Double Descent

Reconciling modern machine learning practice and the bias-variance trade-off [1]

Surprises in High-Dimensional Ridgeless Least Squares Interpolation [2]

[1]Mikhail Belkin, Daniel Hsu, Siyuan Ma, Soumik Mandal
[2]Trevor Hastie, Andrea Montanari, Saharon Rosset, Ryan J. Tibshirani
https://simons.berkeley.edu/sites/default/files/docs/14172/slides1.pdf
http://www.stat.cmu.edu/~ryantibs/talks/ls-inter-2019.pdf

STAT 991
Junhui Cai

Outline

Outline

Outline

Then and now

- Given $\{(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}\}_{i=1}^n \overset{i.i.d.}{\sim} P$
- Find a predictor $h_n : \mathbb{R}^d \to \mathbb{R}$ such that

$$\arg\min_h \mathbb{E}_{(x^*,y^*)\sim P}\left[\ell(h(x^*), y^*)\right]$$

  for some loss function ℓ.
- Empirical risk minimization

$$\arg\min_{h\in\mathcal{H}} \hat{\mathbb{E}}_n\left[\ell(h(x_i), y_i)\right].$$

Problem	Solution: then	Solution: now
Representation	More parameters	More parameters!
Optimization	Convexify	More parameters!!
Generalization	Regularization	More parameters!!!

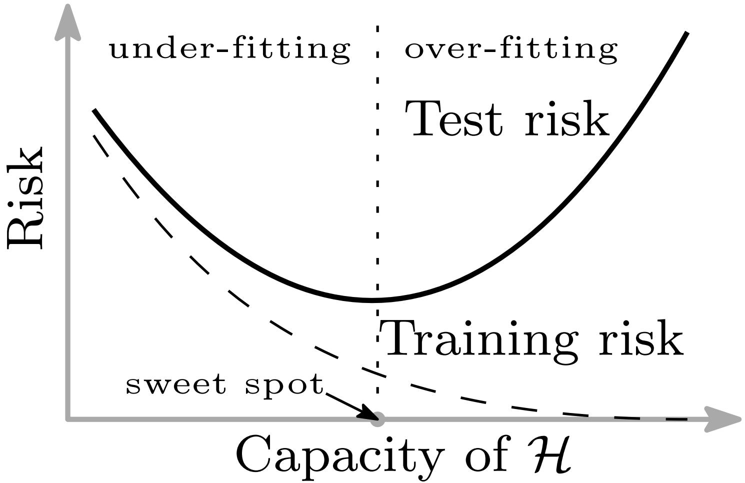
Understanding deep learning requires rethinking generalization (2016)

Zhang, Bengio, Hardt, Recht, Vinyals

Table 1: The training and test accuracy (in percentage) of various models on the CIFAR10 dataset. Performance with and without data augmentation and weight decay are compared. The results of fitting random labels are also included.

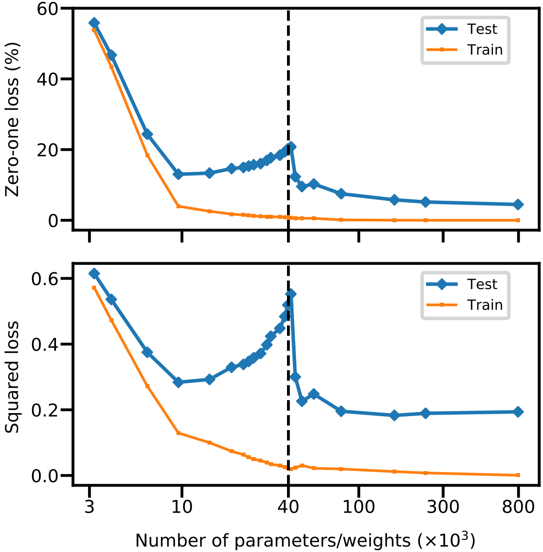
model	# params	random crop	weight decay	train accuracy	test accuracy
Inception	1,649,402	yes	yes	100.0	89.05
		yes	no	100.0	89.31
		no	yes	100.0	86.03
		no	no	100.0	85.75
(fitting random labels)		no	no	100.0	9.78
Inception w/o BatchNorm	1,649,402	no	yes	100.0	83.00
		no	no	100.0	82.00
(fitting random labels)		no	no	100.0	10.12
Alexnet	1,387,786	yes	yes	99.90	81.22
		yes	no	99.82	79.66
		no	yes	100.0	77.36
		no	no	100.0	76.07
(fitting random labels)		no	no	99.82	9.86
MLP 3x512	1,735,178	no	yes	100.0	53.35
		no	no	100.0	52.39
(fitting random labels)		no	no	100.0	10.48
MLP 1x512	1,209,866	no	yes	99.80	50.39
		no	no	100.0	50.51
(fitting random labels)		no	no	99.34	10.61

The bias-variance tradeoff (U-shaped)



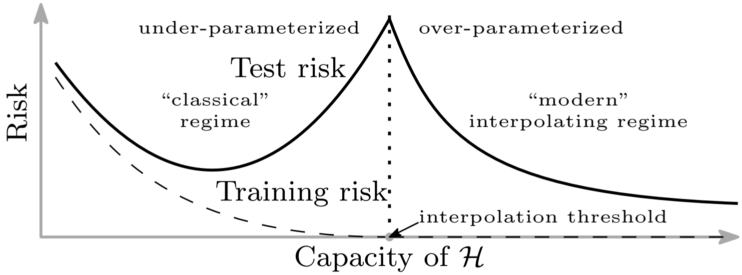
Interpolation, yet not overfitting

on a one-layer neural network

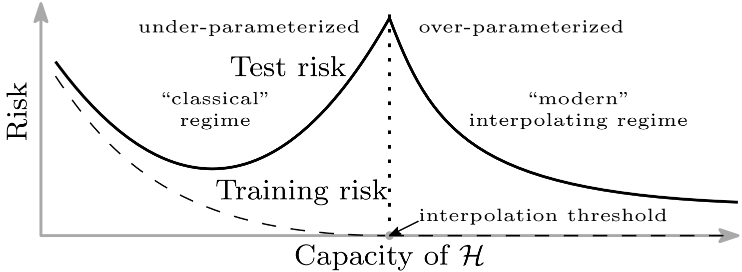


Double descent (W-shaped)

Belkin, Hsu, Ma, and Mandal (2018), "Reconciling modern machine learning and the bias-variance trade-off"

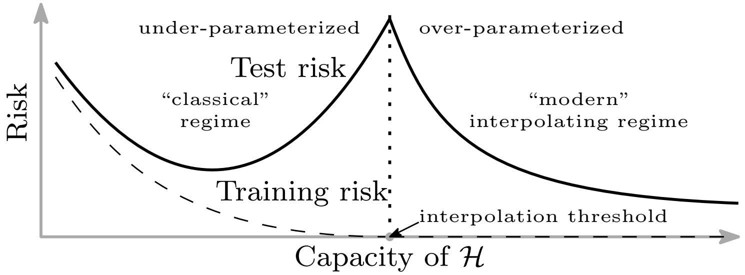


Double descent



- The double descent risk curve appears ubiquitously in a wide spectrum models and datasets.
- High risk at interpolation threshold, but decreasing risk as the function class capacity increasing (lower than the “classical” sweet spot?)
- Intuition: regularization restricts function classes.

Double descent



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

Outline

The double descent curve

- Then and now
- Empirical evidence

Ridgeless regression

- Underparametrized vs Overparametrized
- Misspecified model

Random Feature

Neural networks
Random Fourier features

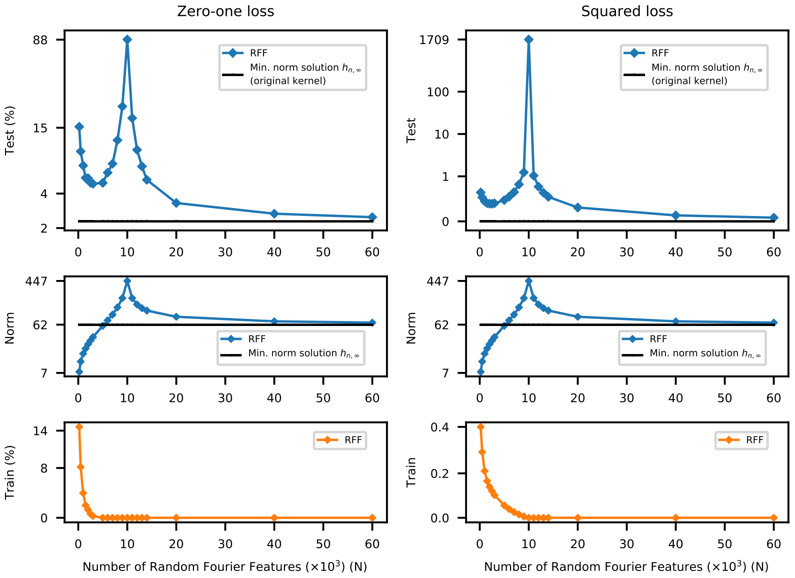
Random Fourier features family $\mathcal{H}_N$ with N parameters.

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

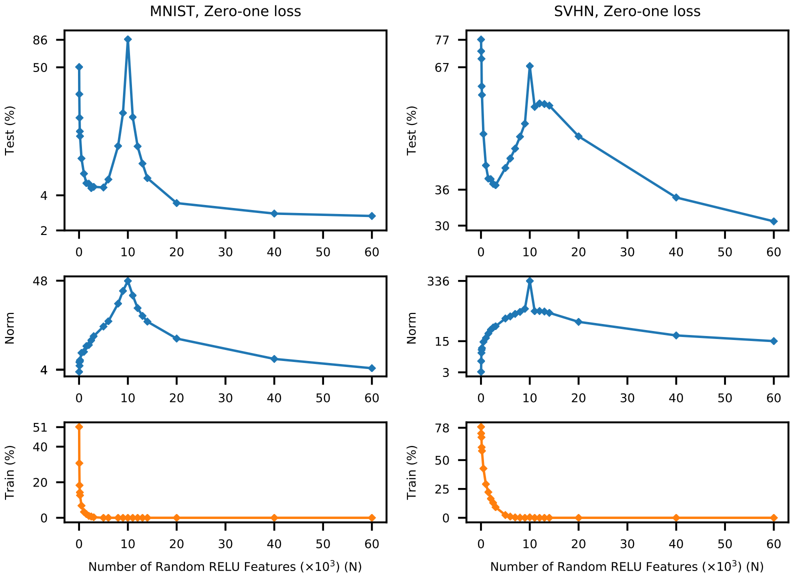
where the vectors $v_1, \ldots, v_N$ are sampled independently from the standard normal distribution in $\mathbb{R}^d$.

- Two-layer neural networks with fixed weights in the first layer
- Find $h_{n,N} = \arg\min_{h \in \mathcal{H}_N} \frac{1}{n} \sum_{i=1}^{n} (h(x_i) - y_i)^2$
- $N > n$ not unique, choose $h_{n,N}$ with smallest $\|a\|_2$

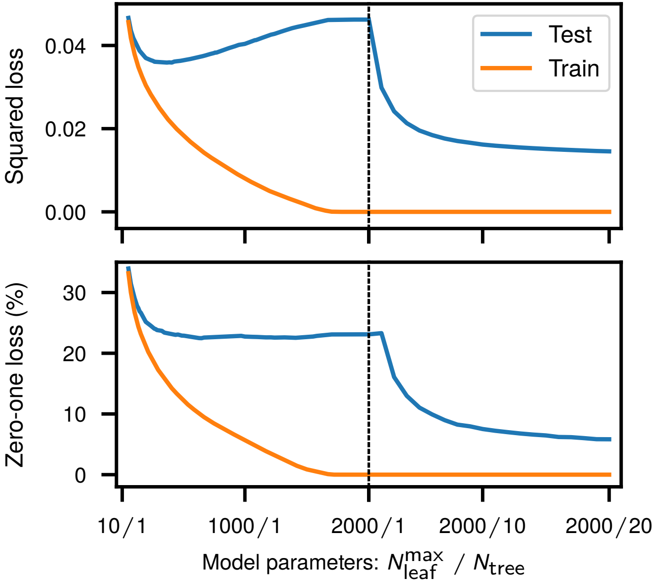
MNIST ($n = 10^4$) with RFF



MNIST ($n = 10^4$) with random ReLU



Tree and ensembles



Outline

Outline

Setup

Given i.i.d. $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \ldots, n$, where

$$(x_i, \epsilon_i) \sim P_x \times P_\epsilon \quad \text{and} \quad y_i = x_i^T \beta + \epsilon_i$$

and $\mathbb{E}(\epsilon_i) = 0$ and $\mathrm{Var}(\epsilon_i) = \sigma^2$

- **Linear model.** $x_i = \Sigma^{1/2} z_i$, where $z_i \in \mathbb{R}^p$ has i.i.d. entries with zero mean and unit variance and $\Sigma \in \mathbb{R}^{p \times p}$ deterministic and positive definite.
- **Nonlinear model.** $x_i = \varphi(W z_i)$, where $z_i \in \mathbb{R}^d$ has i.i.d. entries from $N(0, 1)$, $W \in \mathbb{R}^{p \times d}$ has i.i.d. entries from $N(0, 1/d)$, and φ is an activation function acting componentwise.

For nonlinear features, if $\mathbb{E}[y_i | z_i] = f(\theta; z_i)$,

$$\mathbb{E}[y_i | z_i] \approx \underbrace{\nabla_\theta f(z_i; \theta_0)^T}_{x_i} \underbrace{\theta - \theta_0}_{\beta}$$

Prediction risk

Consider a test point $x_0 \sim P_x$, independent of the training data. For an estimator $\hat{\beta}$ (a function of the training data X, y), we define its out-of-sample prediction risk (or simply, risk) as

$$R_X(\hat{\beta}; \beta) = \mathbb{E}\big[(x_0^T \hat{\beta} - x_0^T \beta)^2 \,|\, X\big] = \mathbb{E}\big[\|\hat{\beta} - \beta\|_\Sigma^2 \,|\, X\big],$$

where $\|x\|_\Sigma^2 = x^T \Sigma x$. Note that our definition of risk is conditional on X (as emphasized by our notation R_X). Note also that we have the bias-variance decomposition

$$R_X(\hat{\beta}; \beta) = \underbrace{\|\mathbb{E}(\hat{\beta}|X) - \beta\|_\Sigma^2}_{B_X(\hat{\beta};\beta)} + \underbrace{\mathrm{tr}[\mathrm{Cov}(\hat{\beta}|X)\Sigma]}_{V_X(\hat{\beta};\beta)}.$$

Ridgeless least square estimator

$$\hat{\beta} = (X^T X)^+ X^T y$$

$$\begin{aligned}
\hat{\beta}_\lambda &= \arg\min_{b \in \mathbb{R}^p} \left\{ \frac{1}{n} \|y - Xb\|_2^2 + \lambda \|b\|_2^2 \right\} \\
&= \lim_{\lambda \to 0^+} \beta_\lambda, \quad \text{where } \beta_\lambda = \arg\min_{b \in \mathbb{R}^p} \left\{ \frac{1}{n} \|y - Xb\|_2^2 + \lambda \|b\|_2^2 \right\} \\
&= \lim_{k \to \infty} \beta^{(k)}, \quad \text{where } \beta^{(k)} = \beta^{(k-1)} + t_k X^T (y - X\beta^{(k-1)})
\end{aligned}$$

The bias and variance of ridgeless estimator is

$$B_X(\hat{\beta}; \beta) = \beta^T \Pi \Sigma \Pi \beta \quad \text{and} \quad V_X(\hat{\beta}; \beta) = \frac{\sigma^2}{n} \mathrm{tr}(\hat{\Sigma}^+ \Sigma),$$

where $\hat{\Sigma} = X^T X / n$ is the (uncentered) sample covariance of X, and $\Pi = I - \hat{\Sigma}^+ \hat{\Sigma}$ is the projection onto the null space of X.

Classical regime ($p < n$)

- $x = \Sigma^{1/2} z$, where $z \in \mathbb{R}^p$ is a random vector with i.i.d. entries with $\mathbb{E}[z_{ij}] = 0$, $\mathbb{E}[z_{ij}^2] = 1$ and $\mathbb{E}[z_{ij}^{2+\delta}] < \infty$
- $\lambda_{\min}(\Sigma) \geq c > 0$, for all n, p and a constant c
- $F_\Sigma = \frac{1}{p} \sum_{i=1}^{p} \delta_{\lambda_i(\Sigma)}$ converges weakly to a measure H.

Theorem (Girko 1990s, Verdu and Tse 1990s)

As $n, p \to \infty$, such that $p/n \to \gamma < 1$, the risk of the least squares estimator satisfies, almost surely,

$$R_X(\hat{\beta}; \beta) \to \sigma^2 \frac{\gamma}{1 - \gamma}.$$

$$R_X(\hat{\beta}; \beta) = \sigma^2 \mathrm{tr}(\hat{\Sigma}^{-1} \Sigma) = \sigma^2 \mathrm{tr}((Z^T Z)^{-1}) \to \sigma^2 \gamma \int \frac{1}{s} dF_\gamma(s).$$

Isotropic with $p > n$

Theorem

Further assume $\Sigma = I$ and $\|\beta\|_2^2 = r^2$ for all n, p. As $n, p \to \infty$, such that $p/n \to \gamma > 1$, the risk of the least squares estimator satisfies, almost surely,

$$R_X(\hat{\beta}; \beta) \to r^2(1 - 1/\gamma) + \frac{\sigma^2}{\gamma - 1}.$$

Bias: consider X is rotationally invariant, then $X \overset{d}{=} XU$ for any orthogonal U. Take U such that $U\beta = re_i$.

$$\begin{aligned}
B_X(\hat{\beta}; \beta) &= \beta^T\big(I - (X^TX)^+X^TX\big)\beta \\
&\overset{d}{=} \beta^T\big(I - U^T(X^TX)^+UU^TX^TXU\big)\beta \\
&= r^2 - (U\beta)^T(X^TX)^+X^TX(U\beta) \\
&\overset{d}{=} r^2\big[1 - \mathrm{tr}\big((X^TX)^+X^TX\big)/p\big] = r^2(1 - n/p).
\end{aligned}$$

Isotropic with $p > n$

Theorem

Further assume $\Sigma = I$ and $\|\beta\|_2^2 = r^2$ for all n, p. As $n, p \to \infty$, such that $p/n \to \gamma > 1$, the risk of the least squares estimator satisfies, almost surely,

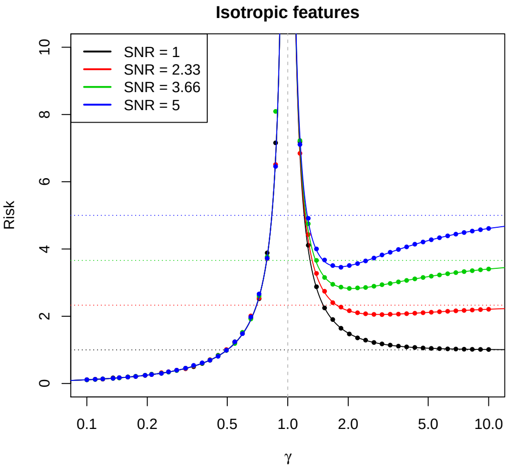
$$R_X(\hat{\beta}; \beta) \to r^2(1 - 1/\gamma) + \frac{\sigma^2}{\gamma - 1}.$$

Variance: let $s_i = \lambda_i(X^T X/n)$, $t_i = \lambda_i(XX^T/p)$.

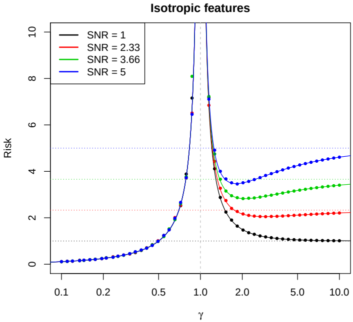
$$V_X(\hat{\beta}; \beta) = \frac{\sigma^2}{n} \sum_{i=1}^{n} \frac{1}{s_i} = \frac{\sigma^2}{p} \sum_{i=1}^{n} \frac{1}{t_i} = \frac{\sigma^2 n}{p} \int \frac{1}{t} \, dF_{XX^T/p}(t)$$
$$\to \frac{\sigma^2}{\gamma - 1}.$$

Risk curve

$$R(\gamma) = \begin{cases} \sigma^2 \frac{\gamma}{1-\gamma} & \text{for } \gamma < 1, \\ r^2\left(1 - \frac{1}{\gamma}\right) + \sigma^2 \frac{1}{\gamma - 1} & \text{for } \gamma > 1. \end{cases}$$



Risk curve: double descent?



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

Outline

Misspecified model

Given Given i.i.d. $(x_i, w_i, y_i) \in \mathbb{R}^p \times \mathbb{R}^q \times \mathbb{R}$, $i = 1, \ldots, n$, where

$$\big((x_i, w_i), \epsilon_i\big) \sim P_{x,w} \times P_\epsilon, \quad \text{and} \quad y_i = x_i^T \beta + w_i^T \theta + \epsilon_i$$

and $\mathbb{E}(\epsilon_i) = 0$ and $\mathrm{Var}(\epsilon_i) = \sigma^2$.
Then the prediction risk is

$$\begin{aligned}
& R_X(\hat{\beta}; \beta, \theta) \\
&= \mathbb{E}\big[\big(x_0^T \hat{\beta} - x_0^T \beta - w_0^T \theta\big)^2 \,|\, X\big] \\
&= \underbrace{\mathbb{E}\big[\big(x_0^T \hat{\beta} - \mathbb{E}(y_0|x_0)\big)^2 \,|\, X\big]}_{R_X^*(\hat{\beta};\beta,\theta)} + \underbrace{\mathbb{E}\big[\big(\mathbb{E}(y_0|x_0) - \mathbb{E}(y_0|x_0, w_0)\big)^2\big]}_{M(\beta,\theta)}.
\end{aligned}$$

Misspecified model: isotropic features

If we are willing to assume $\Sigma = I$, then

$$y_i = x_i^T \beta + \delta_i, \quad i = 1, \ldots, n,$$

where δ_i is independent of x_i, $\mathbb{E}[\delta_i] = 0$ and $\mathbb{E}[\delta_i^2] = \sigma^2 + \|\theta\|_2^2$. Denote

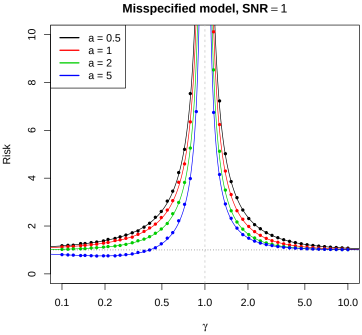
- total signal by $r^2 = \|\beta\|_2^2 + \|\theta\|_2^2$
- fraction of the signal captured by the observed features by $\kappa = \|\beta\|_2^2 / r^2$.

Theorem

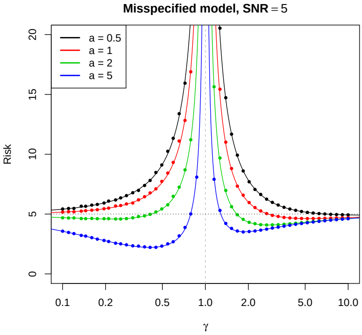
As $n, p \to \infty$, with $p/n \to \gamma$, it holds almost surely that

$$R_X(\hat{\beta}; \beta, \theta) \to \begin{cases} \gamma < 1 : & r^2(1-\kappa) + \left(r^2(1-\kappa) + \sigma^2\right)\frac{\gamma}{1-\gamma} \\ \gamma > 1 : & \\ r^2(1-\kappa) + r^2\kappa\left(1 - \frac{1}{\gamma}\right) + \left(r^2(1-\kappa) + \sigma^2\right)\frac{1}{\gamma-1} & \end{cases}$$

Risk curve

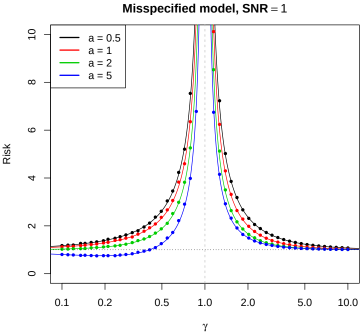


(a) SNR = 1

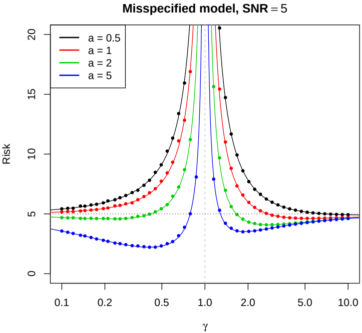


(b) SNR = 5

Risk curve



(c) SNR = 1

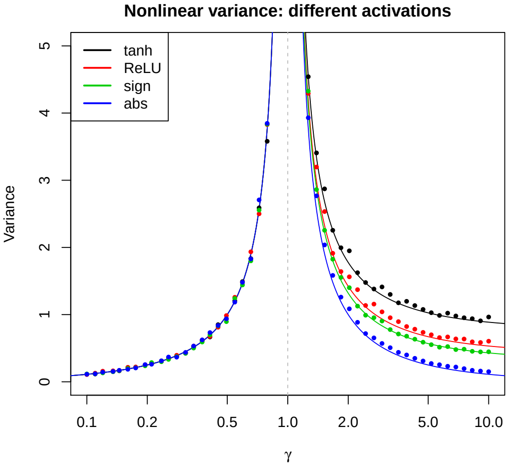


(d) SNR = 5

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

What about nonlinear

$x_i = \varphi(Wz_i)$, where $z_i \in \mathbb{R}^d$ has i.i.d. entries from $N(0,1)$, $W \in \mathbb{R}^{p\times d}$ has i.i.d. entries from $N(0, 1/d)$, and φ is an activation function acting componentwise.



Outline

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

Mei, Montanari

- Given i.i.d $(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$

$$\boldsymbol{x}_i \sim \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d})) \quad \text{and} \quad y_i = f_\star(\boldsymbol{x}_i)$$

- Recall the random features (RF) model

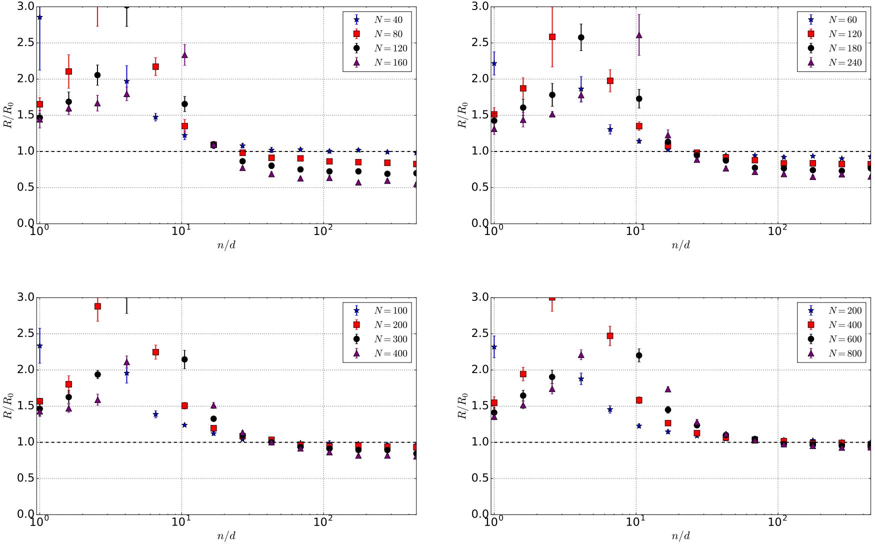
$$\mathcal{F}_{\mathrm{RF}}(\boldsymbol{W}) \equiv \left\{ f(\boldsymbol{x}) = \sum_{i=1}^{N} a_i \, \sigma(\langle \boldsymbol{w}_i, \boldsymbol{x} \rangle) \ : \quad a_i \in \mathbb{R} \ \forall i \leq N \right\}.$$

  where $\boldsymbol{w}_i \sim \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{1}))$

- Kernel ridge regression

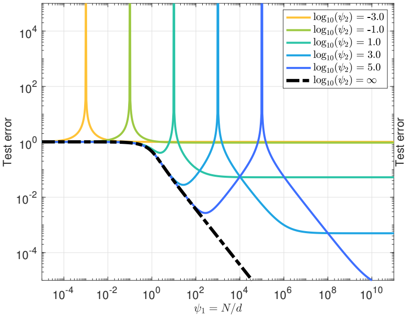
$$\hat{a}(\lambda) = \arg\min_{a \in \mathbb{R}^N} \left\{ \hat{\mathbb{E}}_n \left[(y - \sum_{i=1}^{N} \sigma(\langle \boldsymbol{w}_i, \boldsymbol{x} \rangle))^2 \right] + \frac{N\lambda}{d} \|a\|^2 \right\}$$

RF with ReLU for quadratic $d = 20, 30, 50, 100$

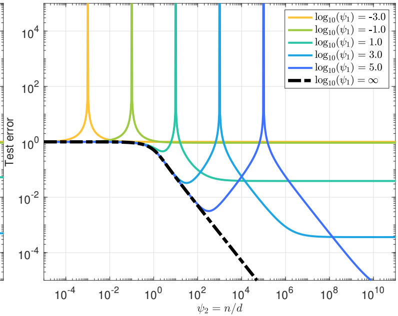


RF with ReLU

- $f_d(\boldsymbol{x}) = \langle \boldsymbol{\beta}_1, \boldsymbol{x} \rangle$ with $\|\boldsymbol{\beta}_1\|_2^2 = 1$.
- SNR: $\|\boldsymbol{\beta}_1\|_2^2 / \tau^2 \equiv \rho = 2$



(e) $\psi_1 = N/d$



(f) $\psi_2 = n/d$

References

[1] Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O. (2016).
Understanding deep learning requires rethinking generalization.
arXiv preprint arXiv:1611.03530.

[2] Belkin, M, Hsu, D., Ma S., Mandal, S. (2019).
Reconciling modern machine-learning practice and the classical bias-variance trade-off.
Proceedings of the National Academy of Sciences, 116(32), 15849-15854.

[3] Hastie, T., Montanari, A., Rosset, S., Tibshirani, R. J. (2019).
Surprises in high-dimensional ridgeless least squares interpolation.
arXiv preprint arXiv:1903.08560.

[4] Ghorbani, B., Mei, S., Misiakiewicz, T., Montanari, A. (2019).
Linearized two-layers neural networks in high dimension.
arXiv preprint arXiv:1904.12191.

[5] Mei, S., Montanari, A. (2019).
The generalization error of random features regression: Precise asymptotics and double descent curve.
arXiv preprint arXiv:1908.05355.

[6] Belkin's slides at Simons Institute (2019)
`https://simons.berkeley.edu/sites/default/files/docs/14172/slides1.pdf`

[6] Tibshirani's slides
`http://www.stat.cmu.edu/~ryantibs/talks/ls-inter-2019.pdf`