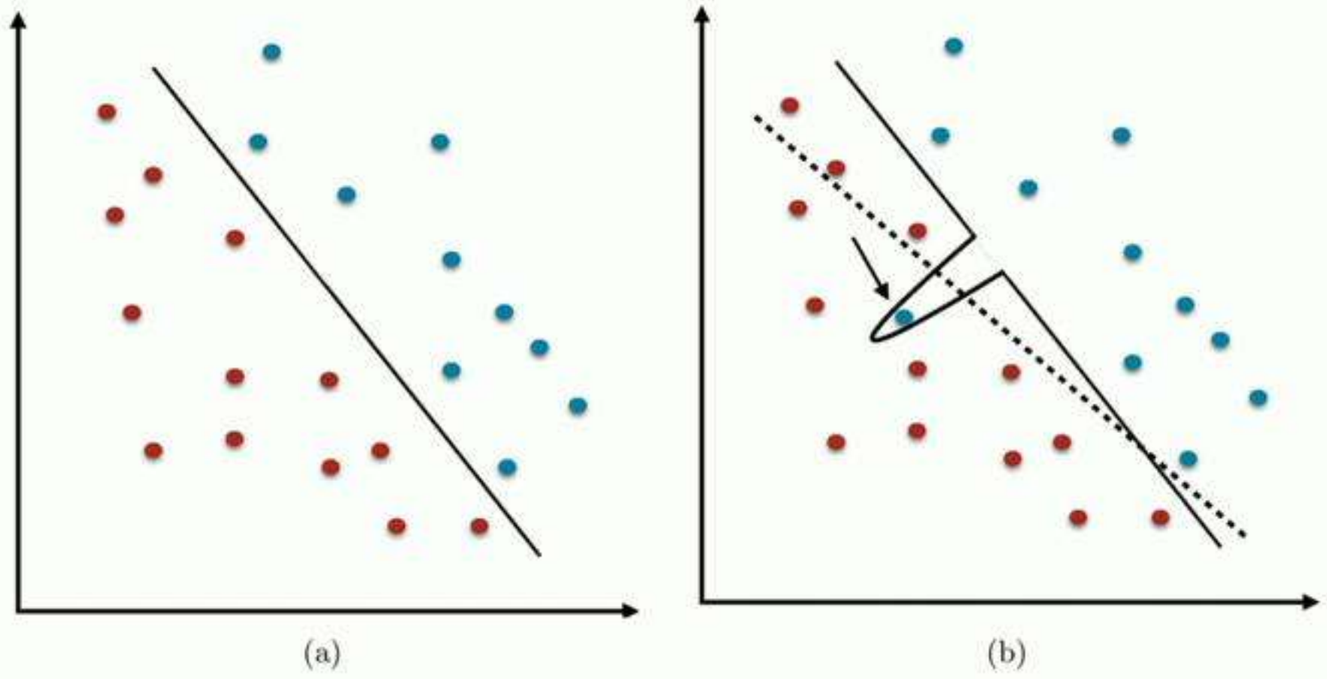


More observations from the field

(Wyner, Olson, Bleich, & Mease, 2017)

AdaBoost + large decision trees / Random forests:

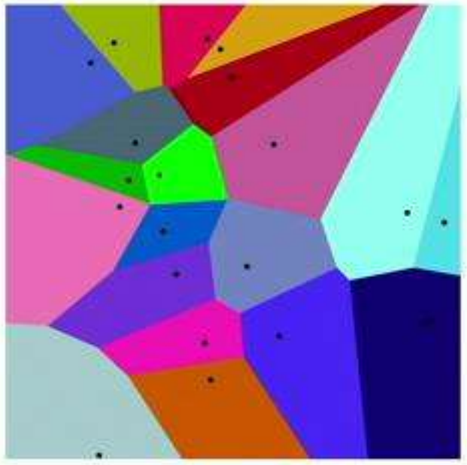
- Interpret as local interpolation methods
- Flexibility -> robustness to label noise



Existing theory about local interpolation

Nearest neighbor (Cover & Hart, 1967)

- Predict with label of nearest training example
- Interpolates training data
- Risk $\to 2 \cdot \mathcal{R}(g^*)$ (sort of)



Hilbert kernel (Devroye, Györfi, & Krzyżak, 1998)

- Special kind of smoothing kernel regression (like Shepard's method)
- Interpolates training data
- Consistent*, but no convergence rates

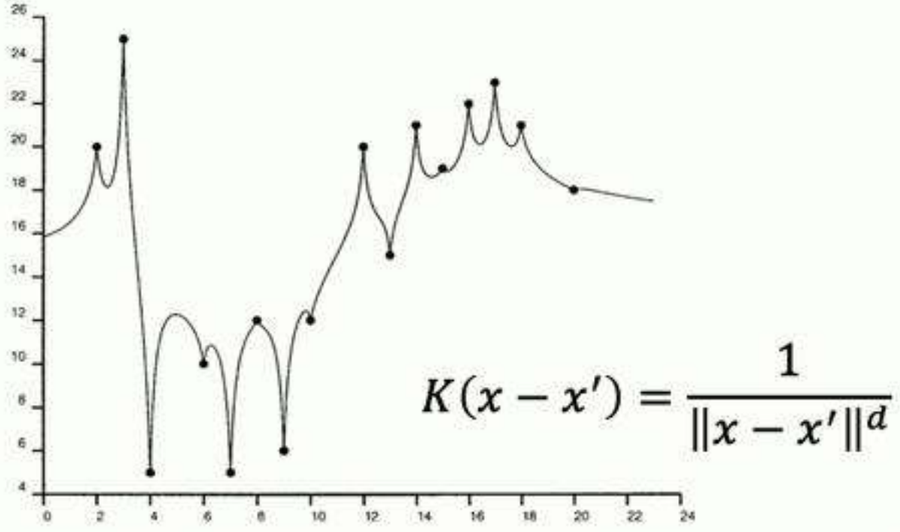


FIG. 2. The Hilbert kernel regression estimate with $a = 1$.

Our goals

- **Further counter the "conventional wisdom" re: interpolation**
  Show interpolation methods can be consistent (or almost consistent) for classification & regression
  - Simplicial interpolation
  - Weighted & interpolated nearest neighbor
  - Neural nets / linear models

Our goals

- **Further counter the "conventional wisdom" re: interpolation**
  Show interpolation methods can be consistent (or almost consistent) for classification & regression
  - Simplicial interpolation
  - Weighted & interpolated nearest neighbor
  - Neural nets / linear models
- Identify some **useful properties of good interpolation** methods
- Suggest **connections to practical methods**

Interpolation via local prediction

Preliminaries

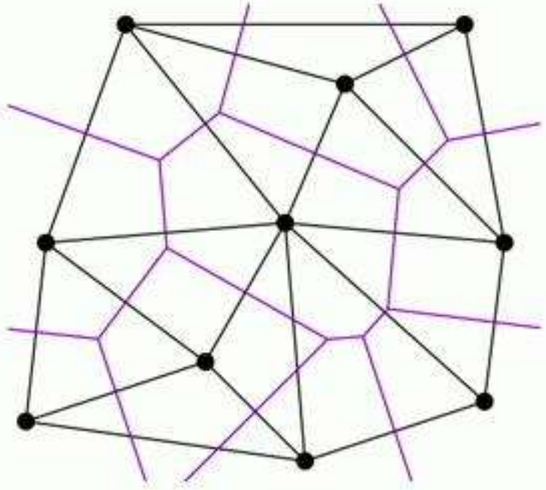
- Construct estimate $\hat{\eta}$ of the **regression function**

$$\eta(x) = \mathbb{E}[\, y' \mid x' = x \,]$$

- For **binary classification** $\mathcal{Y} = \{0,1\}$:
  - $\eta(x) = \Pr(y' = 1 \mid x' = x)$
  - **Optimal classifier**: $f^*(x) = \mathbb{I}_{\eta(x) > \frac{1}{2}}$
  - **Plug-in classifier**: $\hat{f}(x) = \mathbb{I}_{\hat{\eta}(x) > \frac{1}{2}}$ based on estimate $\hat{\eta}$
- **Questions**:
  What is the risk as $n \to \infty$? At what rate does it converge?

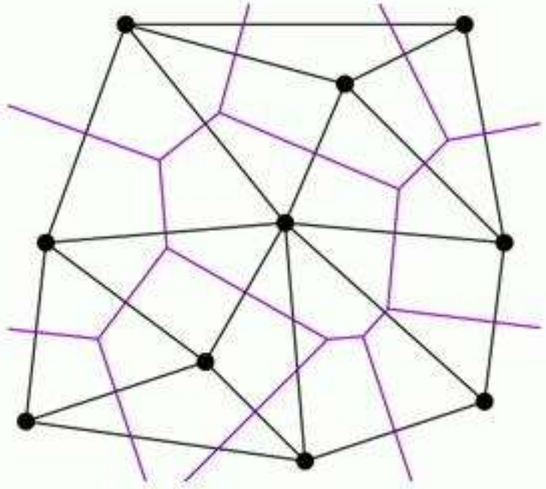
I. Simplicial interpolation

- IID training examples $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^d \times [0,1]$
  - Partition $C := \mathrm{conv}(x_1, \ldots, x_n)$ into simplices with x_i as vertices via Delaunay.
  - Define $\hat{\eta}(x)$ on each simplex by affine interpolation of vertices' labels.
  - Result is piecewise linear on C. (Punt on what happens outside of C.)



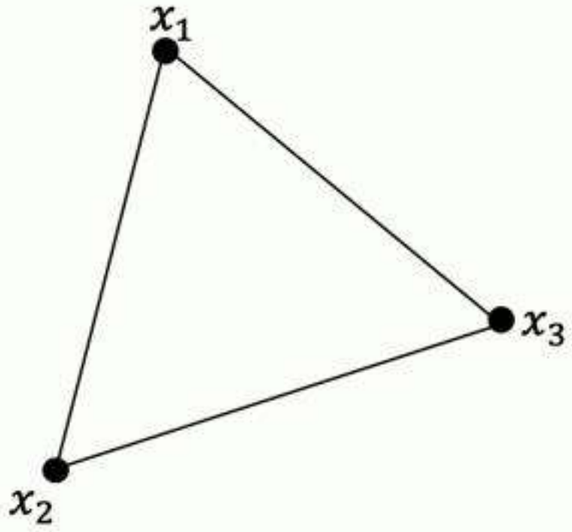
I. Simplicial interpolation

- IID training examples $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^d \times [0,1]$
  - Partition $C := \mathrm{conv}(x_1, \ldots, x_n)$ into simplices with x_i as vertices via Delaunay.
  - Define $\hat{\eta}(x)$ on each simplex by affine interpolation of vertices' labels.
  - Result is piecewise linear on C. (Punt on what happens outside of C.)
- For classification ($y \in \{0,1\}$), $\hat{f}$ is plug-in classifier based on $\hat{\eta}$.



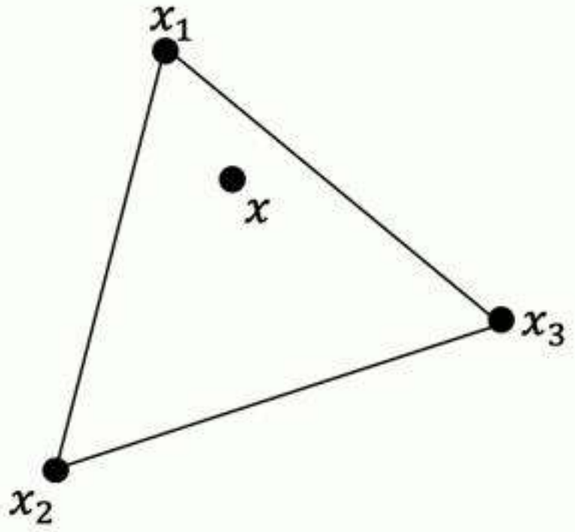
What happens on a single simplex

- Simplex on $x_1, \ldots, x_{d+1}$ with corresponding labels $y_1, \ldots, y_{d+1}$



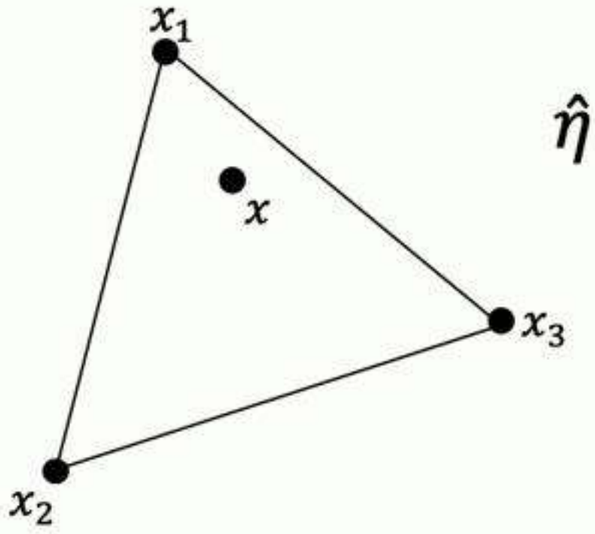
What happens on a single simplex

- Simplex on $x_1, \ldots, x_{d+1}$ with corresponding labels $y_1, \ldots, y_{d+1}$
- Test point x in simplex, with barycentric coordinates $(w_1, \ldots, w_{d+1})$.



What happens on a single simplex

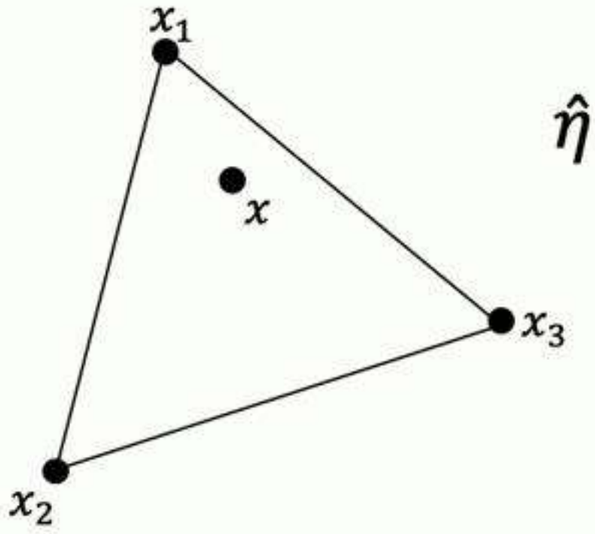
- Simplex on $x_1, \ldots, x_{d+1}$ with corresponding labels $y_1, \ldots, y_{d+1}$
- Test point x in simplex, with barycentric coordinates $(w_1, \ldots, w_{d+1})$.
- Linear interpolation at x (i.e., least squares fit, evaluated at x):



$$\hat{\eta}(x) = \sum_{i=1}^{d+1} w_i y_i$$

What happens on a single simplex

- Simplex on $x_1, \ldots, x_{d+1}$ with corresponding labels $y_1, \ldots, y_{d+1}$
- Test point x in simplex, with barycentric coordinates $(w_1, \ldots, w_{d+1})$.
- Linear interpolation at x (i.e., least squares fit, evaluated at x):



$$\hat{\eta}(x) = \sum_{i=1}^{d+1} w_i y_i$$

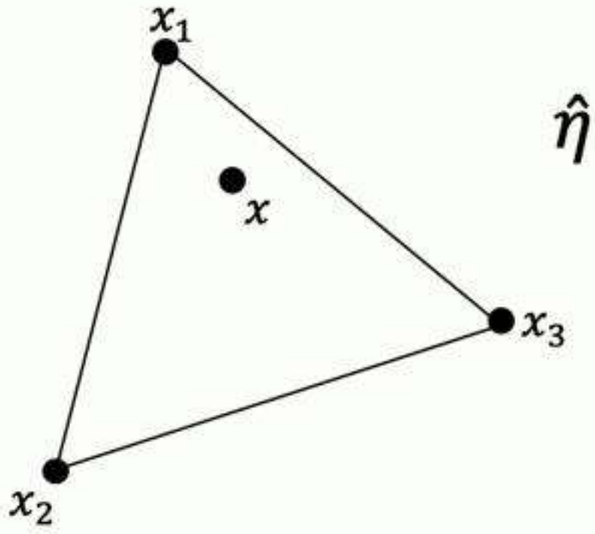
Key idea: aggregates information from all vertices to make prediction.
(C.f. nearest neighbor rule.)

Comparison to nearest neighbor rule

- Suppose $\eta(x) = \Pr(y = 1 \mid x) < 1/2$ for all points in a simplex
  - Optimal prediction of f^* is 0 for all points in simplex.

What happens on a single simplex

- Simplex on $x_1, \ldots, x_{d+1}$ with corresponding labels $y_1, \ldots, y_{d+1}$
- Test point x in simplex, with barycentric coordinates $(w_1, \ldots, w_{d+1})$.
- Linear interpolation at x (i.e., least squares fit, evaluated at x):



$$\hat{\eta}(x) = \sum_{i=1}^{d+1} w_i y_i$$

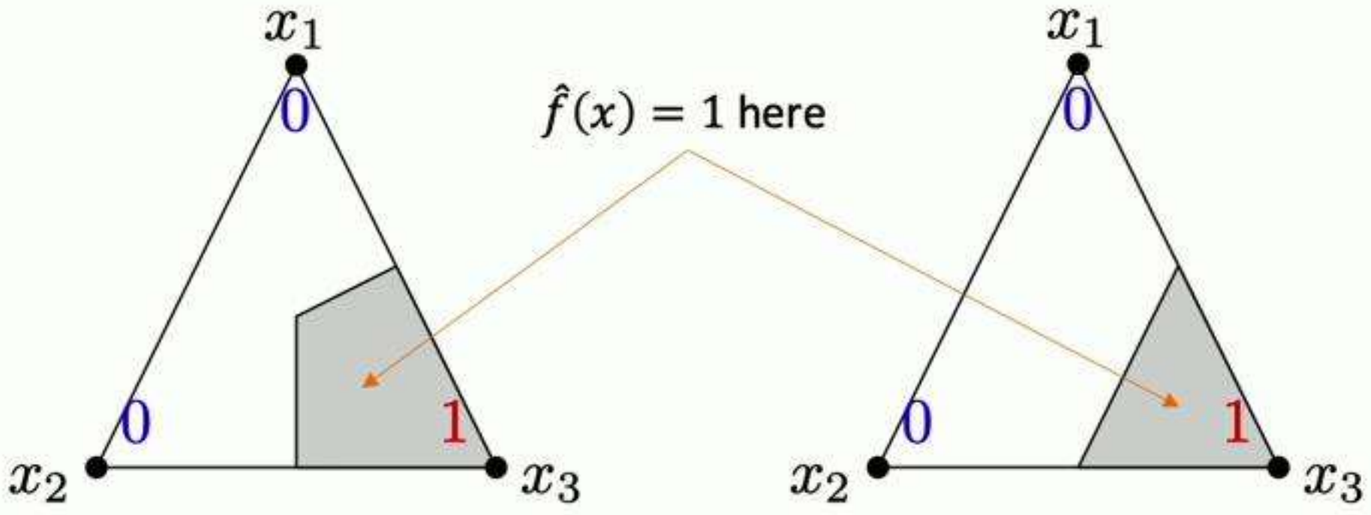
Key idea: aggregates information from all vertices to make prediction.
(C.f. nearest neighbor rule.)

Comparison to nearest neighbor rule

- Suppose $\eta(x) = \Pr(y = 1 \mid x) < 1/2$ for all points in a simplex
  - Optimal prediction of f^* is 0 for all points in simplex.

Comparison to nearest neighbor rule

- Suppose $\eta(x) = \Pr(y = 1 \mid x) < 1/2$ for all points in a simplex
  - Optimal prediction of f^* is 0 for all points in simplex.
- Suppose $y_1 = \cdots = y_d = 0$, but $y_{d+1} = 1$ (due to "label noise")

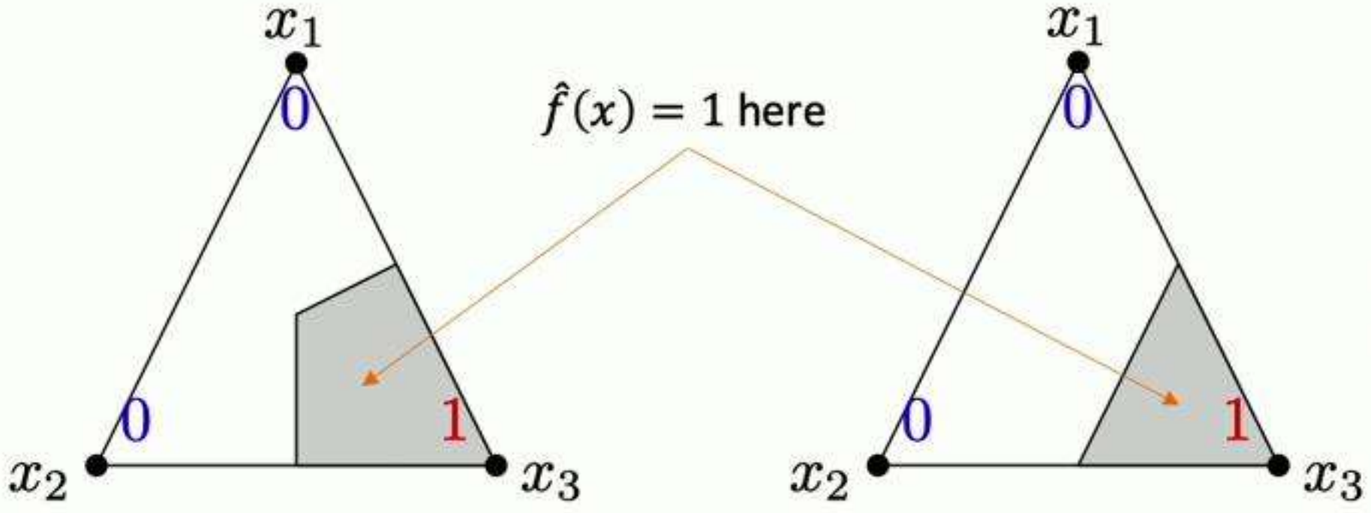


Nearest neighbor rule

Simplicial interpolation

Comparison to nearest neighbor rule

- Suppose $\eta(x) = \Pr(y = 1 \mid x) < 1/2$ for all points in a simplex
  - Optimal prediction of f^* is 0 for all points in simplex.
- Suppose $y_1 = \cdots = y_d = 0$, but $y_{d+1} = 1$ (due to "label noise")



Nearest neighbor rule

Simplicial interpolation

Effect is exponentially more pronounced in high dimensions!

Asymptotic risk for simplicial interpolation

[Belkin, H., Mitra, '18]

Theorem (classification): Assume distribution of x' is uniform on some convex set, and η is bounded away from $1/2$. Then simplicial interpolation's plug-in classifier $\hat{f}$ satisfies

$$\limsup_n \mathbb{E}[\text{zero/one loss}] \leq \left(1 + e^{-\Omega(d)}\right) \cdot \text{OPT}$$

Asymptotic risk for simplicial interpolation

[Belkin, H., Mitra, '18]

Theorem (classification): Assume distribution of x' is uniform on some convex set, and η is bounded away from $1/2$. Then simplicial interpolation's plug-in classifier $\hat{f}$ satisfies

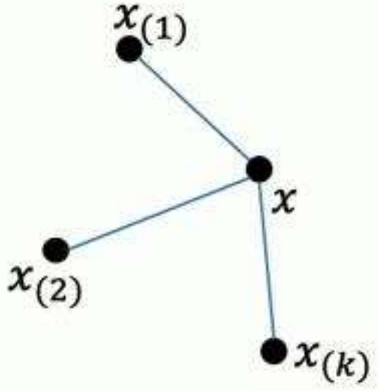
$$\limsup_n \mathbb{E}[\text{zero/one loss}] \leq \left(1 + e^{-\Omega(d)}\right) \cdot \text{OPT}$$

- **C.f. nearest neighbor classifier**: $\limsup_n \mathbb{E}\left[\mathcal{R}(\hat{f})\right] \approx 2 \cdot \mathcal{R}(f^*)$
- **For regression** (squared error):

$$\limsup_n \mathbb{E}[\text{squared error}] \leq \left(1 + O\left(\frac{1}{d}\right)\right) \cdot \text{OPT}$$

II. Weighted & interpolated NN scheme

- For given test point x, let $x_{(1)}, \ldots, x_{(k)}$ be k nearest neighbors in training data, and let $y_{(1)}, \ldots, y_{(k)}$ be corresponding labels.



Define

$$\hat{\eta}(x) = \frac{\sum_{i=1}^{k} w(x, x_{(i)})\, y_{(i)}}{\sum_{i=1}^{k} w(x, x_{(i)})}$$

where

$$w\left(x, x_{(i)}\right) = \left\| x - x_{(i)} \right\|^{-\delta}, \qquad \delta > 0$$

Interpolation: $\hat{\eta}(x) \to y_i$ as $x \to x_i$

Comparison to Hilbert kernel estimate

Weighted & interpolated NN

$$\hat{\eta}(x) = \frac{\sum_{i=1}^{k} w(x, x_{(i)})\, y_{(i)}}{\sum_{i=1}^{k} w(x, x_{(i)})}$$

$$w(x, x_{(i)}) = \|x - x_{(i)}\|^{-\delta}$$

Optimal non-parametric rates

Hilbert kernel (Devroye, Györfi, & Krzyżak, 1998)

$$\hat{\eta}(x) = \frac{\sum_{i=1}^{n} w(x, x_i)\, y_i}{\sum_{i=1}^{n} w(x, x_i)}$$

$$w(x, x_i) = \|x - x_i\|^{-\delta}$$

Consistent, but no non-asymptotic rates

Comparison to Hilbert kernel estimate

Weighted & interpolated NN

$$\hat{\eta}(x) = \frac{\sum_{i=1}^{k} w(x, x_{(i)})\, y_{(i)}}{\sum_{i=1}^{k} w(x, x_{(i)})}$$

$$w(x, x_{(i)}) = \|x - x_{(i)}\|^{-\delta}$$

Optimal non-parametric rates

Hilbert kernel (Devroye, Györfi, & Krzyżak, 1998)

$$\hat{\eta}(x) = \frac{\sum_{i=1}^{n} w(x, x_i)\, y_i}{\sum_{i=1}^{n} w(x, x_i)}$$

$$w(x, x_i) = \|x - x_i\|^{-\delta}$$

Consistent, but no non-asymptotic rates

Localization is essential to get non-asymptotic rate.

Convergence rates for WiNN

[Belkin, H., Mitra, '18]

Theorem: Assume distribution of x' is uniform on some compact set satisfying regularity condition, and η is α-Holder smooth.

For appropriate setting of k, weighted & interpolated NN estimate $\hat{\eta}$ satisfies

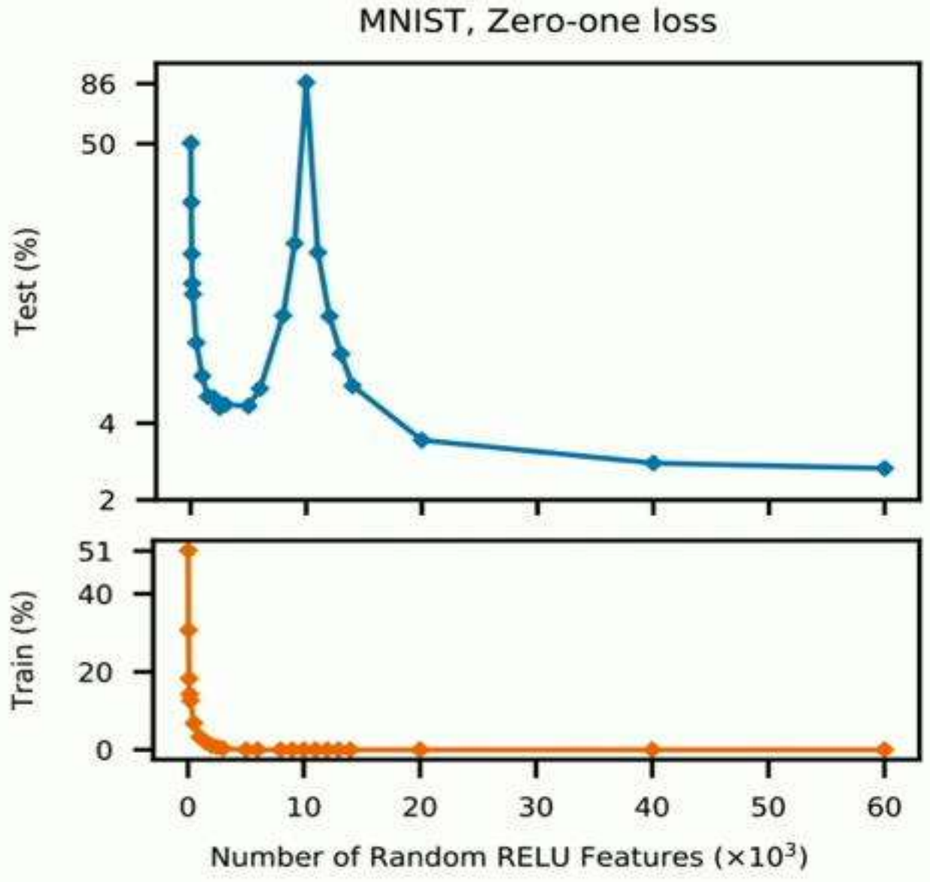
$$\mathbb{E}\left[\left(\hat{\eta}(X) - \eta(X)\right)^2\right] \leq O\left(n^{-2\alpha/(2\alpha+d)}\right)$$

- Consistency + optimal rates of convergence for interpolating method.
- Follow-up work by Belkin, Rakhlin, Tsybakov '19: also for Nadaraya-Watson with compact & singular kernel.

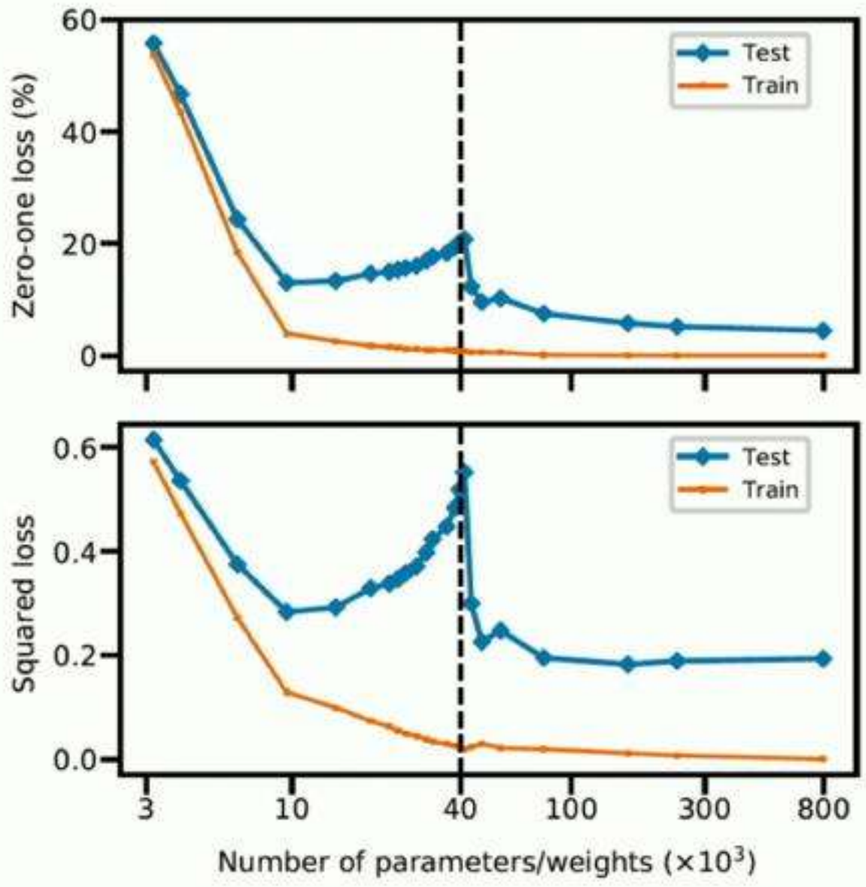
Interpolation via neural nets

Two layer fully-connected neural networks

[Belkin, H., Ma, Mandal, '19]



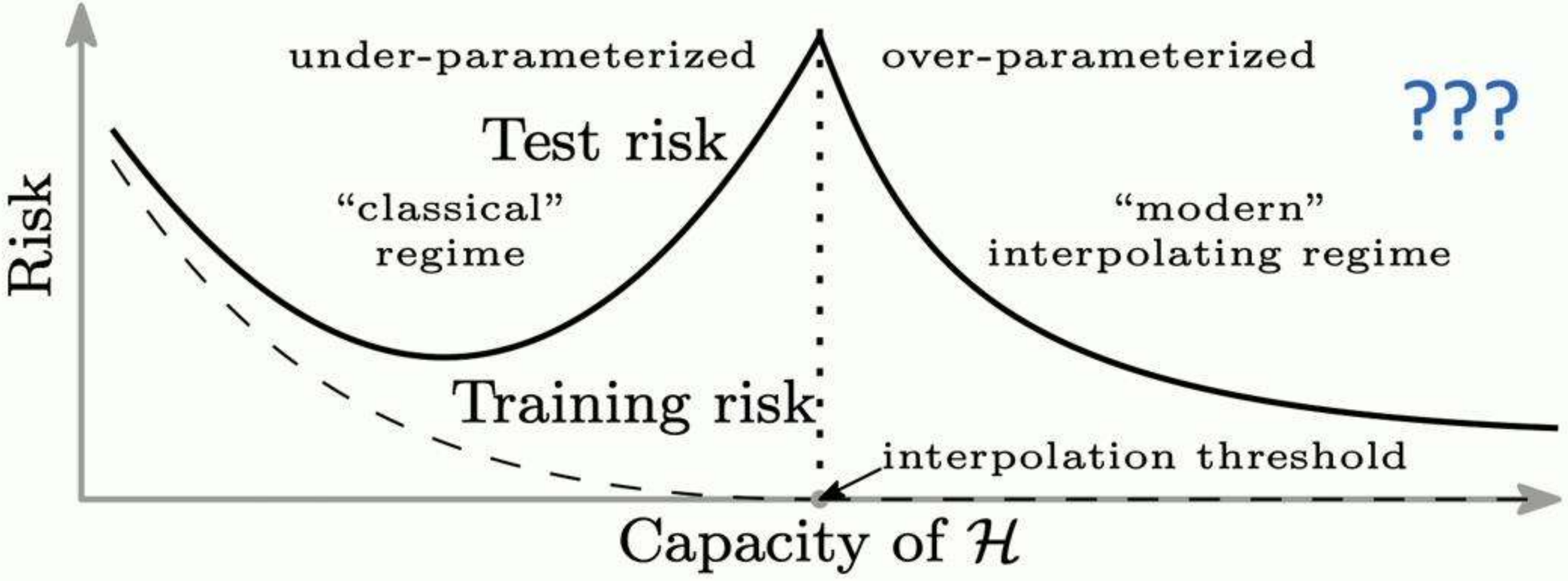
Random first layer



Trained first layer

"Double descent" risk curve

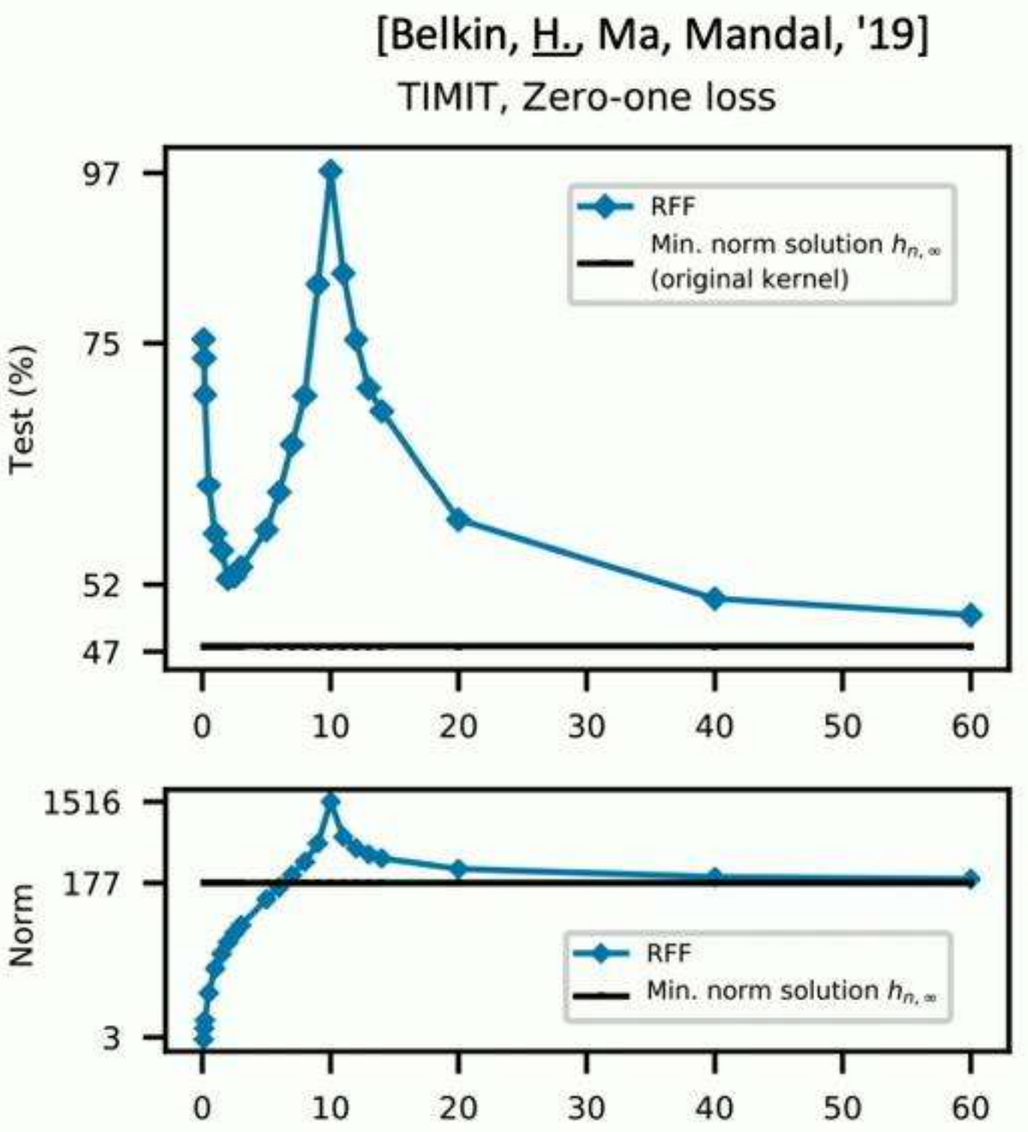
[Belkin, H., Ma, Mandal, '19]



Approximating a kernel machine?

- Effectiveness of interpolation depends on ability to align with the "right" inductive bias
- E.g., low RKHS norm

[Belkin, H., Ma, Mandal, '19]

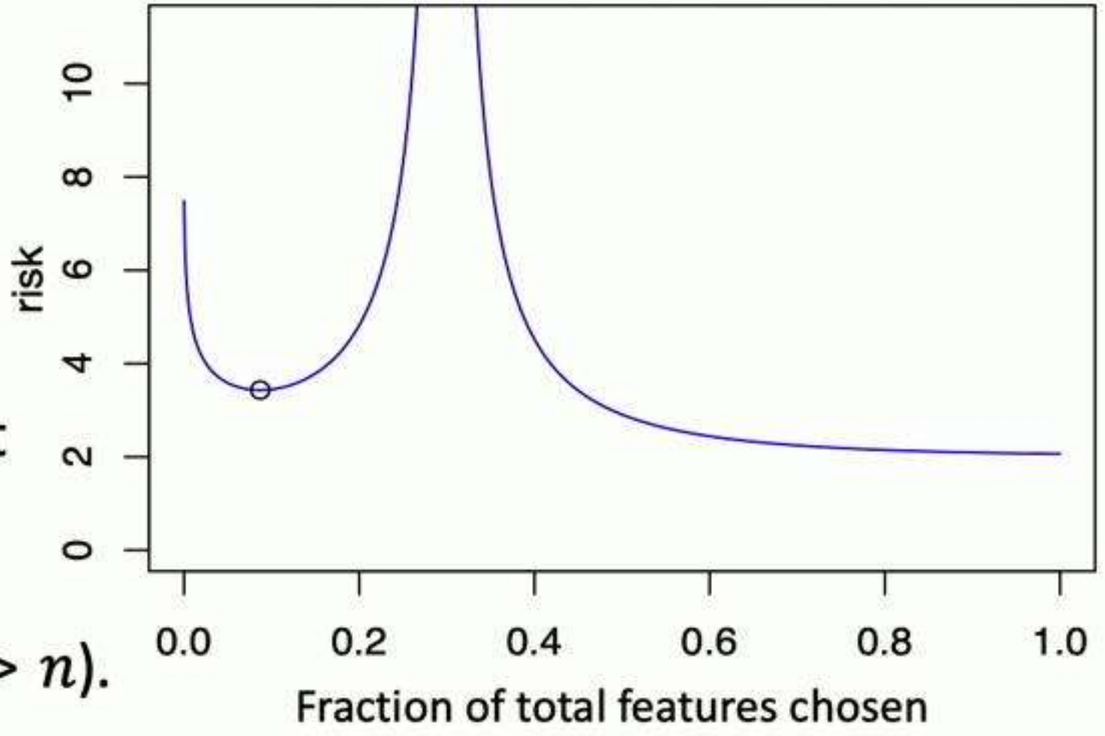


Linear regression w/ weak features [Belkin, H., Xu, '19+; Xu, H., '19+]

Gaussian design linear model with D features
All features are "relevant" but equally weak

Only use p of the features ($1 \leq p \leq D$)
Least squares ($p \leq n$) or least norm ($p \geq n$) fit

Theorem: Under certain eigenvalue condition, minimum is beyond point of interpolation ($p > n$).



Concurrent work by Hastie, Montanari, Rosset, Tibshirani '19 (also for some non-linear models!).

Other analyses of linear models: Muthukumar, Vodrahalli, Sahai, '19; Bartlett, Long, Lugosi, Tsigler, '19.

Conclusions/open problems

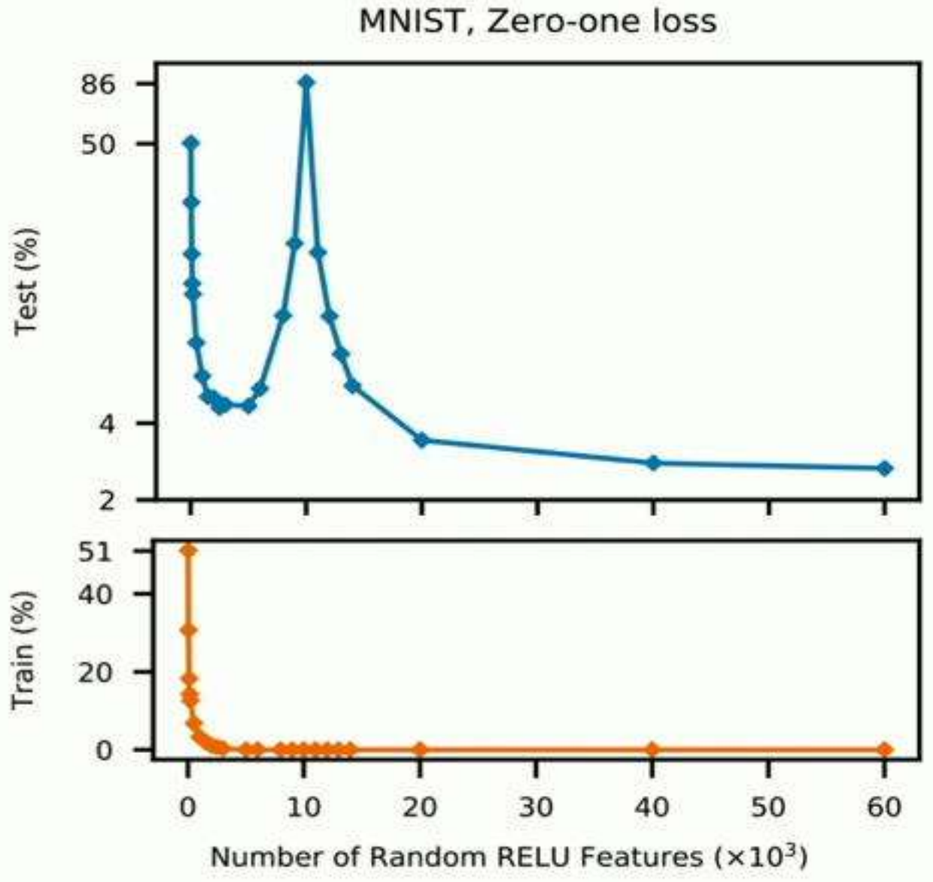
1. Interpolation is compatible with good statistical properties
2. They work by relying (exclusively!) on **inductive bias**: e.g.,
    1. Smoothness from local averaging in high-dimensions.
    2. Low function space norm

Open problems:

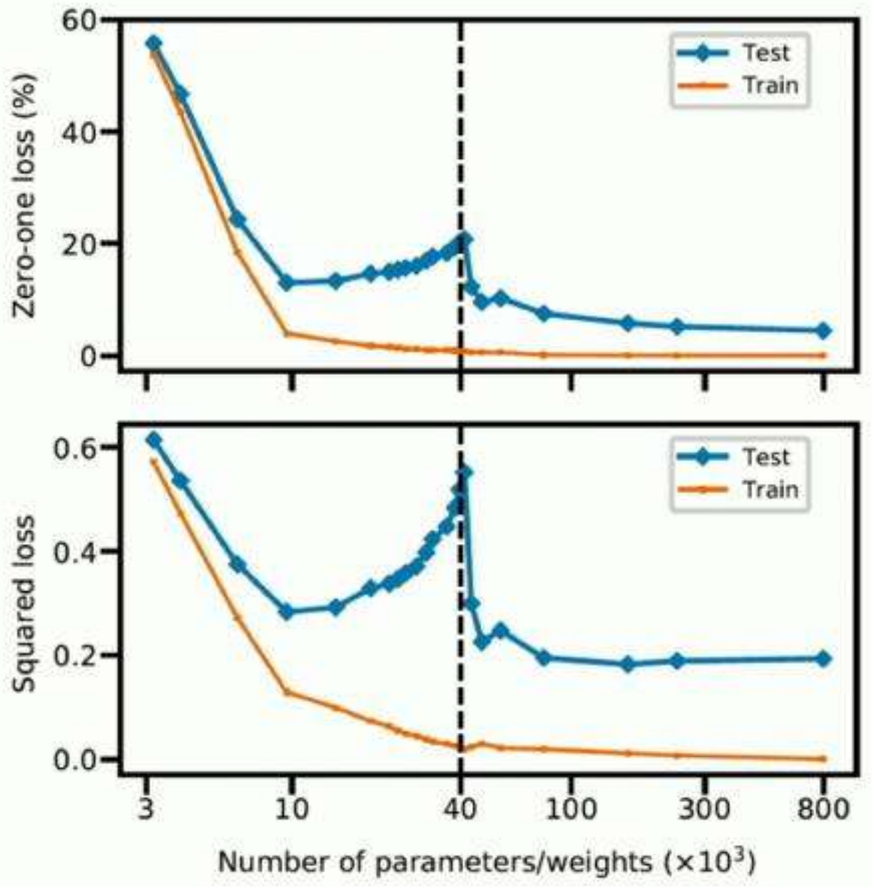
- Characterize inductive biases of other learning algorithms
- Benefits of interpolation?

Two layer fully-connected neural networks

[Belkin, H., Ma, Mandal, '19]



Random first layer



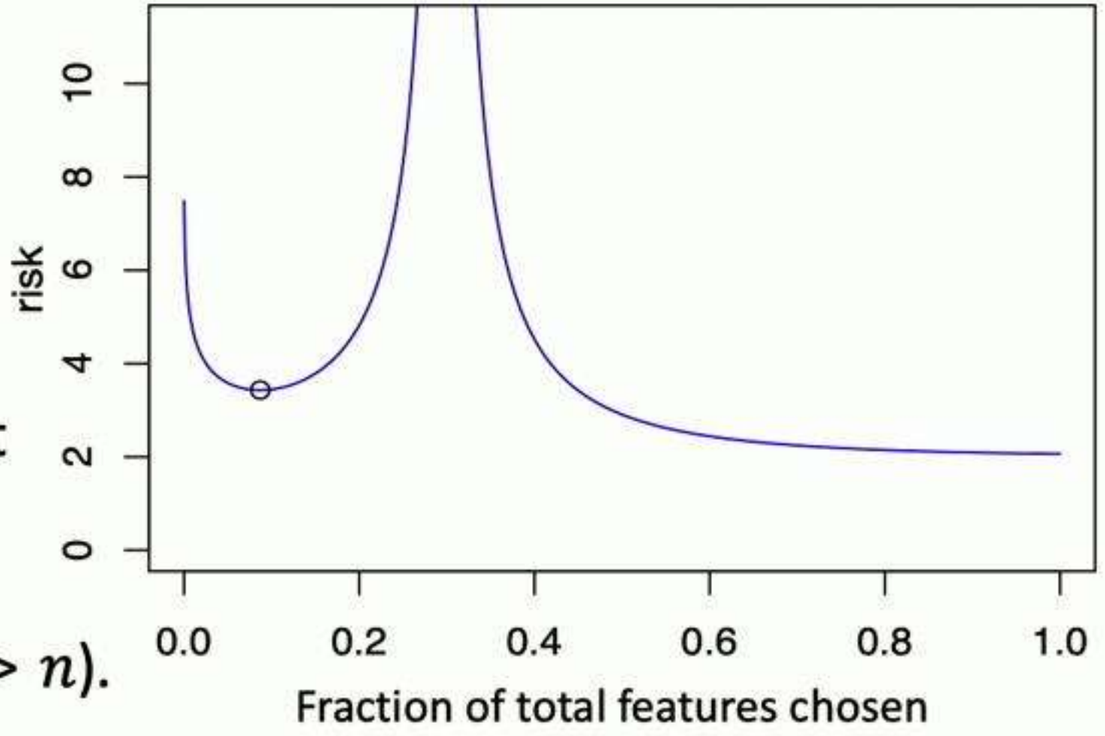
Trained first layer

Linear regression w/ weak features [Belkin, H., Xu, '19+; Xu, H., '19+]

Gaussian design linear model with D features
All features are "relevant" but equally weak

Only use p of the features ($1 \leq p \leq D$)
Least squares ($p \leq n$) or least norm ($p \geq n$) fit

Theorem: Under certain eigenvalue condition, minimum is beyond point of interpolation ($p > n$).



Concurrent work by Hastie, Montanari, Rosset, Tibshirani '19 (also for some non-linear models!).

Other analyses of linear models: Muthukumar, Vodrahalli, Sahai, '19; Bartlett, Long, Lugosi, Tsigler, '19.

On learning methods that memorize data

Sasha Rakhlin

MIT

Aug 26, 2019

Based on (Belkin, R., Tsybakov '18), (Liang and R. '18), (R. and Zhai '18), (Liang, R., Zhai '19)

Can a learning method be successful if it memorizes the training data?

Outline

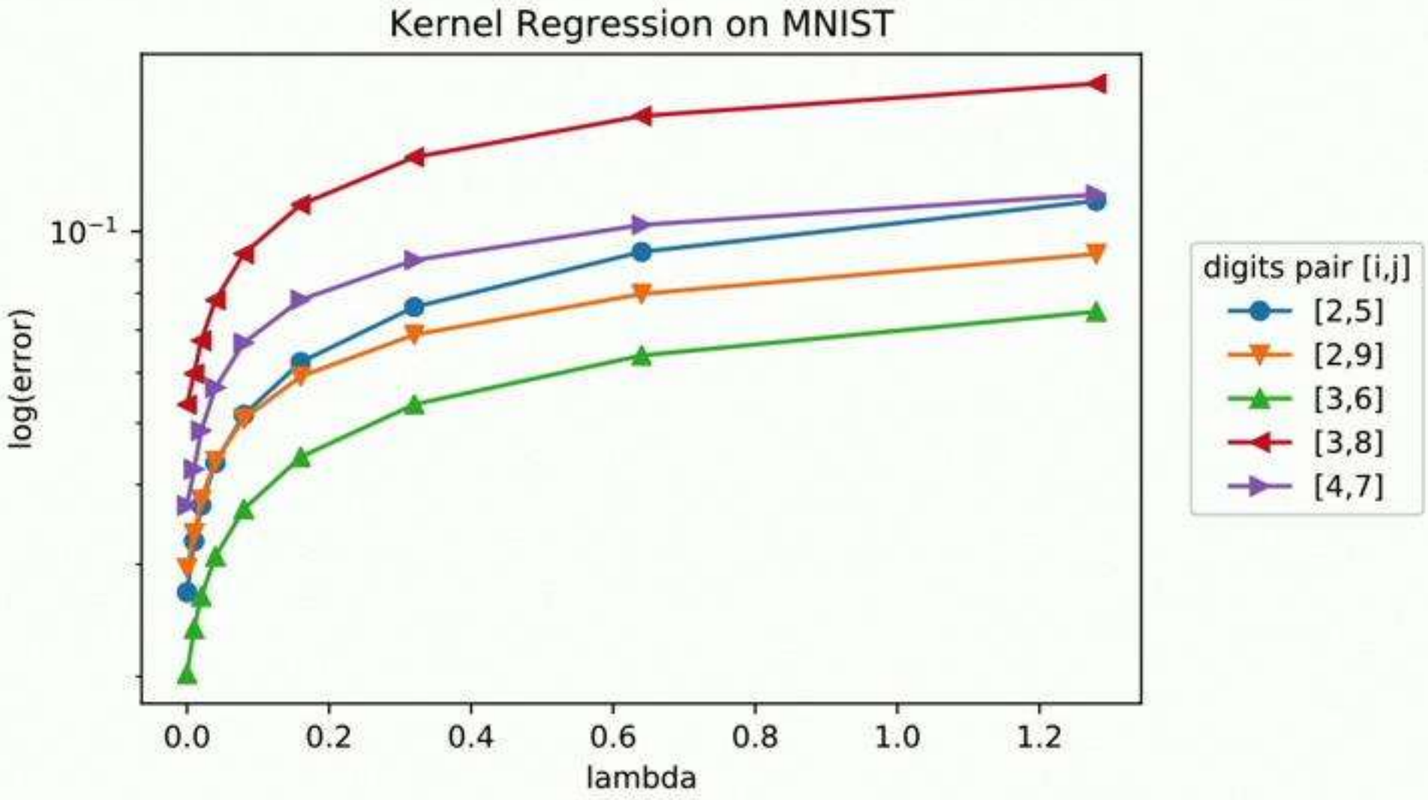
Warmup: The Nadaraya-Watson Estimator

Kernel Ridgeless Regression (minimum-norm interpolation)
- $d = O(1)$ regime
- $d \asymp n^{\alpha}$ regime, $\alpha \in (0, 1)$
- $d \asymp n$ regime

Application to Wide Neural Networks

Extra

An empirical example

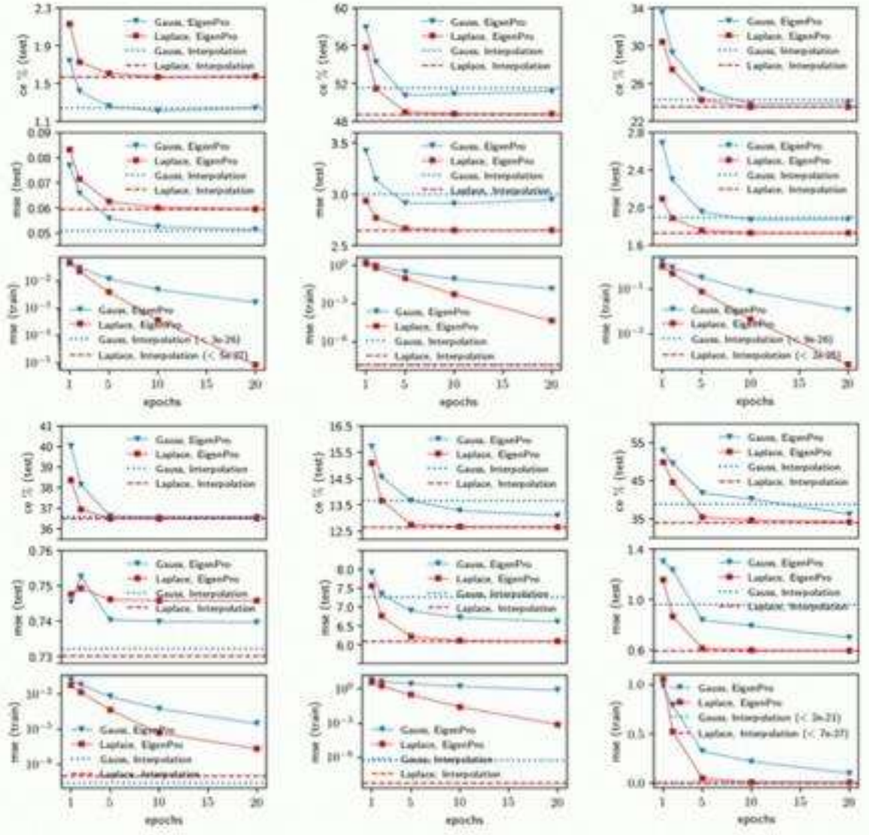


$\lambda = 0$: the interpolated solution, perfect fit on training data.

Isolated phenomenon? No

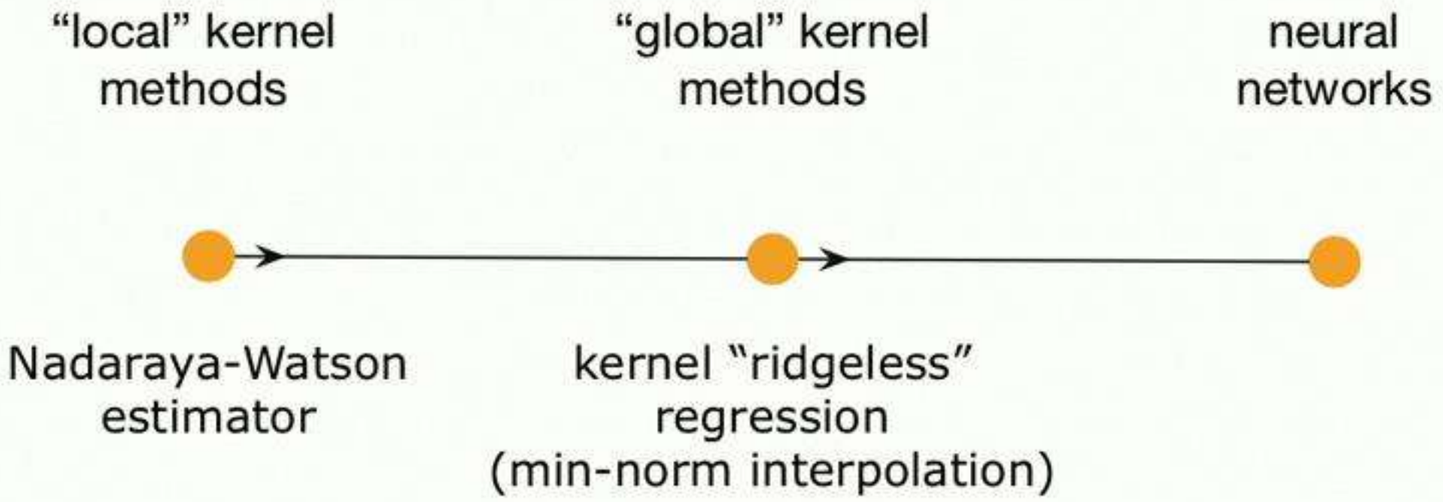
To Understand Deep Learning We Need to Understand Kernel Learning

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



(d) TIMIT ($5 \cdot 10^4$ subsamples) (e) HINT-S ($2 \cdot 10^4$ subsamples) (f) 20 Newsgroups

A study of generalization in the memorization regime



Outline

Warmup: The Nadaraya-Watson Estimator

Kernel Ridgeless Regression (minimum-norm interpolation)
- $d = O(1)$ regime
- $d \asymp n^{\alpha}$ regime, $\alpha \in (0, 1)$
- $d \asymp n$ regime

Application to Wide Neural Networks

Extra

Data $Y_i = f^*(X_i) + \epsilon_i$, $i = 1, \ldots, n$, and f^* is "smooth".

Classical Nadaraya-Watson estimator:

$$\widehat{f}_n(x) = \sum_{i=1}^{n} Y_i W_i(x)$$

where

$$W_i(x) = \frac{K((x - X_i)/h)}{\sum_{i=1}^{n} K((x - X_i)/h)}.$$

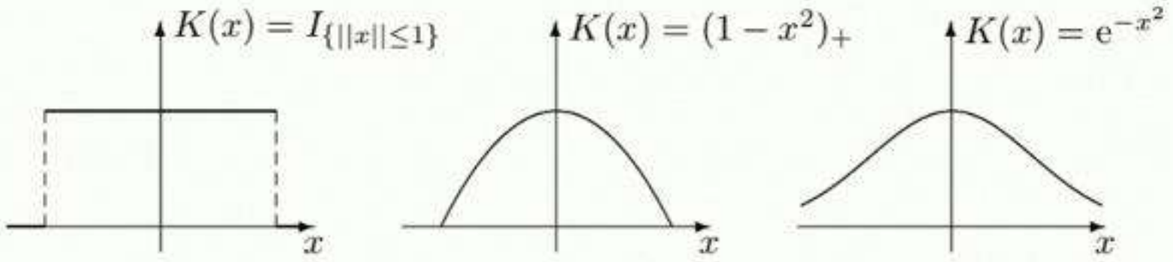


Figure 5.1. Examples for univariate kernels.

(figure from Györfi et al)

NB: "fit to data" controlled by kernel height at 0.

Example:

$$K(x) = \min\{1/\|x\|^a, \tau\}$$

We get $\widehat{f}_n(X_i) \to Y_i$ as $\tau \to \infty$.

Data $Y_i = f^*(X_i) + \epsilon_i$, $i = 1, \ldots, n$, and f^* is "smooth".

Classical Nadaraya-Watson estimator:

$$\widehat{f}_n(x) = \sum_{i=1}^{n} Y_i W_i(x)$$

where

$$W_i(x) = \frac{K((x - X_i)/h)}{\sum_{i=1}^{n} K((x - X_i)/h)}.$$

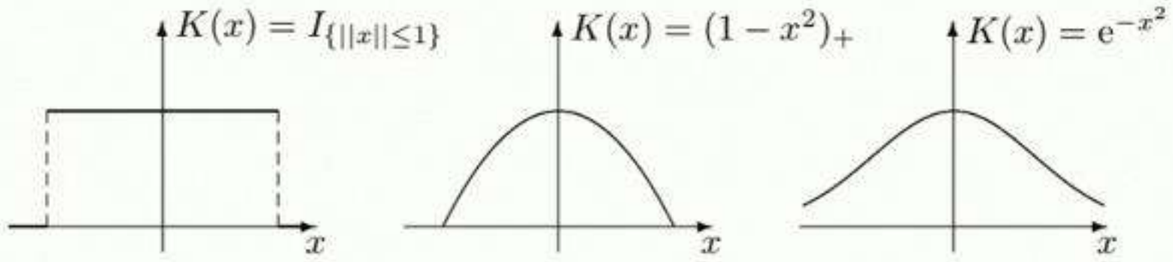


Figure 5.1. Examples for univariate kernels.

(figure from Györfi et al)

NB: "fit to data" controlled by kernel height at 0.

Example:

$$K(x) = \min\{1/|x|^a, \tau\}$$

We get $\widehat{f}_n(X_i) \to Y_i$ as $\tau \to \infty$.

(Devroye et al. (1998)): Nadaraya-Watson with kernel $K(x) = 1/|x|^d$ is

- **consistent:** $|\widehat{f}_n - f^*| \to 0$ in probability (if x has density).
- **pointwise convergence occurs nowhere:** there exists a distribution such that for all x in support, $\widehat{f}_n(x)$ does not converge to $f^*(x)$ almost surely as $n \to \infty$.

Problem: kernel is too global (h cancels) and bias cannot be controlled.

Belkin, Rakhlin, Tsybakov '18 (see also Belkin-Hsu-Mitra '18)

Our fix: truncate kernel as $K(u) \triangleq \|u\|^{-a} \mathbf{I}\{\|u\| \leq 1\}$.

Informal theorem: can choose bandwidth h in a minimax-optimal way such that

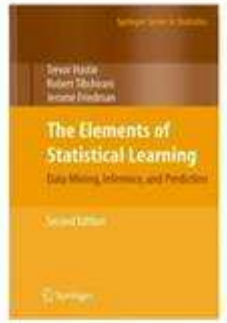
$$\mathbb{E}\left\|\widehat{f}_n - f^*\right\|^2 \leq C n^{-\frac{2\beta}{2\beta+d}},$$

under the assumption that f^* is β-Hölder. Note that $\widehat{f}_n$ is interpolating.

Summary:

- Fit to data does not necessarily say **anyting** about overfitting.
- Data fitting part (governed by parameter τ) is decoupled from the bias-variance trade-off (as given by parameter h).

Classical picture: U-shaped curve



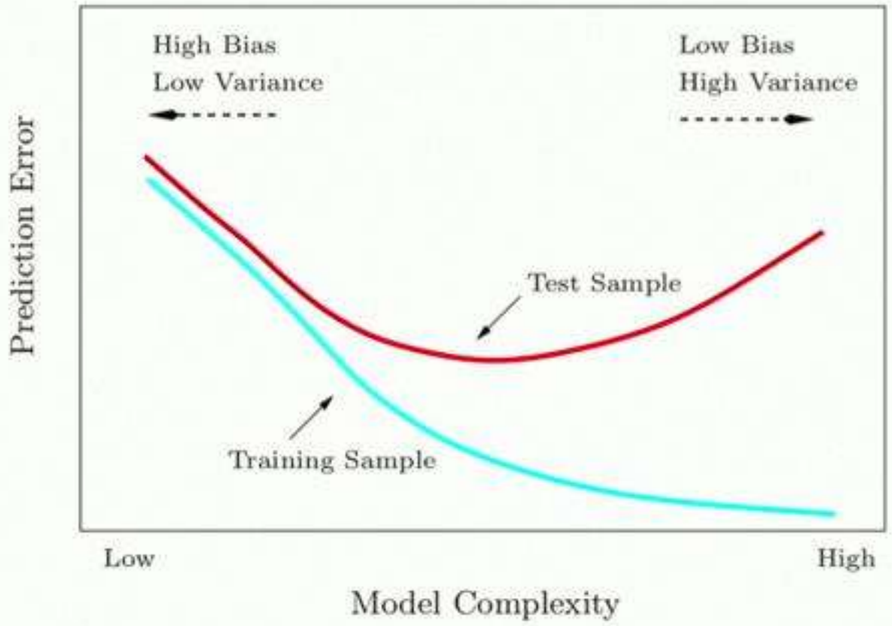


FIGURE 2.11. *Test and training error as a function of model complexity.*

Belkin, Rakhlin, Tsybakov '18 (see also Belkin-Hsu-Mitra '18)

Our fix: truncate kernel as $K(u) \triangleq \|u\|^{-a} \mathbf{I}\{\|u\| \leq 1\}$.

Informal theorem: can choose bandwidth h in a minimax-optimal way such that

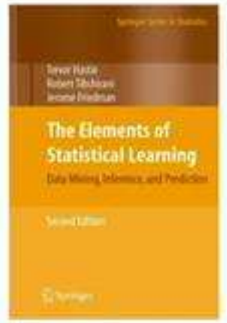
$$\mathbb{E}\left\|\widehat{f}_n - f^*\right\|^2 \leq C n^{-\frac{2\beta}{2\beta+d}},$$

under the assumption that f^* is β-Hölder. Note that $\widehat{f}_n$ is interpolating.

Summary:

- Fit to data does not necessarily say **anyting** about overfitting.
- Data fitting part (governed by parameter τ) is decoupled from the bias-variance trade-off (as given by parameter h).

Classical picture: U-shaped curve



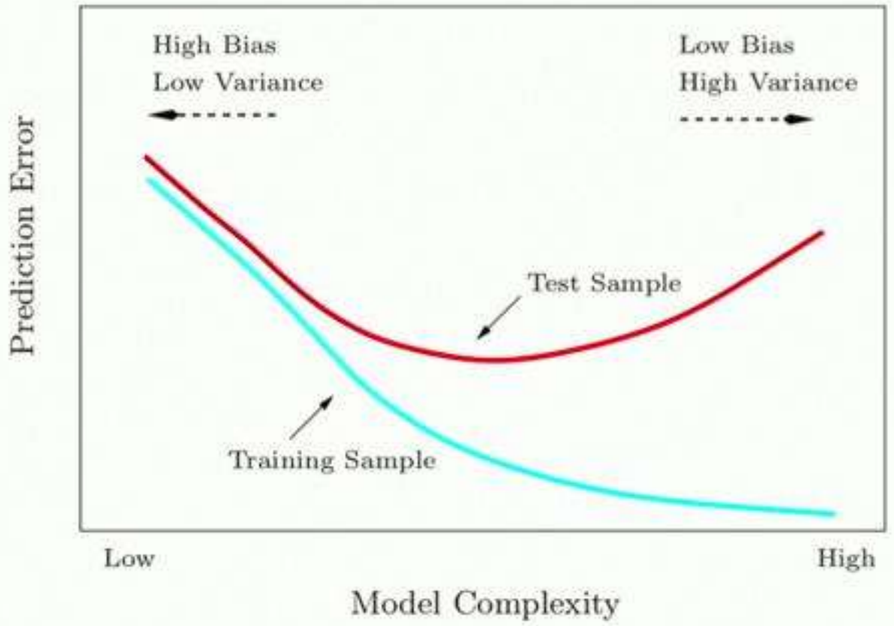


FIGURE 2.11. *Test and training error as a function of model complexity.*

Outline

Warmup: The Nadaraya-Watson Estimator

Kernel Ridgeless Regression (minimum-norm interpolation)
- $d = O(1)$ regime
- $d \asymp n^{\alpha}$ regime, $\alpha \in (0, 1)$
- $d \asymp n$ regime

Application to Wide Neural Networks

Extra

Kernel Ridge Regression:

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2 \ .$$

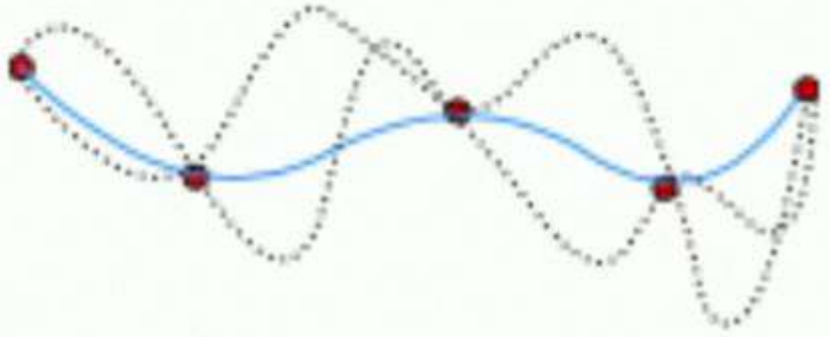
Conventional wisdom: explicit regularization $\lambda \neq 0$ added when the Reproducing Kernel Hilbert Space $\mathcal{H}$ is high- or infinite-dimensional

Puzzle: **Interpolated solutions** ($\lambda = 0$) often perform very well in practice

Belkin, Ma & Mandal (2018), Zhang, Bengio, Hardt, Recht & Vinyals (2016).

Min-norm interpolation:

$$\widehat{f}_n := \operatorname*{argmin}_{f \in \mathcal{H}} |f|_{\mathcal{H}}, \quad \text{s.t.} \quad f(x_i) = y_i, \ \forall i \leq n \ .$$



- "Simplicity" is still enforced, so it is conceivable that minimum-norm interpolant generalizes.
- GD/SGD started from 0 converges to minimum-norm solution.
- Also GD for wide neural networks can be related to kernel ridgeless regression.

Outline

Warmup: The Nadaraya-Watson Estimator

Kernel Ridgeless Regression (minimum-norm interpolation)
- $d = O(1)$ regime
- $d \asymp n^{\alpha}$ regime, $\alpha \in (0, 1)$
- $d \asymp n$ regime

Application to Wide Neural Networks

Extra

(Rakhlin and Zhai '18)

Take Laplace kernel

$$K_\sigma(x, x') = \sigma^{-d} \exp\{-\|x - x'\|/\sigma\}$$

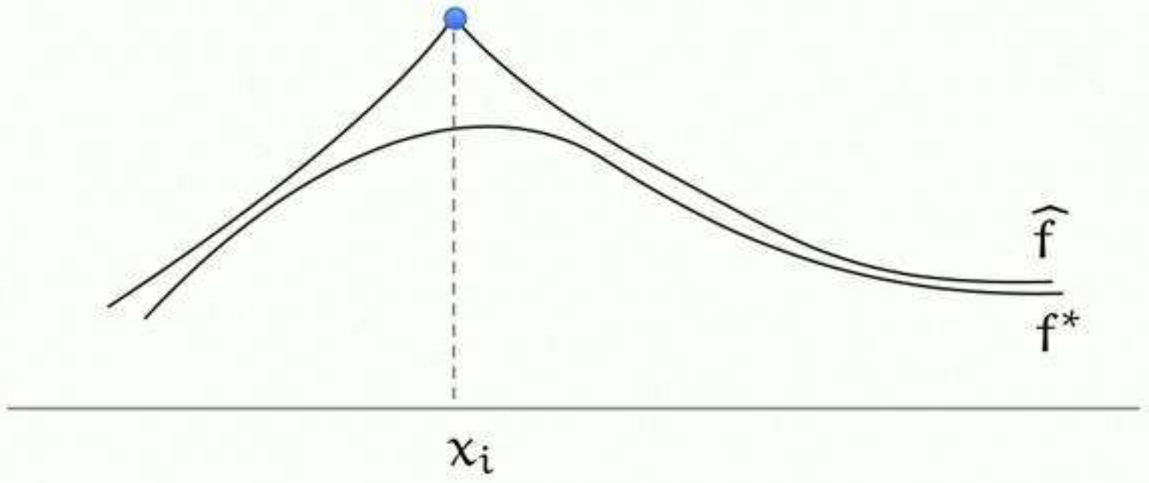
$\widehat{f}_n$ is minimum norm interpolation, as before.

Theorem: for odd dimension d, with probability $1 - O(n^{-1/2})$, for any choice of σ,

$$\mathbb{E}\left\|\widehat{f}_n - f^*\right\|^2 \geq \Omega_d(1).$$

Hence, interpolation with Laplace kernel does not work in constant d.

Lower bound for large σ



(Rakhlin and Zhai '18)

Take Laplace kernel

$$K_\sigma(x, x') = \sigma^{-d} \exp\{-\|x - x'\|/\sigma\}$$

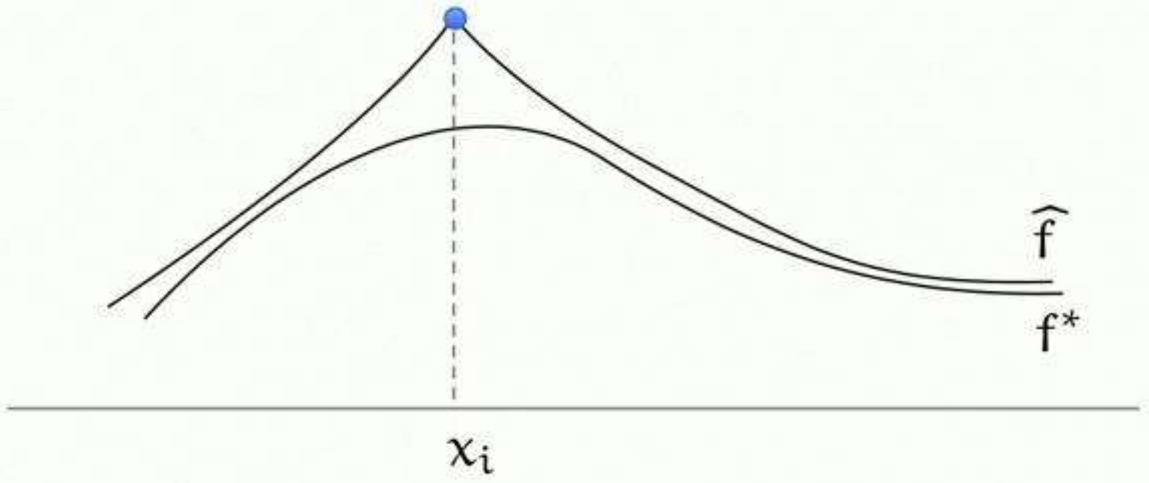
$\widehat{f}_n$ is minimum norm interpolation, as before.

Theorem: for odd dimension d, with probability $1 - O(n^{-1/2})$, for any choice of σ,

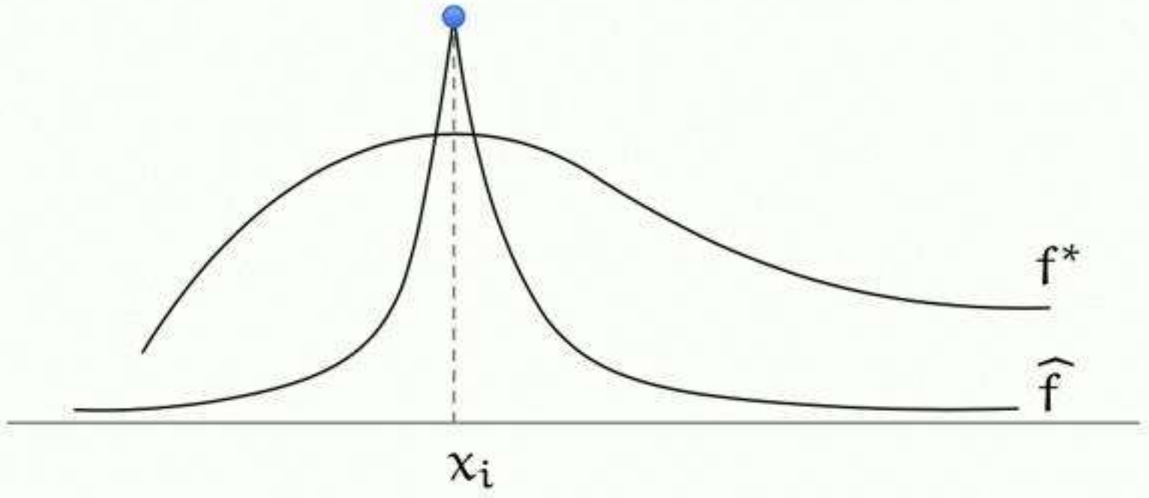
$$\mathbb{E}\left\|\widehat{f}_n - f^*\right\|^2 \geq \Omega_d(1).$$

Hence, interpolation with Laplace kernel does not work in constant d.

Lower bound for large σ



Lower bound for small σ



Interestingly, the two lower bounds cover the range of all σ, showing that Laplace interpolation cannot succeed in low dimension.

Outline

Warmup: The Nadaraya-Watson Estimator

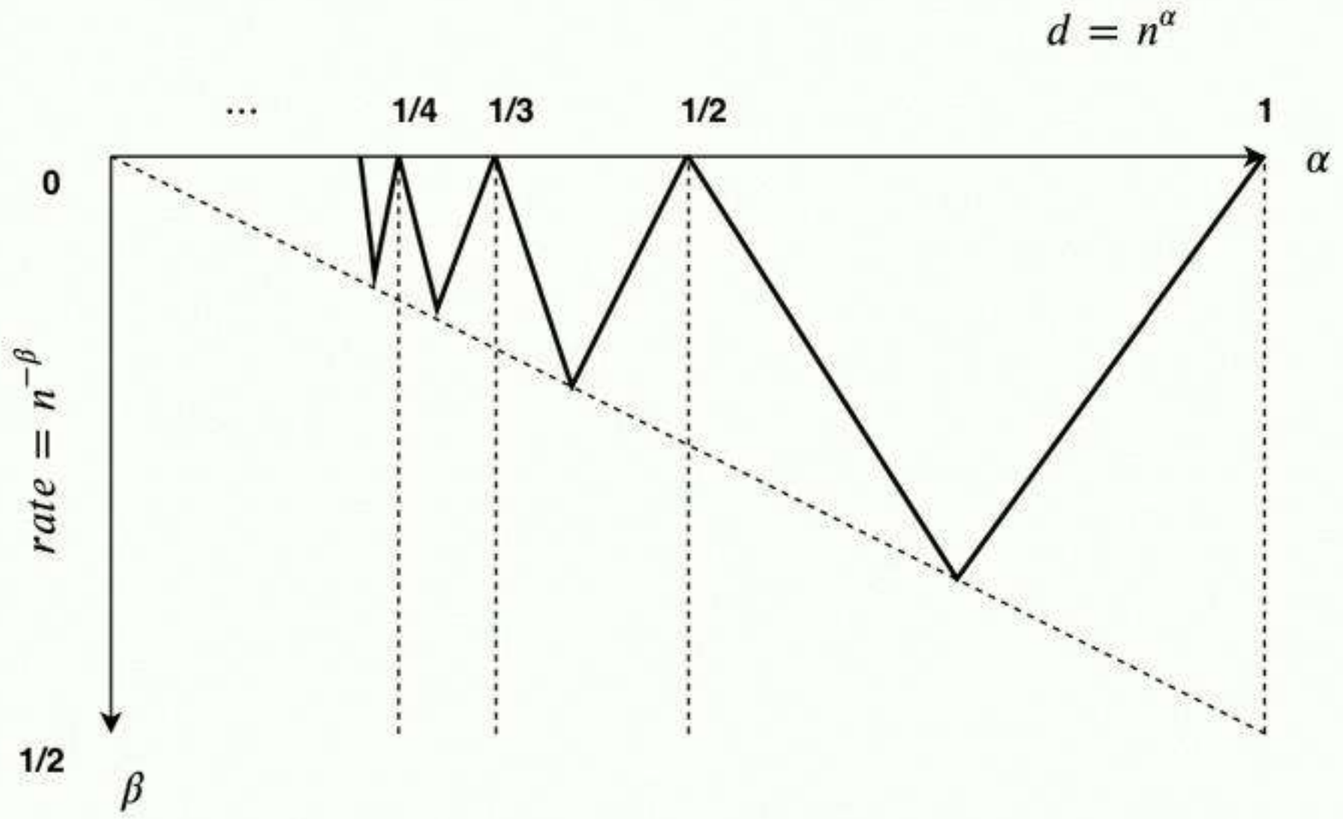
Kernel Ridgeless Regression (minimum-norm interpolation)
- $d = O(1)$ regime
- $d \asymp n^{\alpha}$ regime, $\alpha \in (0, 1)$
- $d \asymp n$ regime

Application to Wide Neural Networks

Extra

(Liang, Rakhlin, Zhai '19)

Upper bound on risk of minimum-norm interpolant has "multiple-descent" shape.



Note: we do not make assumptions on low effective rank of data (unlike (Bartlett et al '19) and (Liang and R. '18)).

Assumptions:

- Kernel is of the form $k(x, x') = h(\langle x, x' \rangle / d)$
- $x_1, \ldots, x_n \in \mathbb{R}^d$ are i.i.d. from $\mathcal{P}^{\otimes d}$ and $\mathcal{P}$ is sub-Gaussian
- Taylor expansion

$$h(z) = \sum_{i=0}^{\infty} \alpha_i z^i \tag{1}$$

converges for all $z \in \mathbb{R}$ with all coefficients $\alpha_i > 0$.

- Regression function $f_*(x) = \mathbb{E}[Y|X = x]$ satisfies

$$f_*(x) = \int K(x, z) \rho_*(z) dz \tag{2}$$

with $\rho_*(z)$ bounded.

(Liang, Rakhlin, Zhai '19)

Let $d = n^{\alpha}$ with $\alpha \in [\frac{1}{k_0+1}, \frac{1}{k_0}]$ for integer $k_0 \geq 1$.

Theorem (Informal).

With probability with high probability over draw of $\mathbf{X}$,

$$\mathbb{E}\left[\|\widehat{f} - f_*\|_{\mathcal{P}_X}^2 \mid \mathbf{X}\right] \leq C \cdot \left(\frac{d^{k_0}}{n} + \frac{n}{d^{k_0+1}}\right) \asymp n^{-\beta}$$

where $\beta := \min\{(k_0+1)\alpha - 1, 1 - k_0\alpha\}$.

Min-norm interpolant has closed form

$$f(x) = K(x, \mathbf{X})K(\mathbf{X})^{-1}Y$$

where $K(x, \mathbf{X}) = [K(x, x_1), \dots, K(x, x_n)]$ and $[K(\mathbf{X})]_{i,j} = K(x_i, x_j)$ is the kernel matrix.

Bias-variance decomposition, conditionally on $\mathbf{X}$

$$\mathbb{E}_Y\left[\|\widehat{f}_n - f_*\|_{\mathcal{P}_X}^2\right] = \underbrace{\mathbb{E}_Y\left[\|\widehat{f}_n - \mathbb{E}_Y[\widehat{f}_n]\|_{\mathcal{P}_X}^2\right]}_{\text{variance}} + \underbrace{\|\mathbb{E}_Y[\widehat{f}_n] - f_*\|_{\mathcal{P}_X}^2}_{\text{bias}^2} \tag{3}$$

Up to $\mathrm{var}(Y)$, the variance term is

$$\mathbb{E}_x\|K(x, \mathbf{X})K(\mathbf{X})^{-1}\|^2$$

and under our assumptions bias^2 is dominated by variance.

Assumptions:

- Kernel is of the form $k(x, x') = h(\langle x, x' \rangle / d)$
- $x_1, \ldots, x_n \in \mathbb{R}^d$ are i.i.d. from $\mathcal{P}^{\otimes d}$ and $\mathcal{P}$ is sub-Gaussian
- Taylor expansion

$$h(z) = \sum_{i=0}^{\infty} \alpha_i z^i \tag{1}$$

converges for all $z \in \mathbb{R}$ with all coefficients $\alpha_i > 0$.

- Regression function $f_*(x) = \mathbb{E}[Y|X = x]$ satisfies

$$f_*(x) = \int K(x, z)\rho_*(z)dz \tag{2}$$

with $\rho_*(z)$ bounded.

Bias-Variance

Min-norm interpolant has closed form

$$f(x) = K(x, \mathbf{X})K(\mathbf{X})^{-1}Y$$

where $K(x, \mathbf{X}) = [K(x, x_1), \dots, K(x, x_n)]$ and $[K(\mathbf{X})]_{i,j} = K(x_i, x_j)$ is the kernel matrix.

Bias-variance decomposition, conditionally on $\mathbf{X}$

$$\mathbb{E}_Y\left[\|\widehat{f}_n - f_*\|_{\mathcal{P}_X}^2\right] = \underbrace{\mathbb{E}_Y\left[\|\widehat{f}_n - \mathbb{E}_Y[\widehat{f}_n]\|_{\mathcal{P}_X}^2\right]}_{\text{variance}} + \underbrace{\|\mathbb{E}_Y[\widehat{f}_n] - f_*\|_{\mathcal{P}_X}^2}_{\text{bias}^2} \tag{3}$$

Up to $\mathrm{var}(Y)$, the variance term is

$$\mathbb{E}_x\|K(x, \mathbf{X})K(\mathbf{X})^{-1}\|^2$$

and under our assumptions bias^2 is dominated by variance.

Roadmap of the proof (I):

Consider truncated kernel

$$h^{[k]}(x) = \sum_{i=0}^{k} \alpha_i x^i, \qquad K_{ij}^{[k]} := h^{[k]}(\langle x_i, x_j \rangle / d)/n.$$

Roadmap of the proof (I):

Consider truncated kernel

$$h^{[k]}(x) = \sum_{i=0}^{k} \alpha_i x^i, \qquad K^{[k]}_{ij} := h^{[k]}(\langle x_i, x_j \rangle / d)/n.$$

Proposition (Restricted Lower Isometry Property for Kernel)**.**

With high prob, for any $k \le k_0$, $K^{[k]}$ has $\binom{d+k}{k}$ nonzero eigenvalues, all of them larger than Cd^{-k}, and the range of $K^{[k]}$ is

$$\{(p(x_1), \cdots, p(x_n)) : p \text{ is a multivar. polynomial of deg. } \le k\}.$$

Roadmap of the proof (II):

- Let $(r_1 \cdots r_d)_{k_0}$ be index in $\{1, 2, \cdots, \binom{k_0+d}{d}\}$
- Taylor expansion gives an $n \times \binom{k_0+d}{d}$ matrix Φ as

$$\Phi_{i,(r_1 \cdots r_d)_{k_0}} = \sqrt{c_{r_1 \cdots r_d} \alpha_{r_1+\cdots+r_d}} p_{r_1 \cdots r_d}(x_i)/d^{(r_1+\cdots+r_n)/2}$$

  such that

$$K^{[k_0]} = \frac{1}{n} \Phi \Phi^\top \ .$$

- Hard to analyze because of correlations.
- Gram-Schimdt Process to orthogonalize $1, z, z^2, \ldots$ with respect to $L_2(\mathcal{P})$. This gives $q_1, q_2, \ldots$ on $\mathbb{R}$.

$$\Psi_{i,(r_1 \cdots r_d)_{k_0}} = \sqrt{c_{r_1 \cdots r_d} \alpha_{r_1+\cdots+r_d}} q_{r_1 \cdots r_d}(x_i)/d^{(r_1+\cdots+r_d)/2} \ .$$

  where $q_{r_1 \cdots r_d}(x) := \prod_{i=1}^{d} q_{r_i}(x[i])$.
- Prove that this process gives $\Phi = \Psi \Lambda$ with bounded Λ in operator norm.

Roadmap of the proof (III):

- Let $f_\gamma(x) := \sum_{\le k_0} \gamma_{r_1 \cdots r_d} q_{r_1 \cdots r_d}(x)$, then

$$\mathbb{E} f_\gamma(x)^4 \lesssim (\mathbb{E} f_\gamma(x)^2)^2$$

- Show a "small-ball" property for functions

$$f_u(x) = \sum_{r_1, \cdots, r_d \ge 0, \sum_i r_i \le k_0} u_{(r_1 \cdots r_d)_{k_0}} \sqrt{c_{r_1 \cdots r_d} \alpha_{r_1 + \cdots + r_d}} q_{r_1 \cdots r_d}(x) / d^{(r_1 + \cdots + r_d)/2} \ .$$

That is, there exist θ, δ, such that for any u

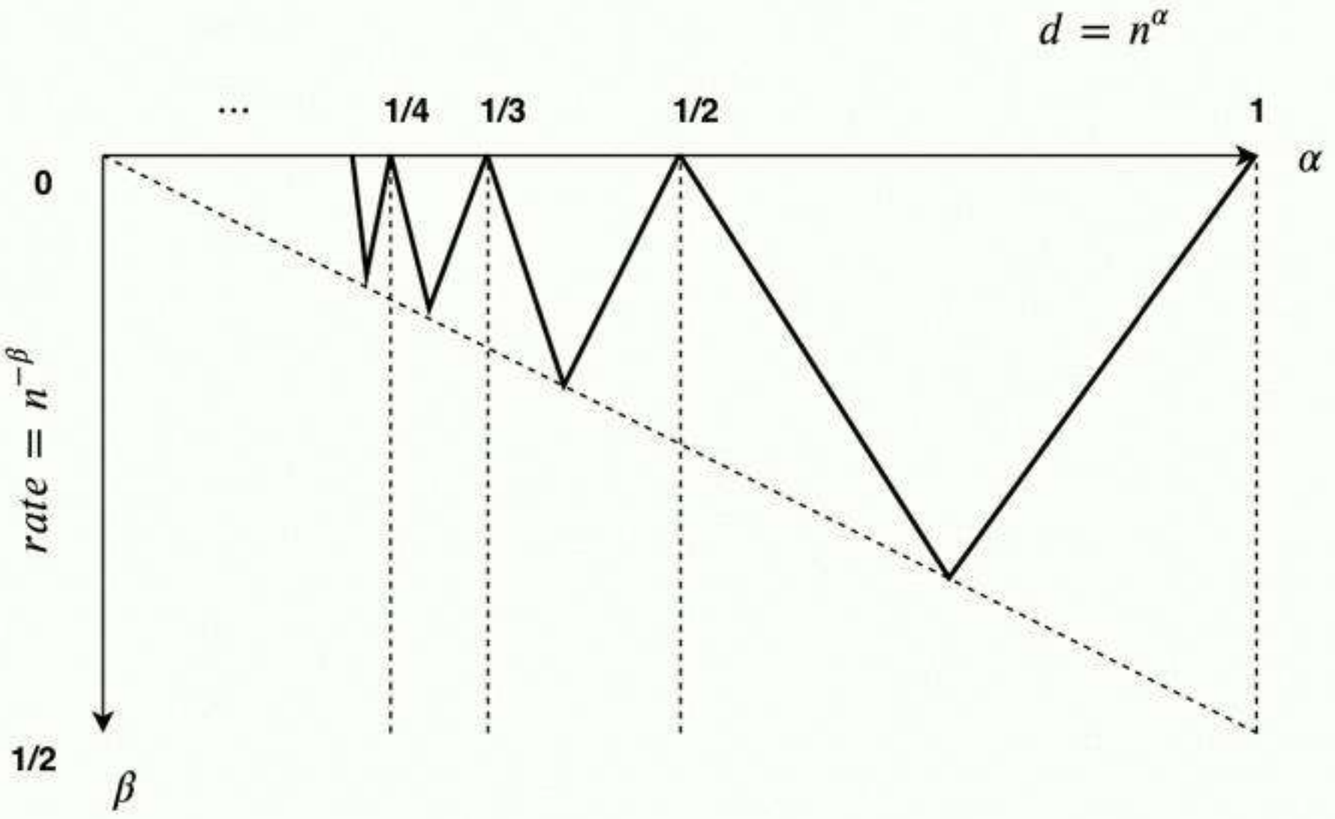
$$\mathbb{P}(f_u(x)^2 > \theta \mathbb{E}[f_u(x)^2]) \ge \delta \ .$$

- Lower bound eigen-values of

$$\Psi^\top \Psi / n > C d^{-k_0}$$

(Liang, Rakhlin, Zhai '19)

Upper bound on risk of minimum-norm interpolant has "multiple-descent" shape.



Outline

Warmup: The Nadaraya-Watson Estimator

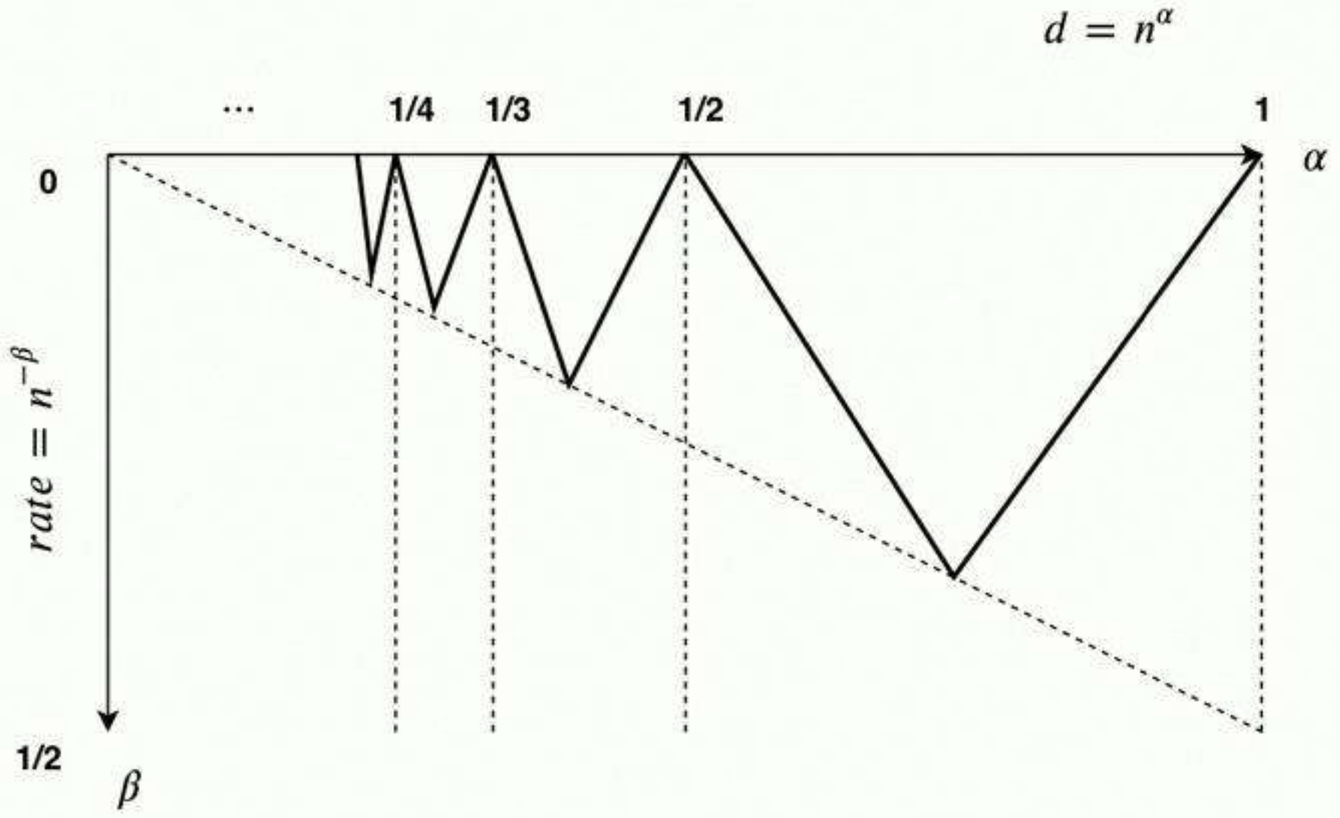
Kernel Ridgeless Regression (minimum-norm interpolation)
- $d = O(1)$ regime
- $d \asymp n^{\alpha}$ regime, $\alpha \in (0, 1)$
- $d \asymp n$ regime

Application to Wide Neural Networks

Extra

(Liang, Rakhlin, Zhai '19)

Upper bound on risk of minimum-norm interpolant has "multiple-descent" shape.



Assumptions

- **(A.1)** High-dim regime: both d, n large, $c \le d/n \le C$
  $\Sigma_d = \mathbf{E}_\mu[\mathbf{x}\mathbf{x}^*]$ satisfies $\|\Sigma_d\| \le C$ and $\mathrm{tr}(\Sigma_d)/d \ge c$.
- **(A.2)** $(8+m)$-moments: $z_i := \Sigma_d^{-1/2} x_i$
  each entries of z_i are i.i.d. mean zero, with bounded $(8+m)$-moments.
- **(A.3)** Noise condition: $\mathbf{E}[(f_*(\mathbf{x}) - \mathbf{y})^2 | \mathbf{x} = x] \le \sigma^2$ for all $x \in \Omega$.
- **(A.4)** Non-linear kernel: for a non-linear smooth function $h(\cdot)$

$$K(x, x') = h\left(\frac{1}{d}\langle x, x'\rangle\right)$$

NB: (A.4) can be replaced with $h\left(\frac{1}{d}\|x - x'\|^2\right)$.

Outline

Warmup: The Nadaraya-Watson Estimator

Kernel Ridgeless Regression (minimum-norm interpolation)
- $d = O(1)$ regime
- $d \asymp n^{\alpha}$ regime, $\alpha \in (0, 1)$
- $d \asymp n$ regime

Application to Wide Neural Networks

Extra

One-hidden-layer NN

$$f(x; W, a) := \frac{1}{\sqrt{m}} \sum_{i=1}^{m} a_i \sigma(w_i^\top \tilde{x}), \tag{5}$$

where $W = (w_1, \cdots, w_m) \in \mathbb{R}^{(d+1)\times m}$ matrix and $a \in \mathbb{R}^m$ and $\tilde{x} = (x^\top, \sqrt{d})^\top$.

Square loss:

$$L = \frac{1}{2n} \sum_{j=1}^{n} (f(x_j; W, a) - y_j)^2 \tag{6}$$

Gradient:

$$\frac{\partial L}{\partial a_i} = \frac{1}{n} \sum_{j=1}^{n} \frac{\sigma(w_i^\top \tilde{x}_j)}{\sqrt{m}} (f(x_j; W, a) - y_j), \tag{7}$$

and

$$\frac{\partial L}{\partial w_i} = \frac{1}{n} \sum_{j=1}^{n} \frac{a_i \tilde{x}_j \sigma'(w_i^\top \tilde{x}_j)}{\sqrt{m}} (f(x_j; W, a) - y_j). \tag{8}$$

Gradient flow yields

$$\frac{df(x; W(t), a(t))}{dt} = -\frac{1}{n}\sum_{j=1}^{n} h^m(x, x_j)(f(x_j; W, a) - y_j)$$

for

$$h^m(x, x') := \frac{1}{m}\left(\sum_{i=1}^{m} \sigma(w_i^\top \tilde{x})\sigma(w_i^\top \tilde{x}') + x^\top x' \sum_{i=1}^{m} a_i^2 \sigma'(w_i^\top \tilde{x})\sigma'(w_i^\top \tilde{x}')\right) \quad (9)$$

and f converges to the minimum-norm interpolant.

With $N(0,1)$ independent init, as $m \to \infty$, H^m converges to

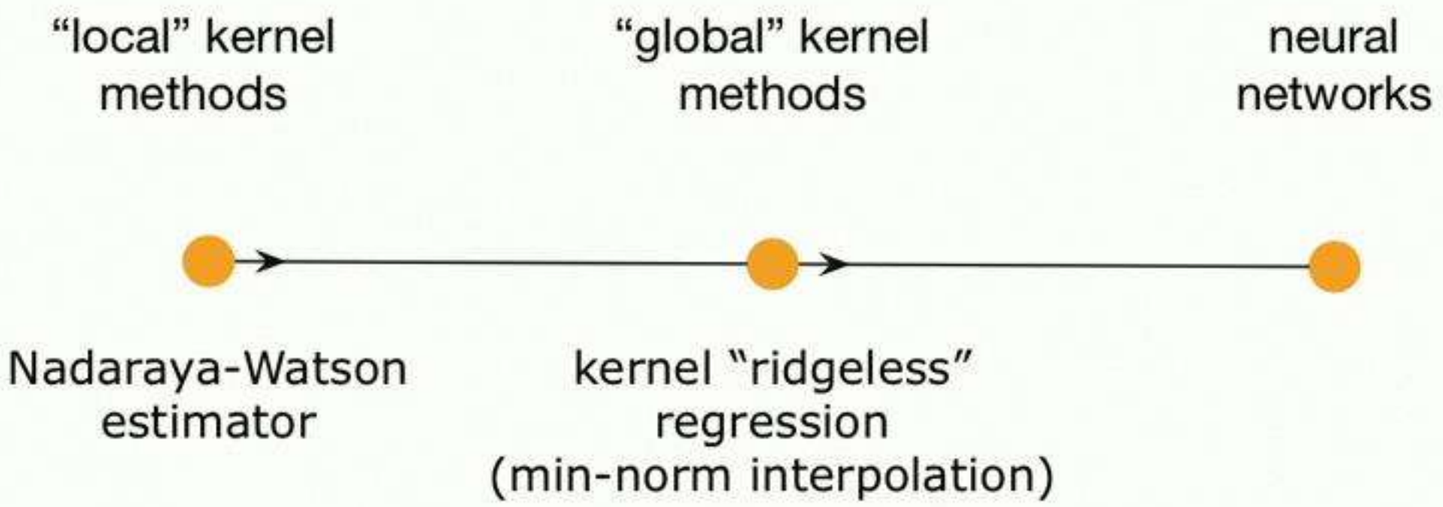
$$h^\infty(x, x') := \frac{1}{4\pi}\|\tilde{x}\|\|\tilde{x}'\|U(\cos\theta_{\tilde{x},\tilde{x}'})$$

where $\theta_{\tilde{x},\tilde{x}'}$ is the angle between $\tilde{x}$ and $\tilde{x}'$ and

$$U(t) := 3t(\pi - \arccos(t)) + \sqrt{1-t^2}$$

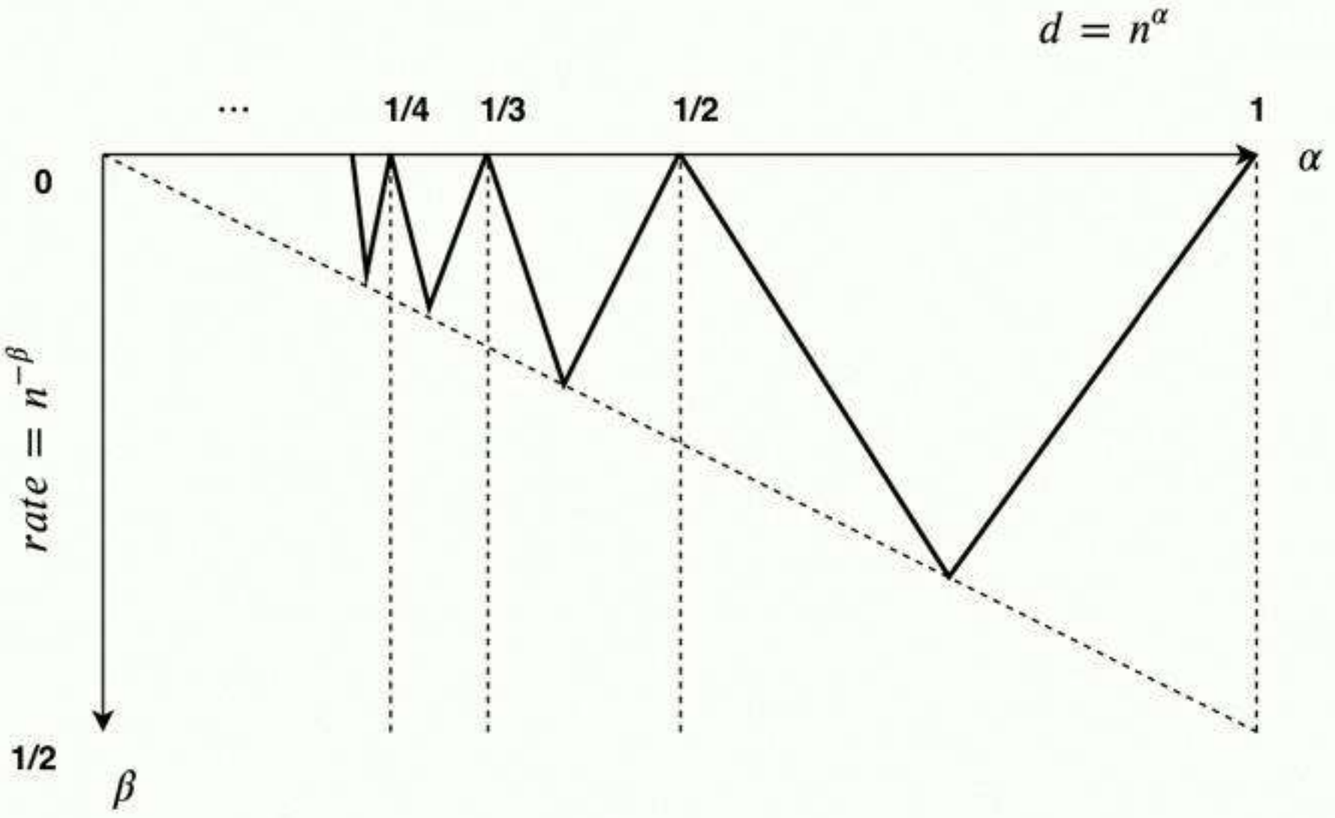
The theorem stated earlier in the talk for the $d \asymp n^\alpha$ regime can be extended to this kernel.

A study of generalization in the memorization regime



(Liang, Rakhlin, Zhai '19)

Upper bound on risk of minimum-norm interpolant has "multiple-descent" shape.



Assumptions:

- Kernel is of the form $k(x, x') = h(\langle x, x' \rangle / d)$
- $x_1, \ldots, x_n \in \mathbb{R}^d$ are i.i.d. from $\mathcal{P}^{\otimes d}$ and $\mathcal{P}$ is sub-Gaussian
- Taylor expansion

$$h(z) = \sum_{i=0}^{\infty} \alpha_i z^i \tag{1}$$

converges for all $z \in \mathbb{R}$ with all coefficients $\alpha_i > 0$.

- Regression function $f_*(x) = \mathbb{E}[Y|X = x]$ satisfies

$$f_*(x) = \int K(x, z) \rho_*(z) dz \tag{2}$$

with $\rho_*(z)$ bounded.

(Liang, Rakhlin, Zhai '19)

Let $d = n^{\alpha}$ with $\alpha \in [\frac{1}{k_0+1}, \frac{1}{k_0}]$ for integer $k_0 \geq 1$.

Theorem (Informal).

With probability with high probability over draw of $\mathbf{X}$,

$$\mathbb{E}\left[\|\widehat{f} - f_*\|_{\mathcal{P}_X}^2 \mid \mathbf{X}\right] \leq C \cdot \left(\frac{d^{k_0}}{n} + \frac{n}{d^{k_0+1}}\right) \asymp n^{-\beta}$$

where $\beta := \min\{(k_0+1)\alpha - 1, 1 - k_0\alpha\}$.

Assumptions:

- Kernel is of the form $k(x, x') = h(\langle x, x' \rangle / d)$
- $x_1, \ldots, x_n \in \mathbb{R}^d$ are i.i.d. from $\mathcal{P}^{\otimes d}$ and $\mathcal{P}$ is sub-Gaussian
- Taylor expansion

$$h(z) = \sum_{i=0}^{\infty} \alpha_i z^i \tag{1}$$

converges for all $z \in \mathbb{R}$ with all coefficients $\alpha_i > 0$.

- Regression function $f_*(x) = \mathbb{E}[Y|X = x]$ satisfies

$$f_*(x) = \int K(x, z)\rho_*(z)dz \tag{2}$$

with $\rho_*(z)$ bounded.

The generalization error of random feature models

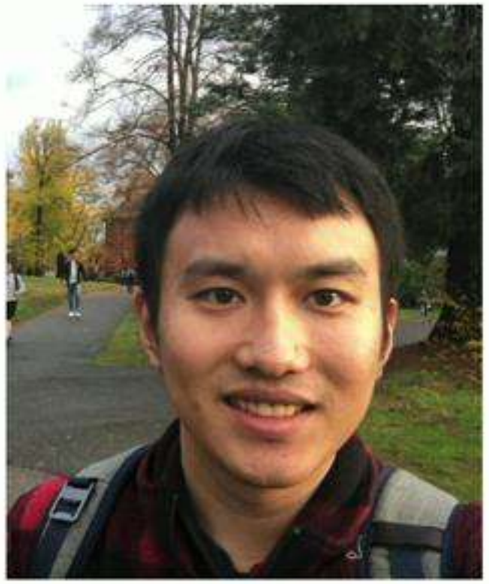
Andrea Montanari

Stanford University

August 26, 2019

3 papers, 6 co-authors

"... Precise asymptotics and double descent"

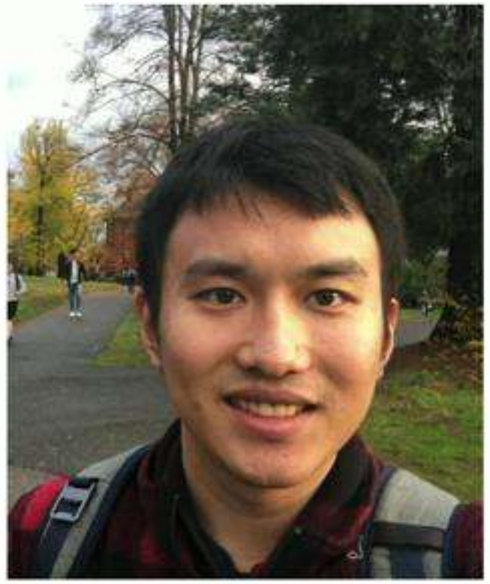


Song Mei

arXiv:1908.05355

"Linearized two-layers neural networks..."

Behrooz Ghorbani

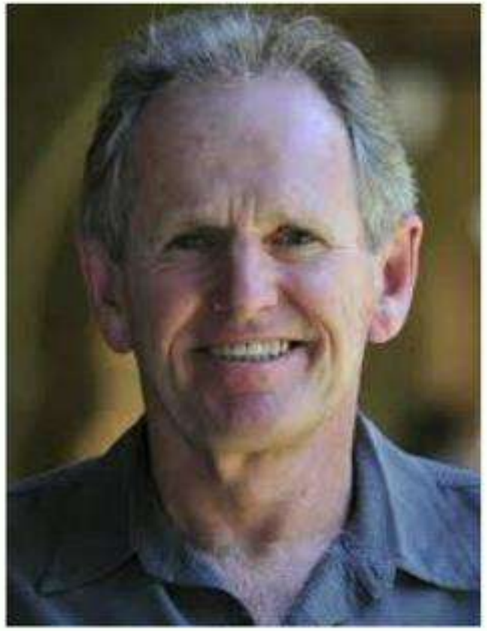


Song Mei

Theodor Misiakiewicz

arXiv:1904.12191

"Surprises in high-dimensional ridgeless ..."



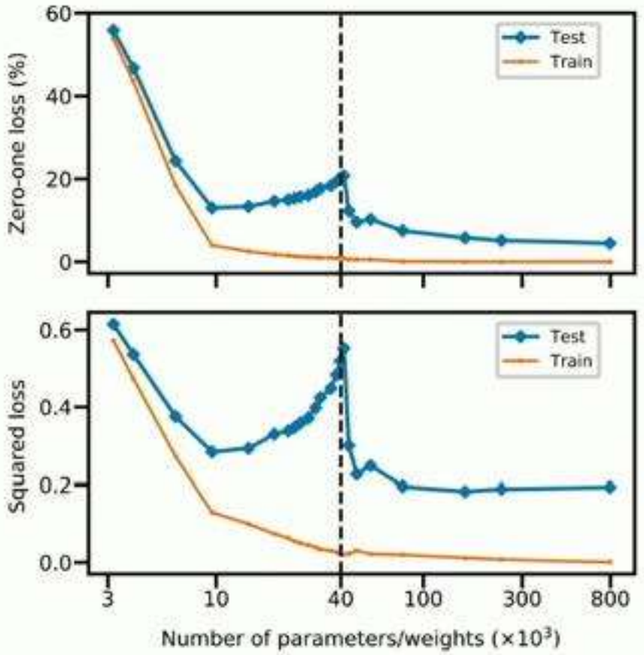
Trevor Hastie

Saharon Rosset

Ryan Tibshirani

arXiv:1903.08560

Motivation

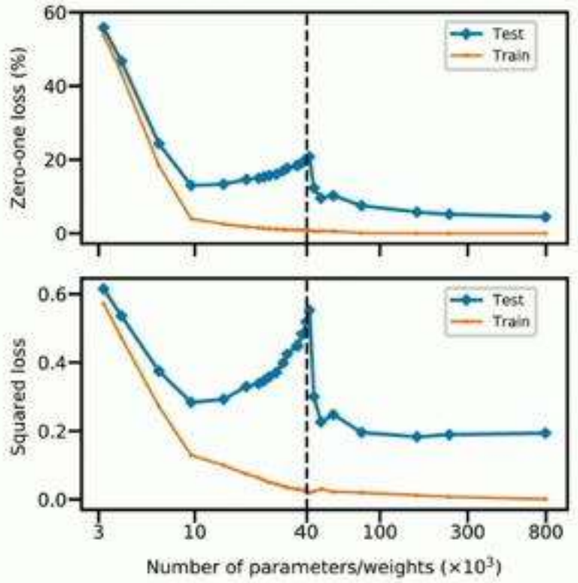


Belkin, Hsu, Mandal, 2019
Two-layers fully connected, MNIST

Do we understand these curves?

Defining the question

Desiderata



- ✓ Peak at the interpolation threshold
- ✓ Global minimum in the overparametrized regime
- ✓ Monotone decreasing in the overparametrized regime
- ✓ Vanishing (explicit) regularization

Ideally: Two-layers neural networks

$$\mathcal{F}_{\mathsf{NN}} \equiv \left\{ f(\boldsymbol{x}) = \sum_{i=1}^{N} a_i \, \sigma(\langle \boldsymbol{w}_i, \boldsymbol{x} \rangle) \; : \quad a_i \in \mathbb{R}, \boldsymbol{w}_i \in \mathbb{R}^d \; \forall i \le N \right\}.$$

- Train by SGD until TrainingError$= 0$
- Compute TestError for this network
- **Perhaps in a couple of years**

Two linearizations...

Rahimi-Recht 2008

$$\mathcal{F}_{\mathsf{RF}}(\boldsymbol{W}) \equiv \left\{ f(\boldsymbol{x}) = \sum_{i=1}^{N} a_i \, \sigma(\langle \boldsymbol{w}_i, \boldsymbol{x} \rangle) \ : \quad a_i \in \mathbb{R} \ \forall i \le N \right\}.$$

- $\boldsymbol{W} = [\boldsymbol{w}_1, \dots, \boldsymbol{w}_N]$ random; $\quad \boldsymbol{w}_i \sim \mathsf{Unif}(\mathbb{S}^{d-1}(1))$, i.i.d.

Jacot-Gabriel-Hongler 2018

$$\mathcal{F}_{\mathsf{NT}}(\boldsymbol{W}) \equiv \left\{ f(\boldsymbol{x}) = \sum_{i=1}^{N} \langle \boldsymbol{a}_i, \boldsymbol{x} \rangle \, \sigma'(\langle \boldsymbol{w}_i, \boldsymbol{x} \rangle) \ : \quad \boldsymbol{a}_i \in \mathbb{R}^d \ \forall i \le N \right\}.$$

- $\boldsymbol{W} = [\boldsymbol{w}_1, \dots, \boldsymbol{w}_N]$ random; $\quad \boldsymbol{w}_i \sim \mathsf{Unif}(\mathbb{S}^{d-1}(1))$, i.i.d.
- Linear expansion of NN around random initialization

The rest of this talk

1. A stylized (Gaussian) model

2. Nonlinear features: Approximation in high dimension

3. Nonlinear features: Generalization curve

A stylized (Gaussian) model

Setting

- Data $\{(y_i, x_i)\}_{i\le n}$, $z_i = \phi(x_i) \in \mathbb{R}^p$

$$y_i = \langle \beta_0, z_i \rangle + \varepsilon_i, \qquad \varepsilon_i \sim \mathsf{N}(0, \tau^2)$$

- 'Ridgeless regression'

$$\widehat{\beta}(\lambda) \equiv \arg\min_{\beta \in \mathbb{R}^p} \left\{ \|y - Z\beta\|_2^2 + \lambda \|\beta\|_2^2 \right\},$$
$$\widehat{\beta}(0) \equiv \lim_{\lambda \to 0+} \widehat{\beta}(\lambda)$$

- Test error

$$R(\beta_0) = \mathbb{E}\{(\langle \widehat{\beta}, z^{\text{new}} \rangle - \langle \beta_0, z^{\text{new}} \rangle)^2\}$$

A stylized model

- **Gaussian features**

$$z_i \sim \mathsf{N}(\mathbf{0}, \boldsymbol{\Sigma})$$

Setting

- Data $\{(y_i, x_i)\}_{i\le n}$, $z_i = \phi(x_i) \in \mathbb{R}^p$

$$y_i = \langle \beta_0, z_i \rangle + \varepsilon_i, \qquad \varepsilon_i \sim \mathsf{N}(0, \tau^2)$$

- 'Ridgeless regression'

$$\widehat{\beta}(\lambda) \equiv \arg\min_{\beta \in \mathbb{R}^p} \left\{ \|y - Z\beta\|_2^2 + \lambda \|\beta\|_2^2 \right\},$$
$$\widehat{\beta}(0) \equiv \lim_{\lambda \to 0+} \widehat{\beta}(\lambda)$$

- Test error

$$R(\beta_0) = \mathbb{E}\{(\langle \widehat{\beta}, z^{\text{new}} \rangle - \langle \beta_0, z^{\text{new}} \rangle)^2\}$$

Precise asymptotics

Theorem (Hastie, M., Rosset, Tibshirani, 2019)

If $p/n \to \gamma$, $\boldsymbol{\Sigma} = \boldsymbol{I}_p$, $\|\boldsymbol{\beta}_0\|_2 \to r^2$, *then*

$$R(\boldsymbol{\beta}_0) \to \begin{cases} \tau^2 \frac{\gamma}{1-\gamma} & \text{for } \gamma < 1, \\ \tau^2 \frac{1}{\gamma - 1} + r^2\left(1 - \frac{1}{\gamma}\right) & \text{for } \gamma > 1. \end{cases}$$

Further:

- *If* $\Sigma \neq \boldsymbol{I}_p$ ...
- *If* $z_i = \Sigma^{1/2} x_i$, x_i *with iid components* ...

[Belkin, Hsu, Xu, 2019; Bartlett, Long, Lugosi, Tsigler, 2019]

Setting

- Data $\{(y_i, x_i)\}_{i\le n}$, $z_i = \phi(x_i) \in \mathbb{R}^p$

$$y_i = \langle \beta_0, z_i \rangle + \varepsilon_i, \qquad \varepsilon_i \sim \mathsf{N}(0, \tau^2)$$

- 'Ridgeless regression'

$$\widehat{\beta}(\lambda) \equiv \arg\min_{\beta \in \mathbb{R}^p} \left\{ \|y - Z\beta\|_2^2 + \lambda \|\beta\|_2^2 \right\},$$
$$\widehat{\beta}(0) \equiv \lim_{\lambda \to 0+} \widehat{\beta}(\lambda)$$

- Test error

$$R(\beta_0) = \mathbb{E}\{(\langle \widehat{\beta}, z^{\text{new}} \rangle - \langle \beta_0, z^{\text{new}} \rangle)^2\}$$

Precise asymptotics

Theorem (Hastie, M., Rosset, Tibshirani, 2019)

If $p/n \to \gamma$, $\boldsymbol{\Sigma} = \boldsymbol{I}_p$, $\|\boldsymbol{\beta}_0\|_2 \to r^2$, *then*

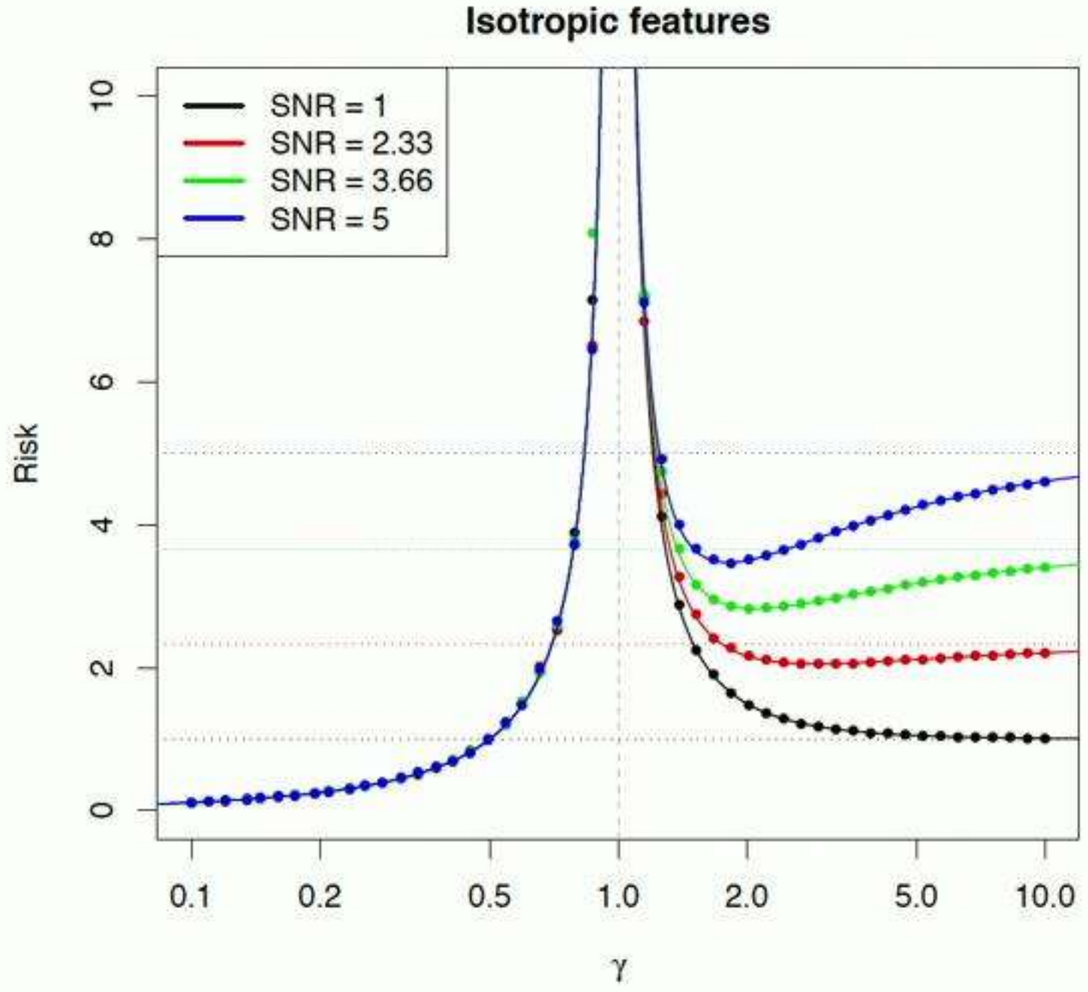
$$R(\boldsymbol{\beta}_0) \to \begin{cases} \tau^2 \frac{\gamma}{1-\gamma} & \text{for } \gamma < 1, \\ \tau^2 \frac{1}{\gamma - 1} + r^2\left(1 - \frac{1}{\gamma}\right) & \text{for } \gamma > 1. \end{cases}$$

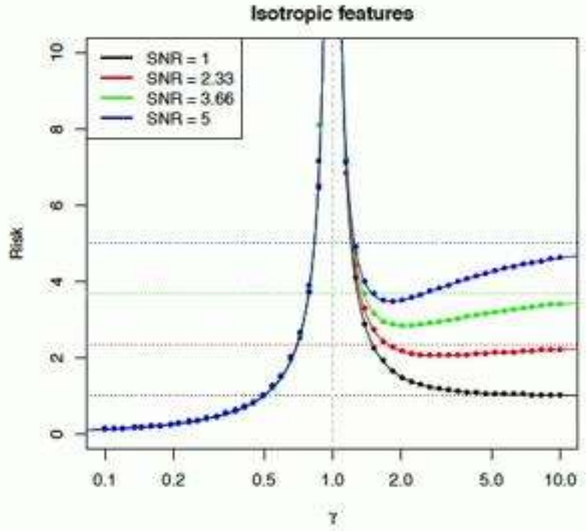
Further:

- *If* $\boldsymbol{\Sigma} \neq \boldsymbol{I}_p$ ...
- *If* $\boldsymbol{z}_i = \boldsymbol{\Sigma}^{1/2} \boldsymbol{x}_i$, $\boldsymbol{x}_i$ *with iid components* ...

[Belkin, Hsu, Xu, 2019; Bartlett, Long, Lugosi, Tsigler, 2019]

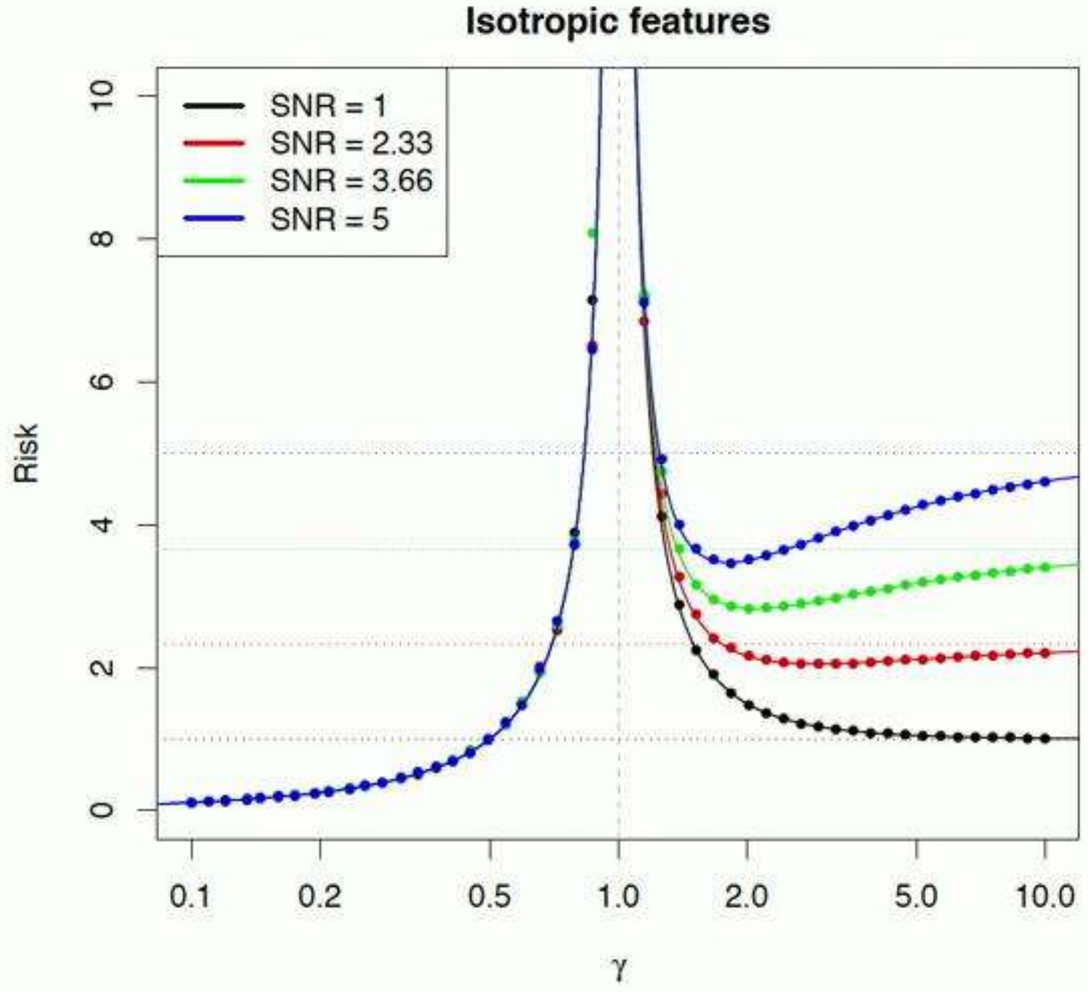
$\boldsymbol{\Sigma} = \boldsymbol{I}_p$, $\mathrm{SNR} = \|\boldsymbol{\beta}_0\|_2^2/\tau^2$





- ✓ Peak at the interpolation threshold
- ✗ Global minimum in the overparametrized regime
- ✗ Monotone decreasing in the overparametrized regime
- ✓ Vanishing (explicit) regularization
- ▸ Dimension = Number of parameters

$\boldsymbol{\Sigma} = \boldsymbol{I}_p$, $\mathrm{SNR} = \|\boldsymbol{\beta}_0\|_2^2/\tau^2$



Precise asymptotics

Theorem (Hastie, M., Rosset, Tibshirani, 2019)

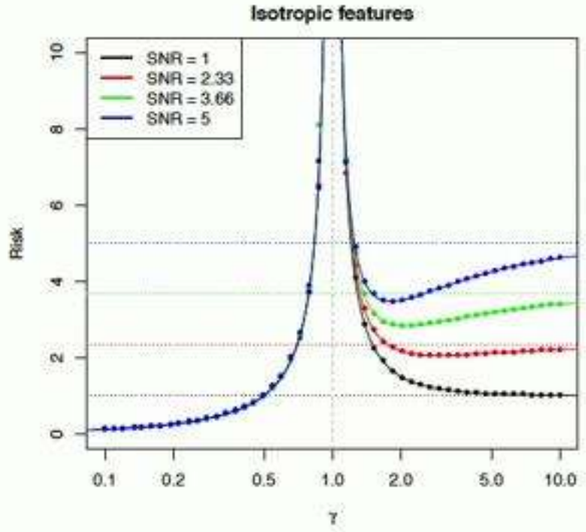
If $p/n \to \gamma$, $\boldsymbol{\Sigma} = \boldsymbol{I}_p$, $\|\boldsymbol{\beta}_0\|_2 \to r^2$, *then*

$$R(\boldsymbol{\beta}_0) \to \begin{cases} \tau^2 \frac{\gamma}{1-\gamma} & \text{for } \gamma < 1, \\ \tau^2 \frac{1}{\gamma - 1} + r^2\left(1 - \frac{1}{\gamma}\right) & \text{for } \gamma > 1. \end{cases}$$

Further:

- *If* $\boldsymbol{\Sigma} \neq \boldsymbol{I}_p$ ...
- *If* $\boldsymbol{z}_i = \boldsymbol{\Sigma}^{1/2} \boldsymbol{x}_i$, $\boldsymbol{x}_i$ *with iid components* ...

[Belkin, Hsu, Xu, 2019; Bartlett, Long, Lugosi, Tsigler, 2019]



- ✓ Peak at the interpolation threshold
- ✗ Global minimum in the overparametrized regime
- ✗ Monotone decreasing in the overparametrized regime
- ✓ Vanishing (explicit) regularization
- ▸ Dimension = Number of parameters

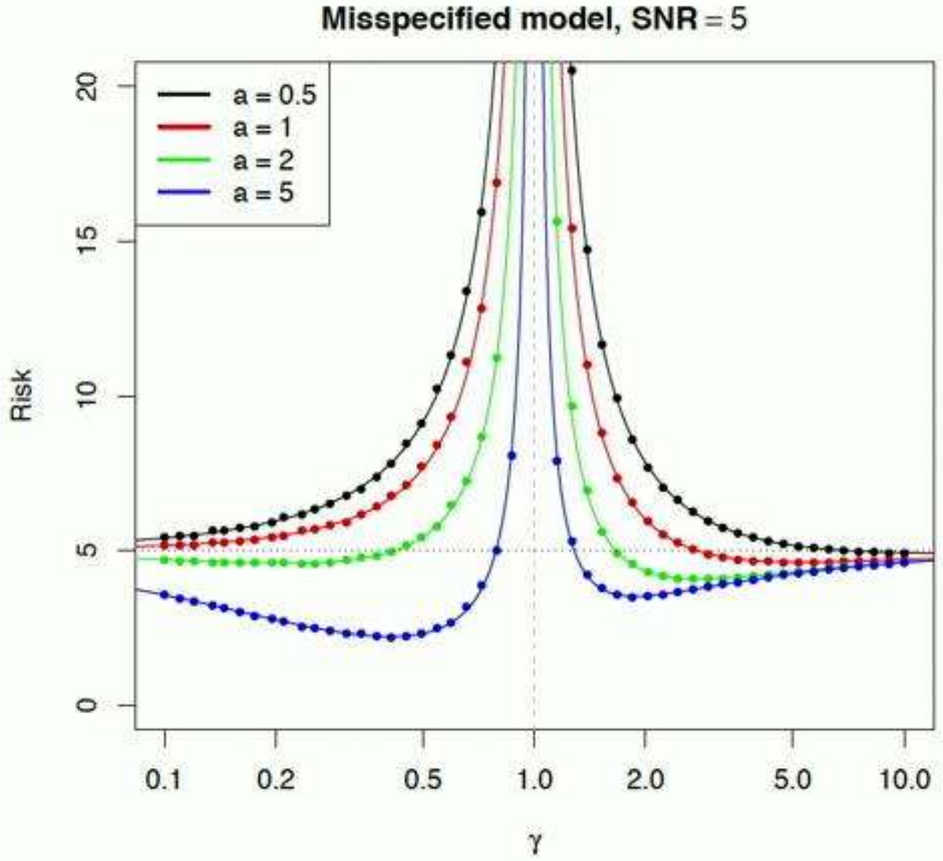
Misspecified model

- $$y_i = \langle \beta_0, z_i \rangle + \langle \tilde{\beta}_0, \tilde{z}_i \rangle + \varepsilon_i, \qquad \varepsilon_i \sim \mathsf{N}(0, \tau^2)$$

- 'Ridgeless regression'

$$\widehat{\beta}(\lambda) \equiv \arg\min_{a \in \mathbb{R}^p} \left\{ \|y - Z\beta\|_2^2 + \lambda \|\beta\|_2^2 \right\},$$

Misspecified Gaussian model



- ✓ Global minimum in the overparametrized regime
- ✗ Monotone decreasing in the overparametrized regime
- ▶ Dimension $\neq$ Number of parameters

Nonlinear features: Approximation in high dimension

Setting

Nonparametric regression

Setting

- $\{(y_i, \boldsymbol{x}_i)\}_{i \le n}$ i.i.d. $\qquad \boldsymbol{x}_i \sim \mathsf{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$,

$$y_i = f_\star(\boldsymbol{x}_i) + \varepsilon_i \qquad \varepsilon_i \sim \mathsf{N}(0, \tau^2)$$

- Random features $\qquad \boldsymbol{w}_i \sim \mathsf{Unif}(\mathbb{S}^{d-1}(1))$,

$$\mathcal{F}_{\mathsf{RF}}(\boldsymbol{W}) \equiv \left\{ f(\boldsymbol{x}) = \sum_{i=1}^{N} a_i \, \sigma(\langle \boldsymbol{w}_i, \boldsymbol{x} \rangle) \ : \quad a_i \in \mathbb{R} \ \forall i \le N \right\}.$$

- Ridge

$$\widehat{\boldsymbol{a}}(\lambda) = \mathrm{argmin}_{\boldsymbol{a} \in \mathbb{R}^N} \left\{ \widehat{\mathbb{E}}_n \Big[\Big(y - \sum_{i=1}^{N} a_i \sigma(\langle \boldsymbol{w}_i, \boldsymbol{x} \rangle \Big)^2 \Big] + \frac{N\lambda}{d} \, \|\boldsymbol{a}\|_2^2 \right\}.$$

Setting

- Test error

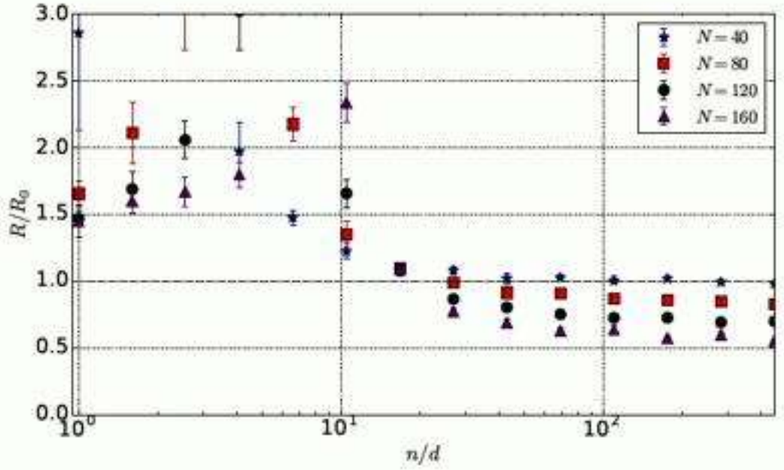
$$R_{\mathsf{RF}}(\hat{f}; f) \equiv \mathbb{E}\{(\hat{f}(\boldsymbol{x}) - f_\star(\boldsymbol{x}))^2\}\,.$$

- First question

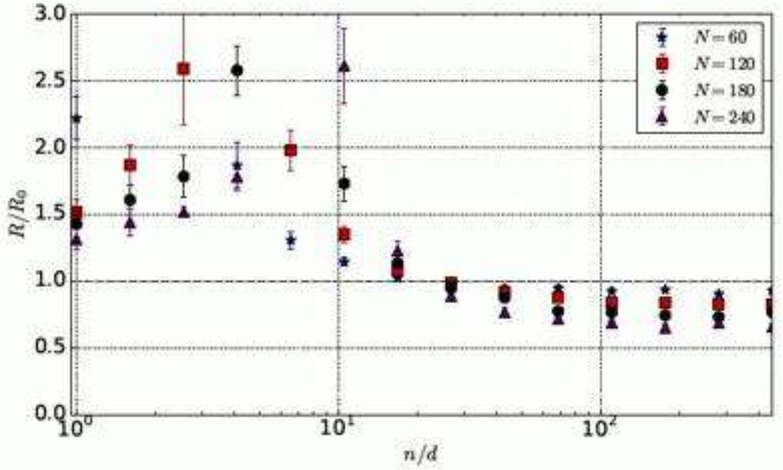
What can I fit with infinite data $n = \infty$?

An experiment

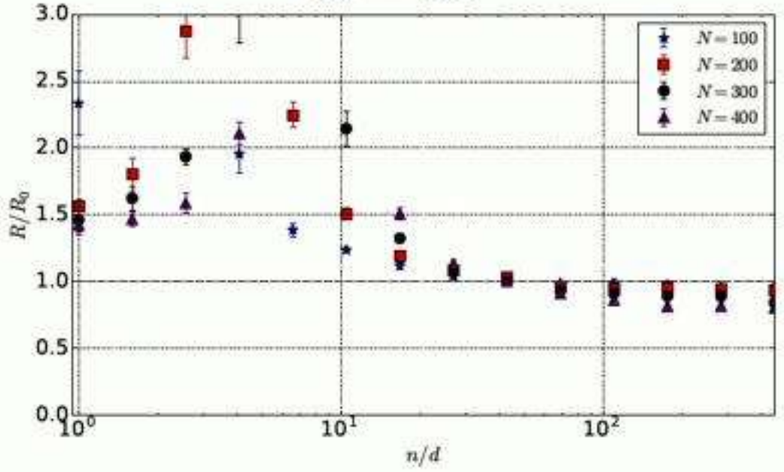
$d = 20$



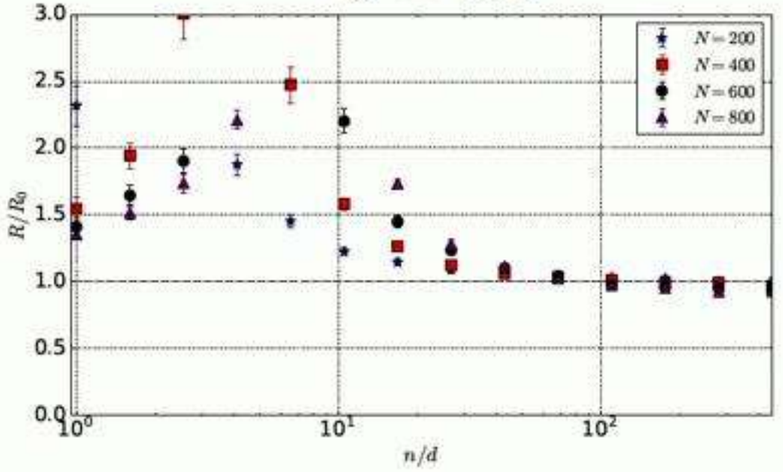
$d = 30$



$d = 50$

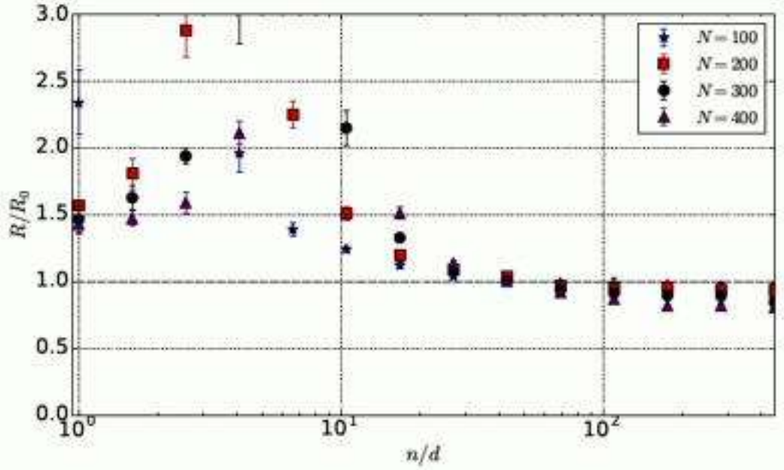


$d = 100$

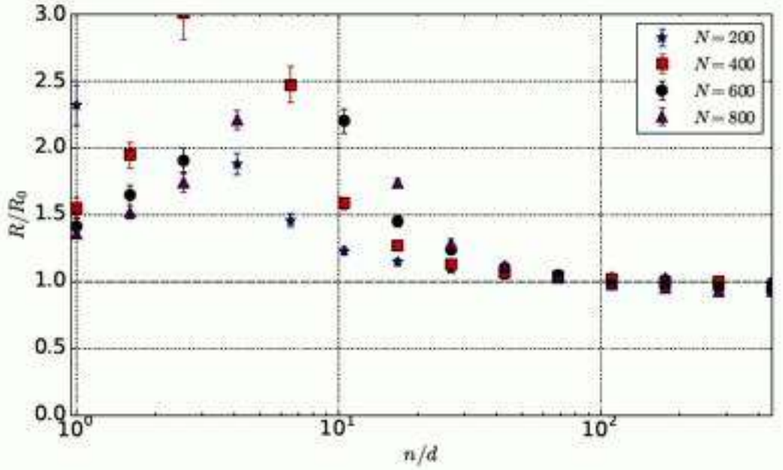


An experiment

$d = 50$



$d = 100$



- RF no better than trivial predictor $f_0(x) = 0$
- A 'difficult' target $f_\star$?

$$f_{\star,2}(x) = \sum_{i=1}^{d/2} x_i^2 - \sum_{i=d/2+1}^{d} x_i^2$$

A theorem ($\mathsf{P}_{\le \ell}$ projects on low degree)

Theorem (Ghorbani, Mei, Misiakiewicz, M. 2019)

Assume $N \le d^{\ell+1-\delta}$ for a fixed integer ℓ and some $\delta > 0$.
Then, with high probability:

$$R_{\mathsf{RF}}(f_\star) = R_{\mathsf{RF}}(\mathsf{P}_{\le \ell} f_\star) + \|\mathsf{P}_{>\ell} f_\star\|_{L^2}^2 + o(\|f_\star\|_{L^2}^2)\,.$$

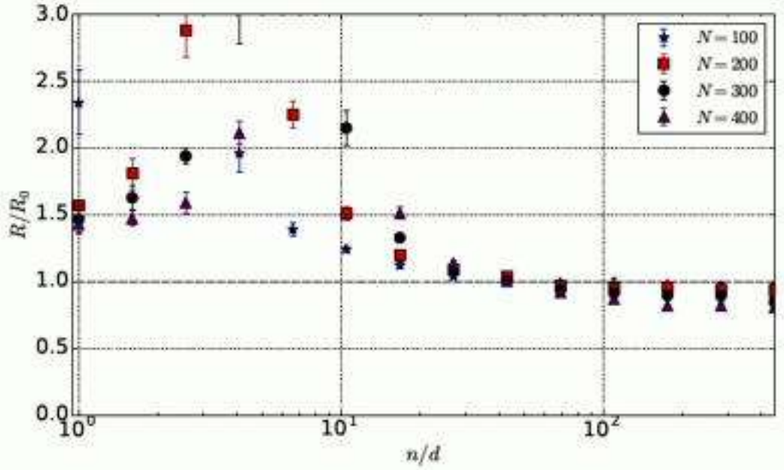
Example: If # neurons $N \ll d^2$, then

$$\begin{aligned} R_{\mathsf{RF}}(f_\star, \boldsymbol{W}) &\gtrsim R_{\text{lin. reg.}}(f_\star) \\ &\equiv \min_{b,\boldsymbol{\beta}} \mathbb{E}\{(f_\star(\boldsymbol{x}) - b - \langle \boldsymbol{\beta}, \boldsymbol{x}\rangle)\}\,. \end{aligned}$$

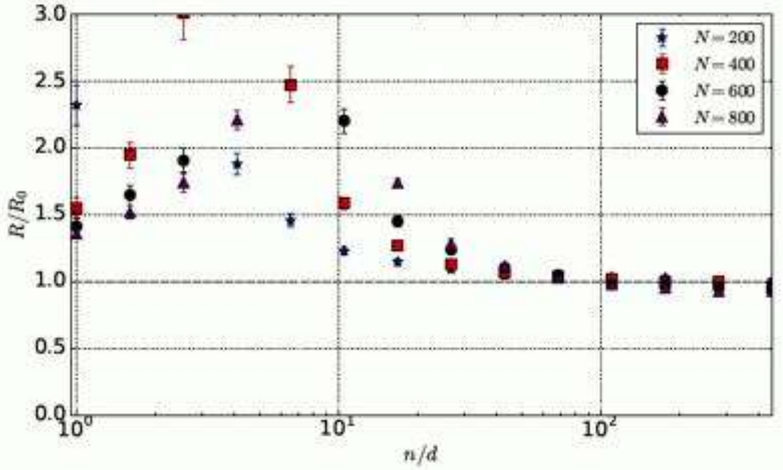
[Bach 2017;...]

An experiment

$d = 50$



$d = 100$



- RF no better than trivial predictor $f_0(x) = 0$
- A 'difficult' target $f_\star$?

$$f_{\star,2}(x) = \sum_{i=1}^{d/2} x_i^2 - \sum_{i=d/2+1}^{d} x_i^2$$

A theorem ($\mathsf{P}_{\le \ell}$ projects on low degree)

Theorem (Ghorbani, Mei, Misiakiewicz, M. 2019)

Assume $N \le d^{\ell+1-\delta}$ for a fixed integer ℓ and some $\delta > 0$.
Then, with high probability:

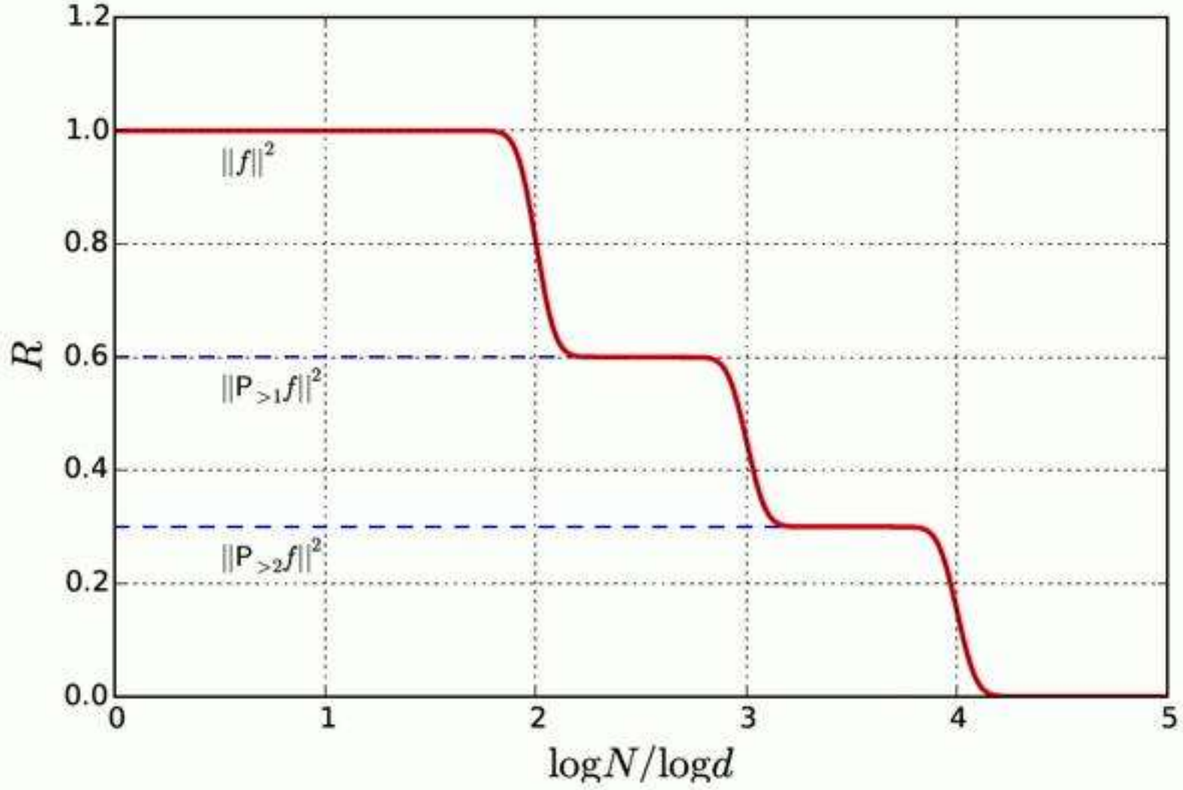
$$R_{\mathsf{RF}}(f_\star) = R_{\mathsf{RF}}(\mathsf{P}_{\le \ell} f_\star) + \|\mathsf{P}_{>\ell} f_\star\|_{L^2}^2 + o(\|f_\star\|_{L^2}^2)\,.$$

Example: If # neurons $N \ll d^2$, then

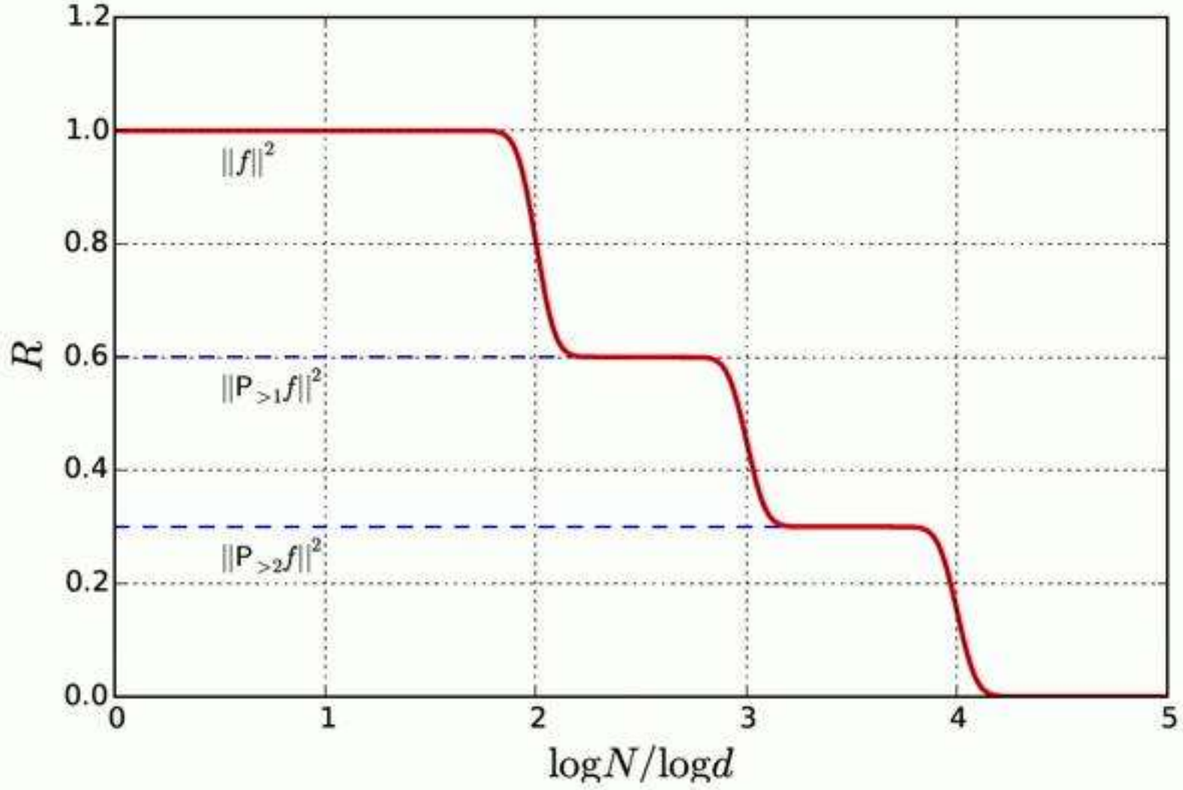
$$\begin{aligned} R_{\mathsf{RF}}(f_\star, \boldsymbol{W}) &\gtrsim R_{\text{lin. reg.}}(f_\star) \\ &\equiv \min_{b,\boldsymbol{\beta}} \mathbb{E}\{(f_\star(\boldsymbol{x}) - b - \langle \boldsymbol{\beta}, \boldsymbol{x}\rangle)\}\,. \end{aligned}$$

[Bach 2017;...]

Similar theorem for NT



What happens at finite n?



What happens at finite n?

Nonlinear features: Generalization curve

Setting

- Random features ridge regression (as above.)
- Proportional regime $n, N, d \to \infty$,

$$\frac{N}{d} \to \psi_1, \quad \frac{n}{d} \to \psi_2 \, .$$

- $f_\star(\boldsymbol{x}) = \langle \boldsymbol{\beta}_0, \boldsymbol{x} \rangle$ (Higher order terms $\approx$ noise)

Precise asymptotics

Theorem

Assume $f_\star(\boldsymbol{x}) = \langle \boldsymbol{\beta}_0, \boldsymbol{x}\rangle$ *and define (for* $G \sim \mathsf{N}(0,1)$*)*

$$b_\star^2 = \mathbb{E}[\sigma(G)^2] - \mathbb{E}[\sigma(G)],^2 - b_1^2, \qquad \zeta \equiv \frac{b_1^2}{b_\star^2}\,.$$

Then, for any $\lambda > 0$*, we have*

$$R_{\mathsf{RF}}(\hat{f}_\lambda, f_\star) = \|\boldsymbol{\beta}_0\|_2^2 \mathscr{B}(\zeta, \psi_1, \psi_2, \lambda/\overline{b}_\star^2) + \tau^2 \mathscr{V}(\zeta, \psi_1, \psi_2, \lambda/\overline{b}_\star^2) + o_d(1)\,,$$

where $\mathscr{B}(\zeta, \psi_1, \psi_2, \overline{\lambda})$*,* $\mathscr{V}(\zeta, \psi_1, \psi_2, \overline{\lambda})$ *are explicitly given below.*

Explicit formulae

Let $(\nu_1(\xi), \nu_2(\xi))$ be the unique solution of

$$\nu_1 = \psi_1\Big(-\xi - \nu_2 - \frac{\zeta^2\nu_2}{1-\zeta^2\nu_1\nu_2}\Big)^{-1},$$
$$\nu_2 = \psi_2\Big(-\xi - \nu_1 - \frac{\zeta^2\nu_1}{1-\zeta^2\nu_1\nu_2}\Big)^{-1};$$

Let

$$\chi \equiv \nu_1(\boldsymbol{i}(\psi_1\psi_2\overline{\lambda})^{1/2})\cdot\nu_2(\boldsymbol{i}(\psi_1\psi_2\overline{\lambda})^{1/2}),$$

and

$$\begin{aligned}
\mathscr{E}_0(\zeta,\psi_1,\psi_2,\overline{\lambda}) \equiv\ & -\chi^5\zeta^6 + 3\chi^4\zeta^4 + (\psi_1\psi_2 - \psi_2 - \psi_1 + 1)\chi^3\zeta^6 - 2\chi^3\zeta^4 - 3\chi^3\zeta^2 \\
& + (\psi_1 + \psi_2 - 3\psi_1\psi_2 + 1)\chi^2\zeta^4 + 2\chi^2\zeta^2 + \chi^2 + 3\psi_1\psi_2\chi\zeta^2 - \psi_1\psi_2\,, \\
\mathscr{E}_1(\zeta,\psi_1,\psi_2,\overline{\lambda}) \equiv\ & \psi_2\chi^3\zeta^4 - \psi_2\chi^2\zeta^2 + \psi_1\psi_2\chi\zeta^2 - \psi_1\psi_2\,, \\
\mathscr{E}_2(\zeta,\psi_1,\psi_2,\overline{\lambda}) \equiv\ & \chi^5\zeta^6 - 3\chi^4\zeta^4 + (\psi_1 - 1)\chi^3\zeta^6 + 2\chi^3\zeta^4 + 3\chi^3\zeta^2 + (-\psi_1 - 1)\chi^2\zeta^4 - 2\chi^2\zeta^2 - \chi^2\,.
\end{aligned}$$

We then have

$$\mathscr{B}(\zeta,\psi_1,\psi_2,\overline{\lambda}) \equiv \frac{\mathscr{E}_1(\zeta,\psi_1,\psi_2,\overline{\lambda})}{\mathscr{E}_0(\zeta,\psi_1,\psi_2,\overline{\lambda})}\,, \qquad \mathscr{V}(\zeta,\psi_1,\psi_2,\overline{\lambda}) \equiv \frac{\mathscr{E}_2(\zeta,\psi_1,\psi_2,\overline{\lambda})}{\mathscr{E}_0(\zeta,\psi_1,\psi_2,\overline{\lambda})}\,.$$

Random matrix theory for kernel inner product random matrices

[Cheng, Singer, 2013; Do, Vu 2013; Fan, M. 2019; . . .]

Explicit formulae

Let $(\nu_1(\xi), \nu_2(\xi))$ be the unique solution of

$$\nu_1 = \psi_1\Big(-\xi - \nu_2 - \frac{\zeta^2\nu_2}{1-\zeta^2\nu_1\nu_2}\Big)^{-1},$$
$$\nu_2 = \psi_2\Big(-\xi - \nu_1 - \frac{\zeta^2\nu_1}{1-\zeta^2\nu_1\nu_2}\Big)^{-1};$$

Let

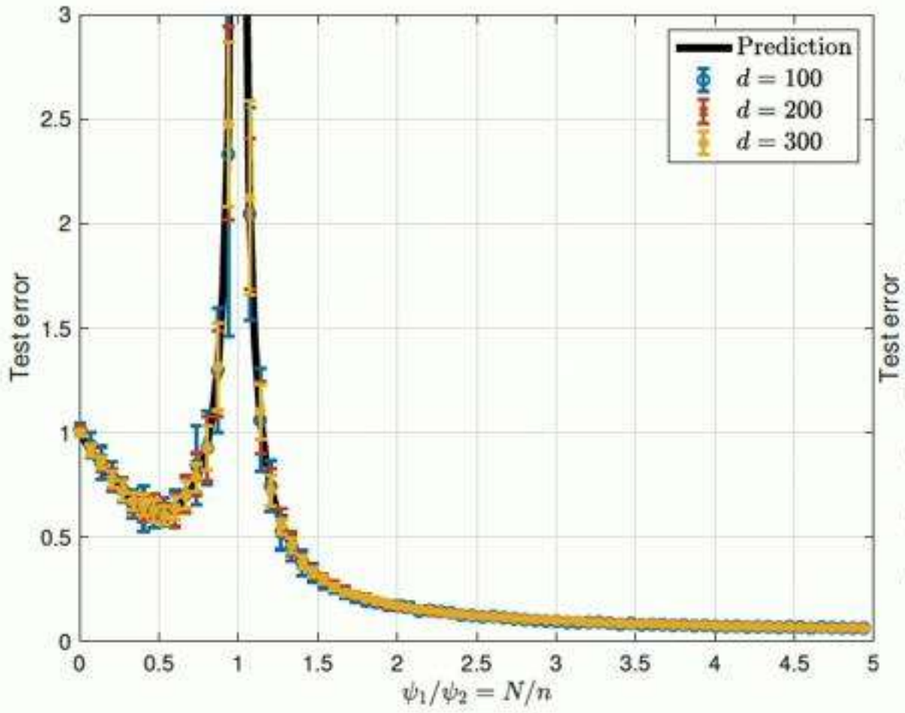
$$\chi \equiv \nu_1(\boldsymbol{i}(\psi_1\psi_2\overline{\lambda})^{1/2})\cdot\nu_2(\boldsymbol{i}(\psi_1\psi_2\overline{\lambda})^{1/2}),$$

and

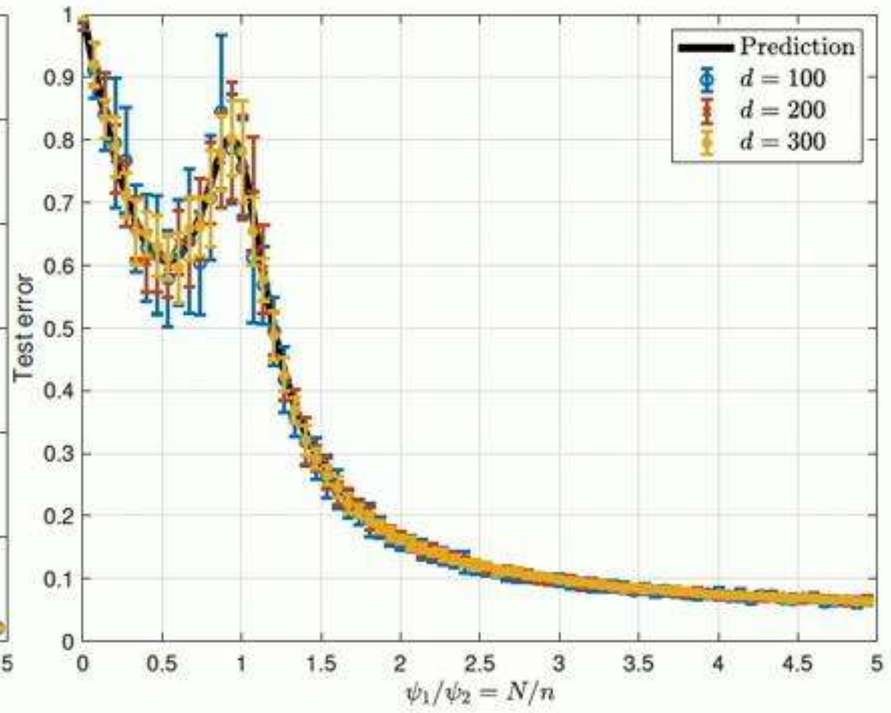
$$\begin{aligned}
\mathscr{E}_0(\zeta,\psi_1,\psi_2,\overline{\lambda}) \equiv\ & -\chi^5\zeta^6 + 3\chi^4\zeta^4 + (\psi_1\psi_2 - \psi_2 - \psi_1 + 1)\chi^3\zeta^6 - 2\chi^3\zeta^4 - 3\chi^3\zeta^2 \\
& + (\psi_1 + \psi_2 - 3\psi_1\psi_2 + 1)\chi^2\zeta^4 + 2\chi^2\zeta^2 + \chi^2 + 3\psi_1\psi_2\chi\zeta^2 - \psi_1\psi_2\,, \\
\mathscr{E}_1(\zeta,\psi_1,\psi_2,\overline{\lambda}) \equiv\ & \psi_2\chi^3\zeta^4 - \psi_2\chi^2\zeta^2 + \psi_1\psi_2\chi\zeta^2 - \psi_1\psi_2\,, \\
\mathscr{E}_2(\zeta,\psi_1,\psi_2,\overline{\lambda}) \equiv\ & \chi^5\zeta^6 - 3\chi^4\zeta^4 + (\psi_1 - 1)\chi^3\zeta^6 + 2\chi^3\zeta^4 + 3\chi^3\zeta^2 + (-\psi_1 - 1)\chi^2\zeta^4 - 2\chi^2\zeta^2 - \chi^2\,.
\end{aligned}$$

We then have

$$\mathscr{B}(\zeta,\psi_1,\psi_2,\overline{\lambda}) \equiv \frac{\mathscr{E}_1(\zeta,\psi_1,\psi_2,\overline{\lambda})}{\mathscr{E}_0(\zeta,\psi_1,\psi_2,\overline{\lambda})}\,, \qquad \mathscr{V}(\zeta,\psi_1,\psi_2,\overline{\lambda}) \equiv \frac{\mathscr{E}_2(\zeta,\psi_1,\psi_2,\overline{\lambda})}{\mathscr{E}_0(\zeta,\psi_1,\psi_2,\overline{\lambda})}\,.$$

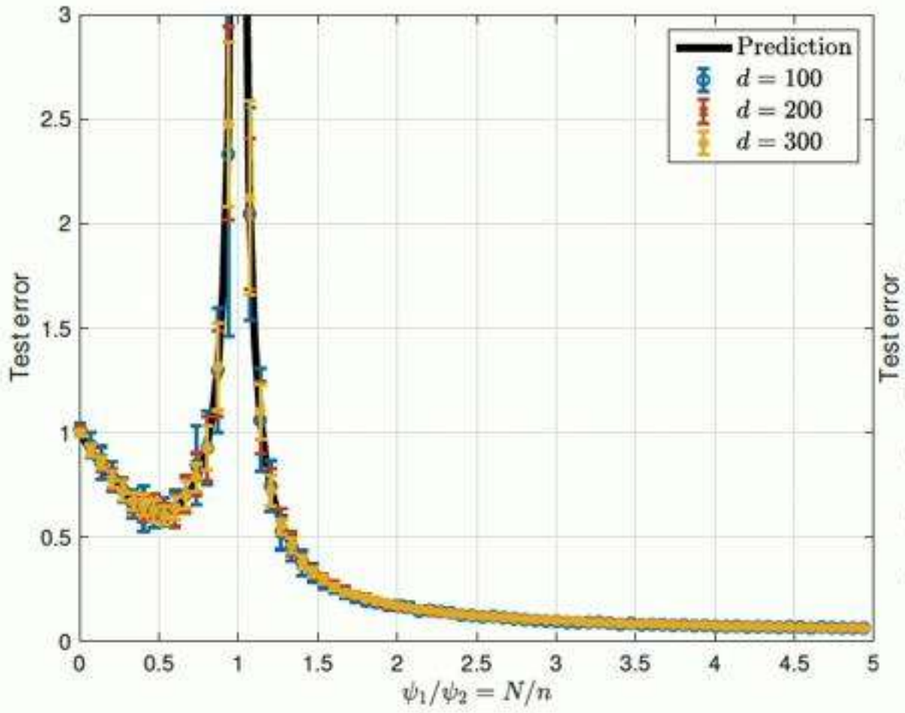


$\lambda = 0+$

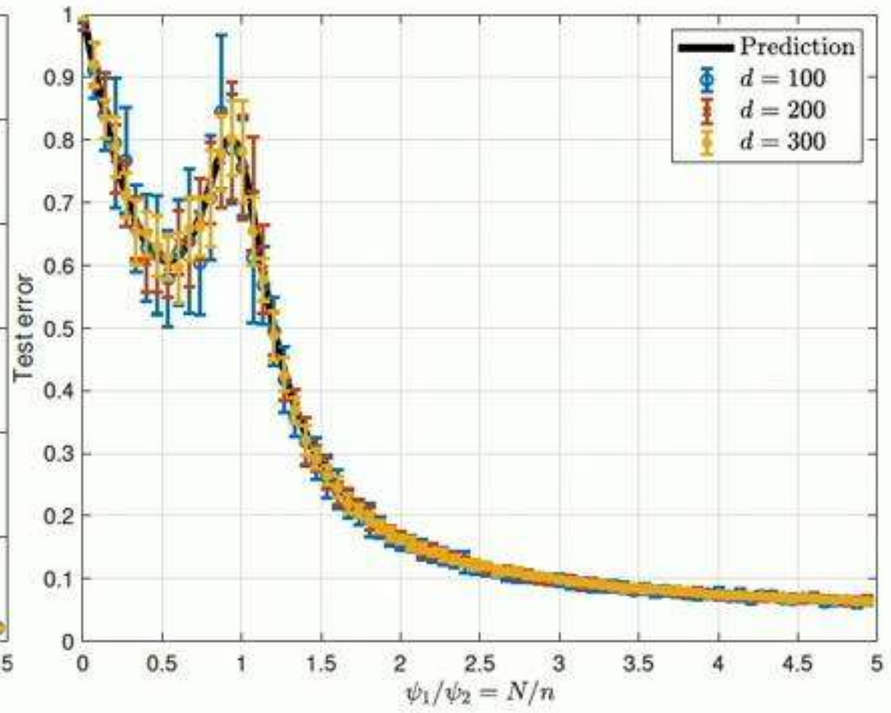


$\lambda > 0$

- ✓ Peak at the interpolation threshold
- ✓ Global minimum in the overparametrized regime
- ✓ Monotone decreasing in the overparametrized regime



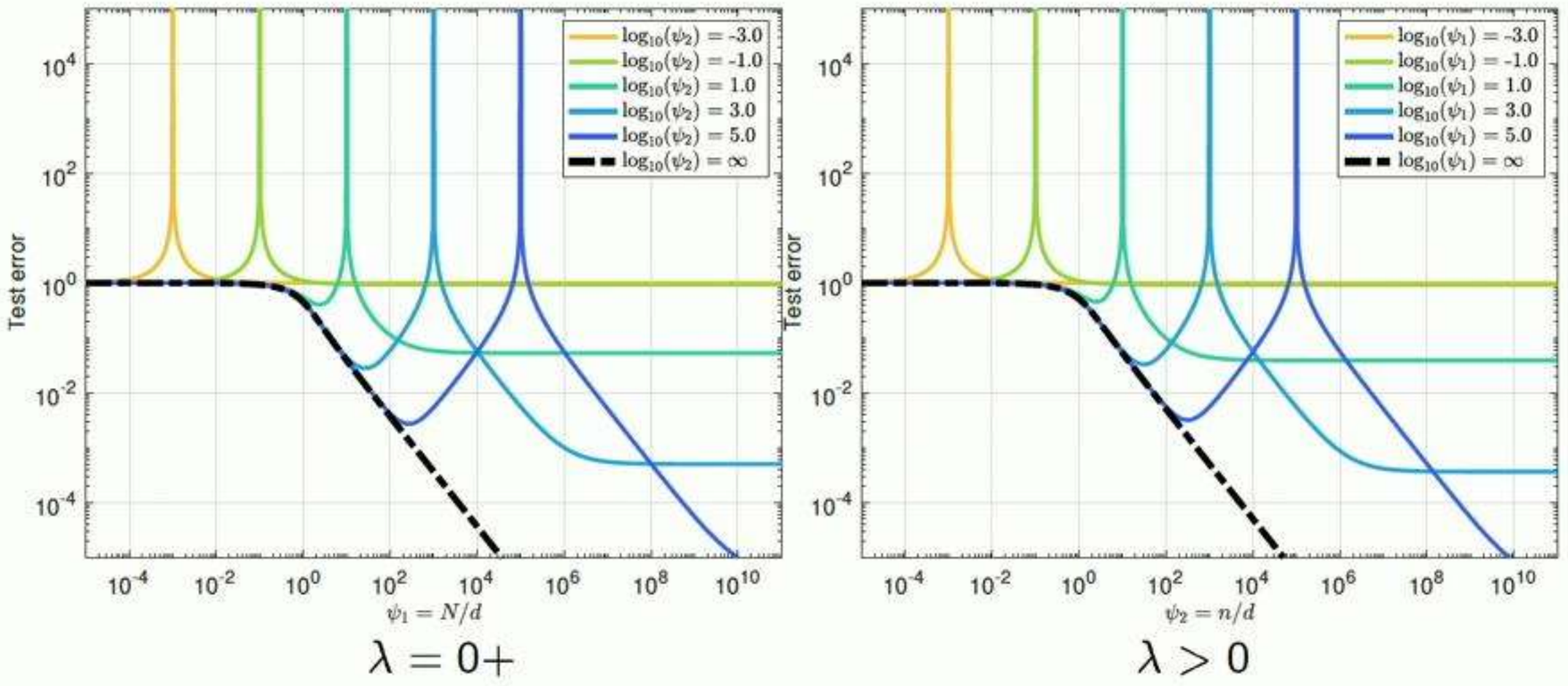
$\lambda = 0+$



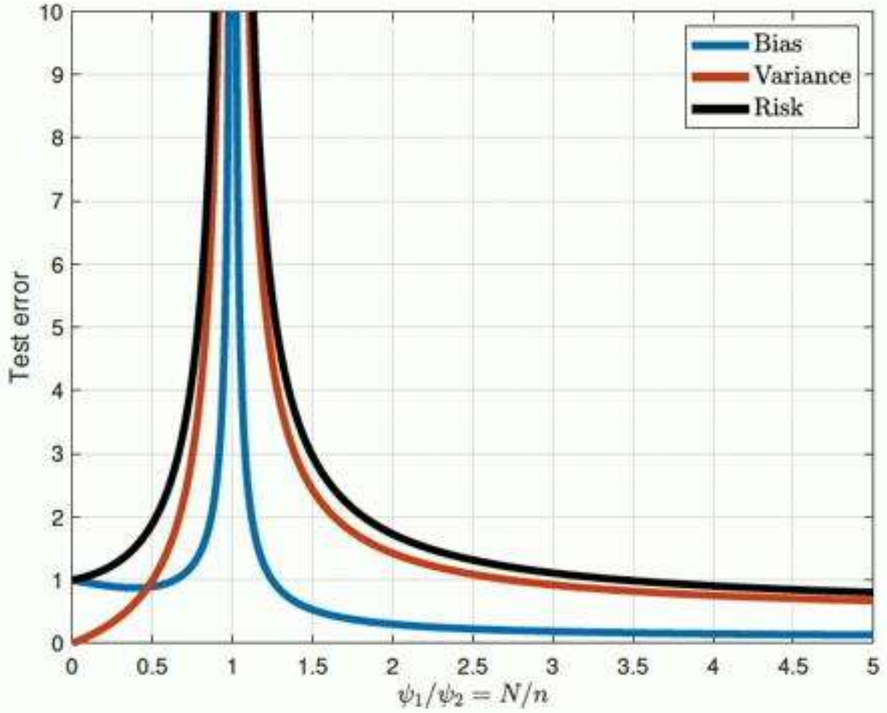
$\lambda > 0$

- ✓ Peak at the interpolation threshold
- ✓ Global minimum in the overparametrized regime
- ✓ Monotone decreasing in the overparametrized regime

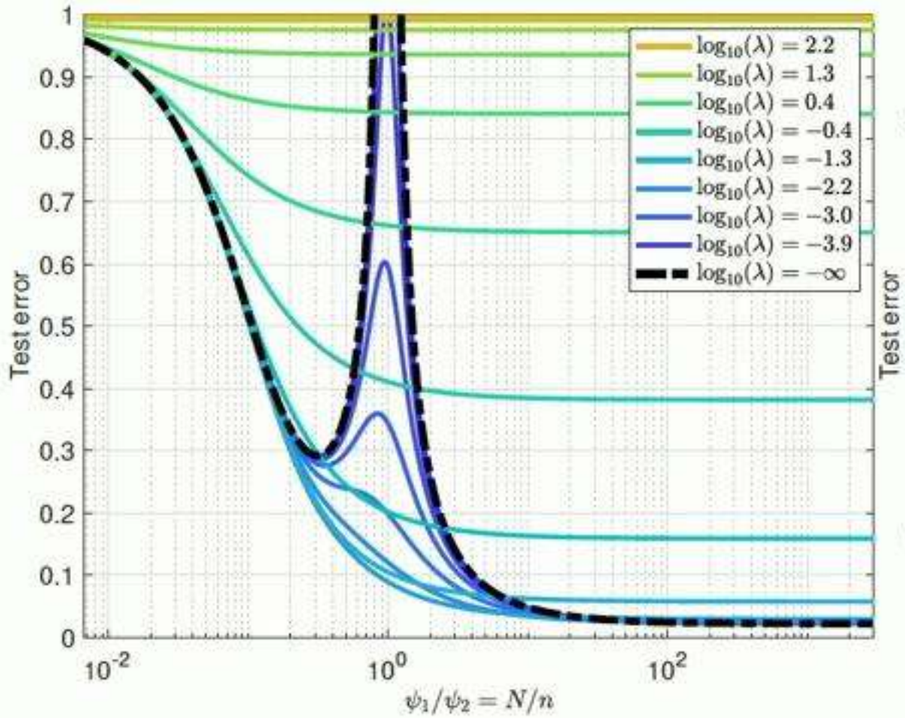
Insigths



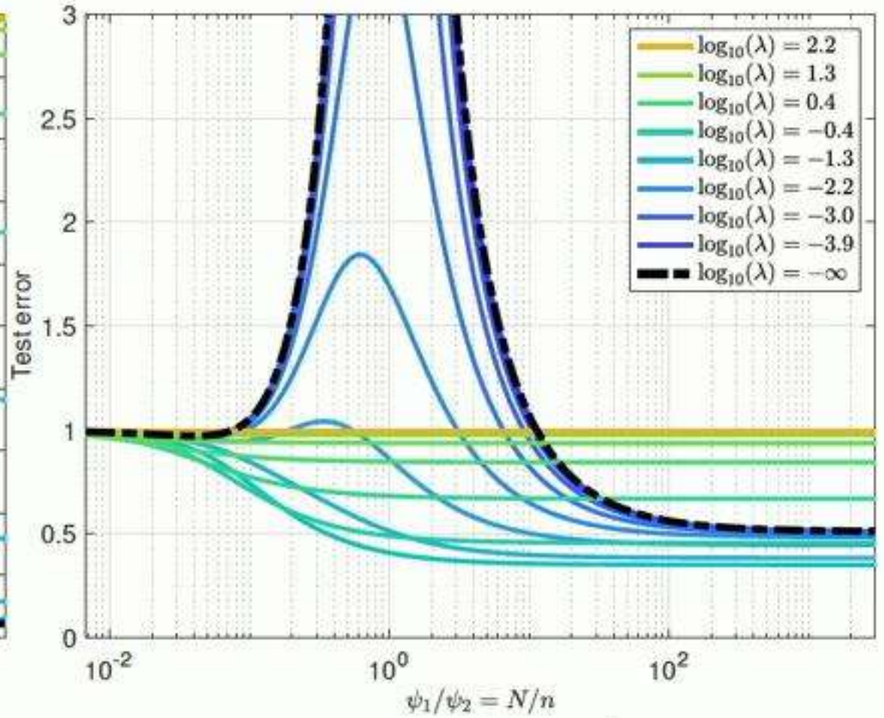
- Singularity at the interpolation threshold
- Minimum risk at extreme overparametrization $N/n \to \infty$.



- Both bias and variance are singular

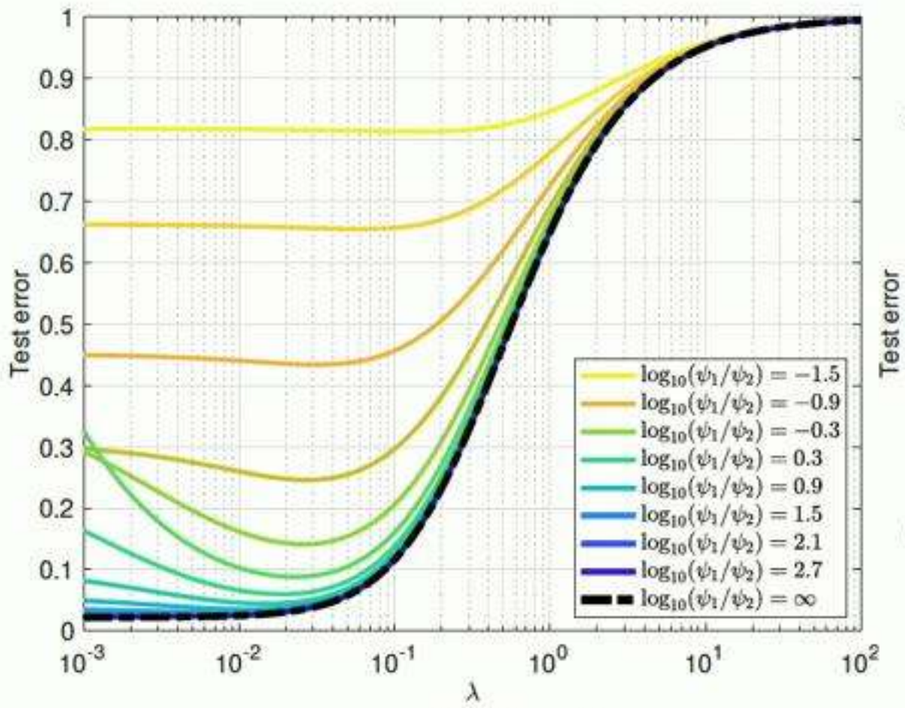


SNR= 5

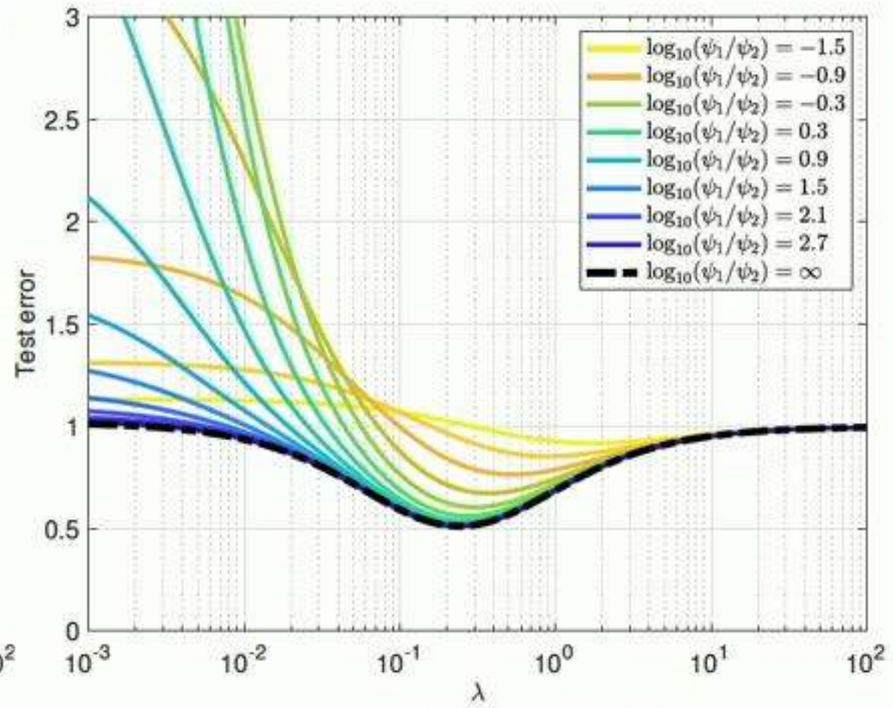


SNR= 1/5

- Same at $\lambda > 0$ fixed: Minimum at $N/n \to \infty$.

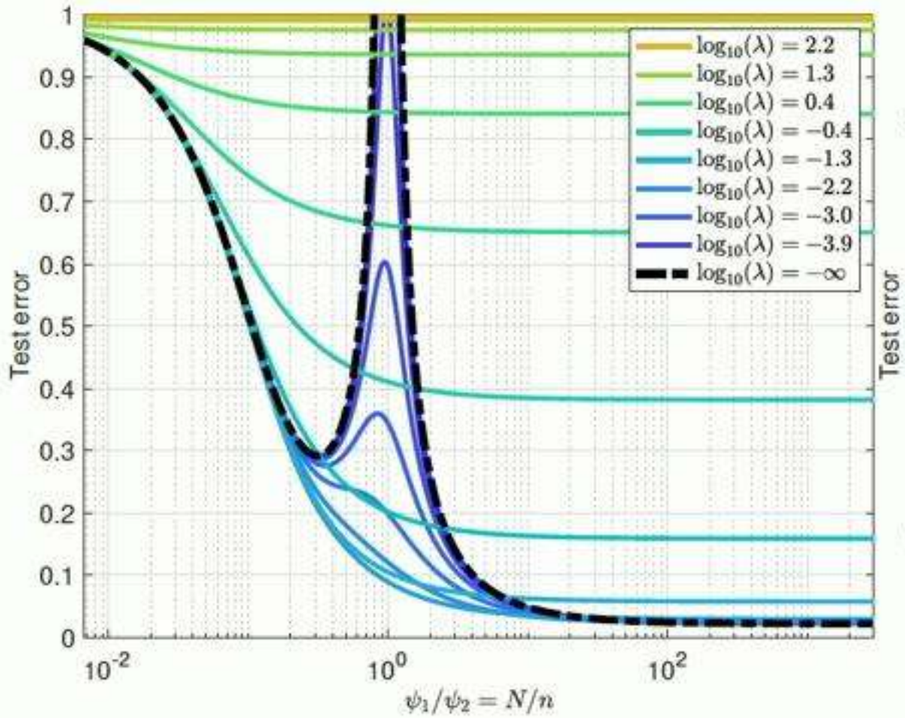


SNR= 5

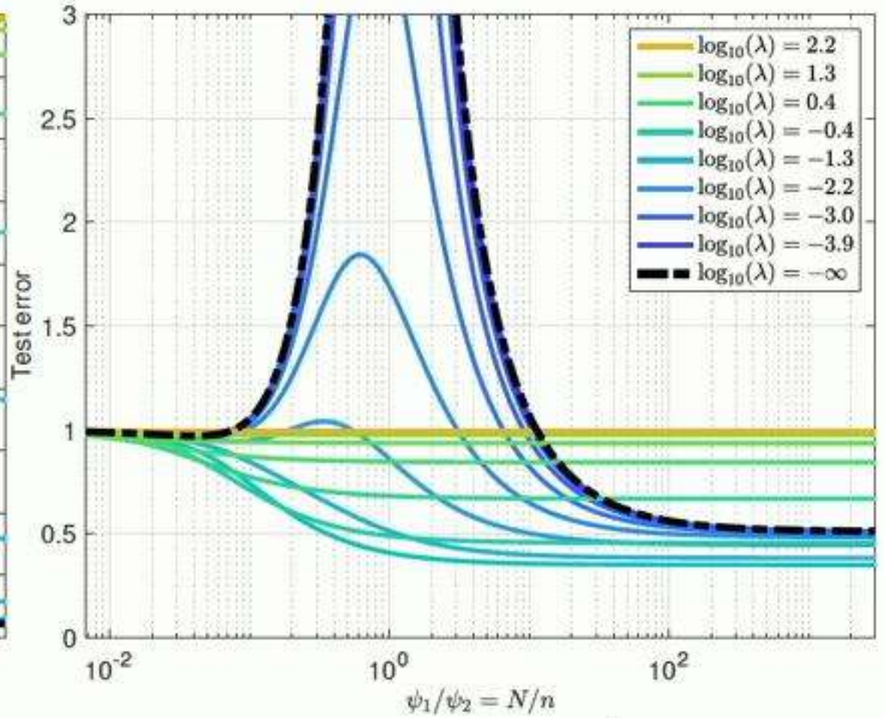


SNR= 1/10

- High SNR: Minimum at $\lambda = 0+$.
- Low SNR: Minimum at $\lambda > 0$.

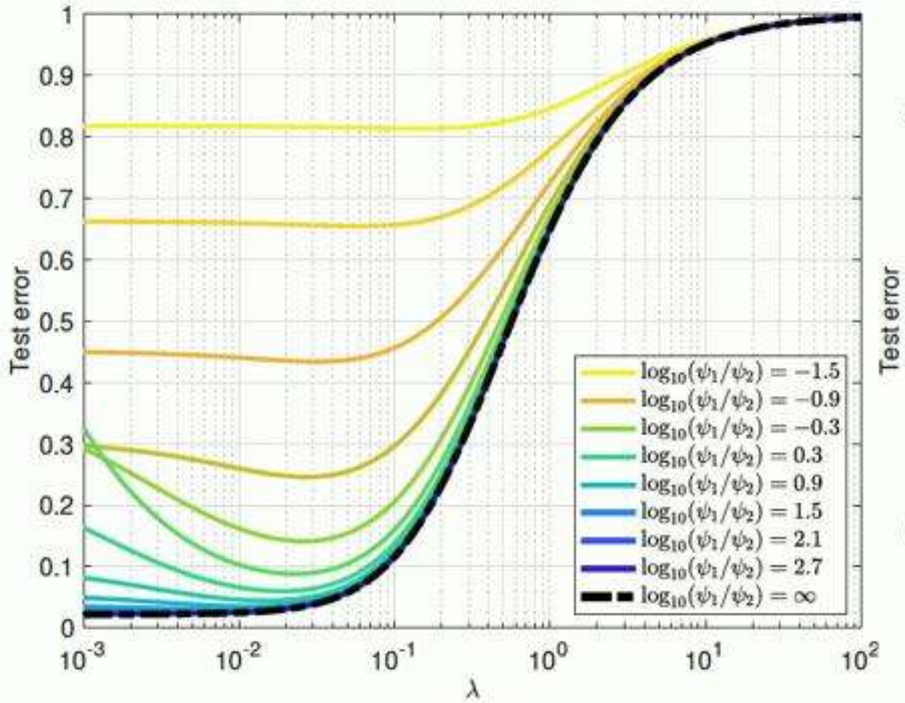


SNR= 5

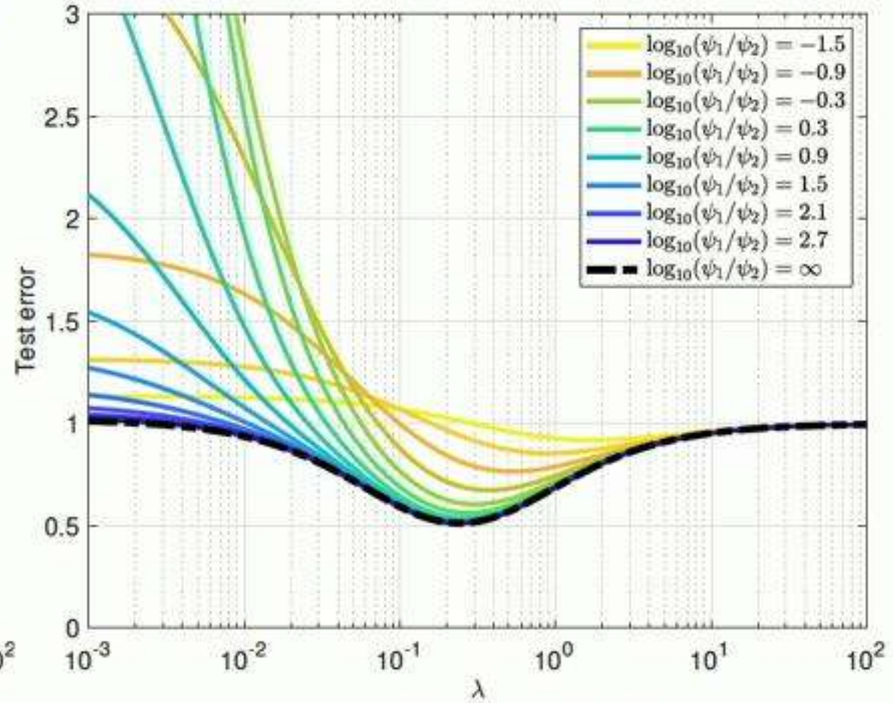


SNR= 1/5

- Same at $\lambda > 0$ fixed: Minimum at $N/n \to \infty$.



SNR= 5



SNR= 1/10

- High SNR: Minimum at $\lambda = 0+$.
- Low SNR: Minimum at $\lambda > 0$.

Conclusion

Overparametrization $\not\Rightarrow$ Poor generalization

- Mechanism #1: Early stopping/mean field

  [Mei, M, Nguyen 2018; Chizat, Bach 2018]

- Mechanism #2: High dimension/interpolation

  [This talk]

- Open questions:
  - Fully trained networks?
  - When is extreme overparametrization optimal?

Thanks!

Overparametrization $\not\Rightarrow$ Poor generalization

- Mechanism #1: Early stopping/mean field

  [Mei, M, Nguyen 2018; Chizat, Bach 2018]

- Mechanism #2: High dimension/interpolation

  [This talk]

- Open questions:
  - Fully trained networks?
  - When is extreme overparametrization optimal?

Thanks!

Explicit formulae

Let $(\nu_1(\xi), \nu_2(\xi))$ be the unique solution of

$$\nu_1 = \psi_1\Big(-\xi - \nu_2 - \frac{\zeta^2\nu_2}{1-\zeta^2\nu_1\nu_2}\Big)^{-1},$$
$$\nu_2 = \psi_2\Big(-\xi - \nu_1 - \frac{\zeta^2\nu_1}{1-\zeta^2\nu_1\nu_2}\Big)^{-1};$$

Let

$$\chi \equiv \nu_1(\boldsymbol{i}(\psi_1\psi_2\overline{\lambda})^{1/2})\cdot\nu_2(\boldsymbol{i}(\psi_1\psi_2\overline{\lambda})^{1/2}),$$

and

$$\begin{aligned}
\mathscr{E}_0(\zeta,\psi_1,\psi_2,\overline{\lambda}) \equiv\ & -\chi^5\zeta^6 + 3\chi^4\zeta^4 + (\psi_1\psi_2 - \psi_2 - \psi_1 + 1)\chi^3\zeta^6 - 2\chi^3\zeta^4 - 3\chi^3\zeta^2 \\
& + (\psi_1 + \psi_2 - 3\psi_1\psi_2 + 1)\chi^2\zeta^4 + 2\chi^2\zeta^2 + \chi^2 + 3\psi_1\psi_2\chi\zeta^2 - \psi_1\psi_2\,, \\
\mathscr{E}_1(\zeta,\psi_1,\psi_2,\overline{\lambda}) \equiv\ & \psi_2\chi^3\zeta^4 - \psi_2\chi^2\zeta^2 + \psi_1\psi_2\chi\zeta^2 - \psi_1\psi_2\,, \\
\mathscr{E}_2(\zeta,\psi_1,\psi_2,\overline{\lambda}) \equiv\ & \chi^5\zeta^6 - 3\chi^4\zeta^4 + (\psi_1 - 1)\chi^3\zeta^6 + 2\chi^3\zeta^4 + 3\chi^3\zeta^2 + (-\psi_1 - 1)\chi^2\zeta^4 - 2\chi^2\zeta^2 - \chi^2\,.
\end{aligned}$$

We then have

$$\mathscr{B}(\zeta,\psi_1,\psi_2,\overline{\lambda}) \equiv \frac{\mathscr{E}_1(\zeta,\psi_1,\psi_2,\overline{\lambda})}{\mathscr{E}_0(\zeta,\psi_1,\psi_2,\overline{\lambda})}\,, \qquad \mathscr{V}(\zeta,\psi_1,\psi_2,\overline{\lambda}) \equiv \frac{\mathscr{E}_2(\zeta,\psi_1,\psi_2,\overline{\lambda})}{\mathscr{E}_0(\zeta,\psi_1,\psi_2,\overline{\lambda})}\,.$$

Precise asymptotics

Theorem

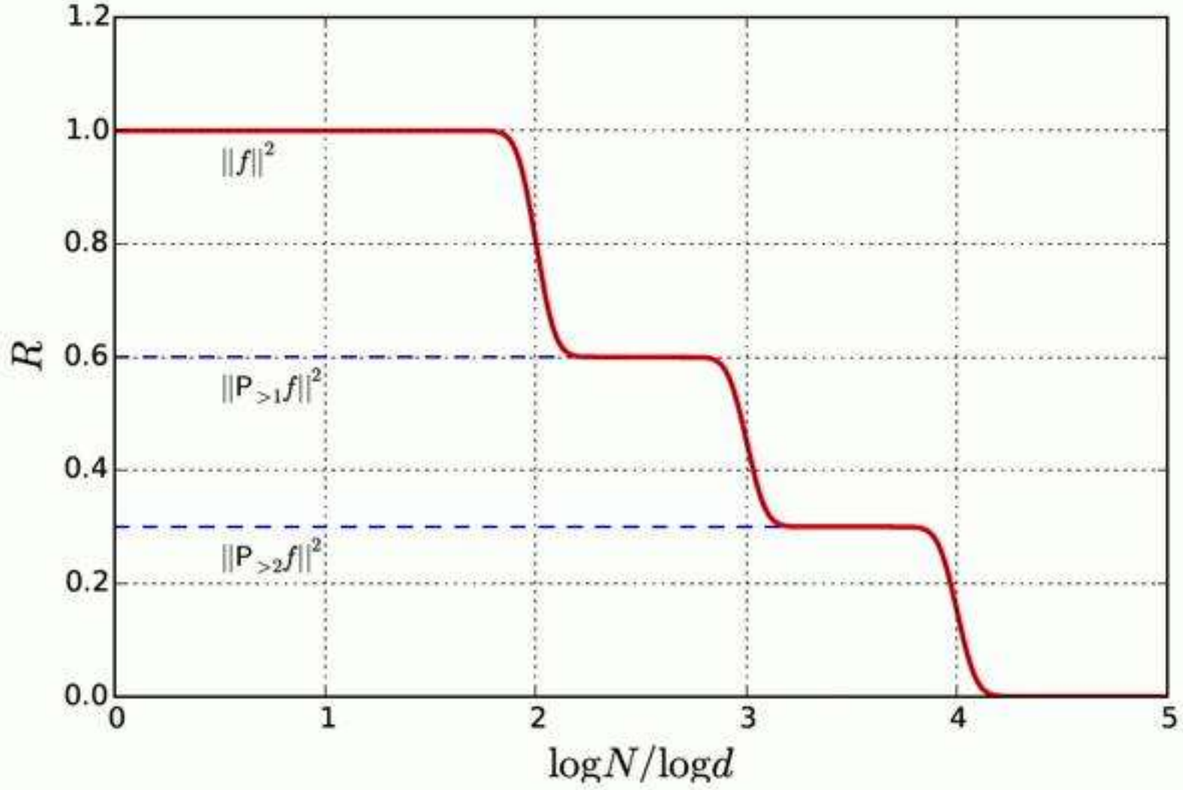
Assume $f_\star(\boldsymbol{x}) = \langle \boldsymbol{\beta}_0, \boldsymbol{x}\rangle$ *and define (for* $G \sim \mathsf{N}(0,1)$*)*

$$b_\star^2 = \mathbb{E}[\sigma(G)^2] - \mathbb{E}[\sigma(G)],^2 - b_1^2, \qquad \zeta \equiv \frac{b_1^2}{b_\star^2}\,.$$

Then, for any $\lambda > 0$*, we have*

$$R_{\mathsf{RF}}(\hat{f}_\lambda, f_\star) = \|\boldsymbol{\beta}_0\|_2^2 \mathscr{B}(\zeta, \psi_1, \psi_2, \lambda/\overline{b}_\star^2) + \tau^2 \mathscr{V}(\zeta, \psi_1, \psi_2, \lambda/\overline{b}_\star^2) + o_d(1)\,,$$

where $\mathscr{B}(\zeta, \psi_1, \psi_2, \overline{\lambda})$*,* $\mathscr{V}(\zeta, \psi_1, \psi_2, \overline{\lambda})$ *are explicitly given below.*



What happens at finite n?