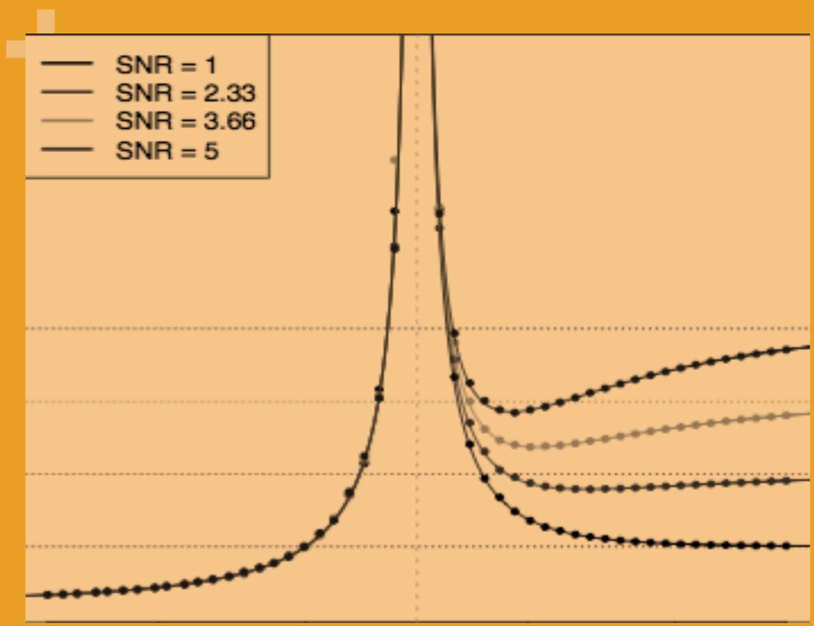


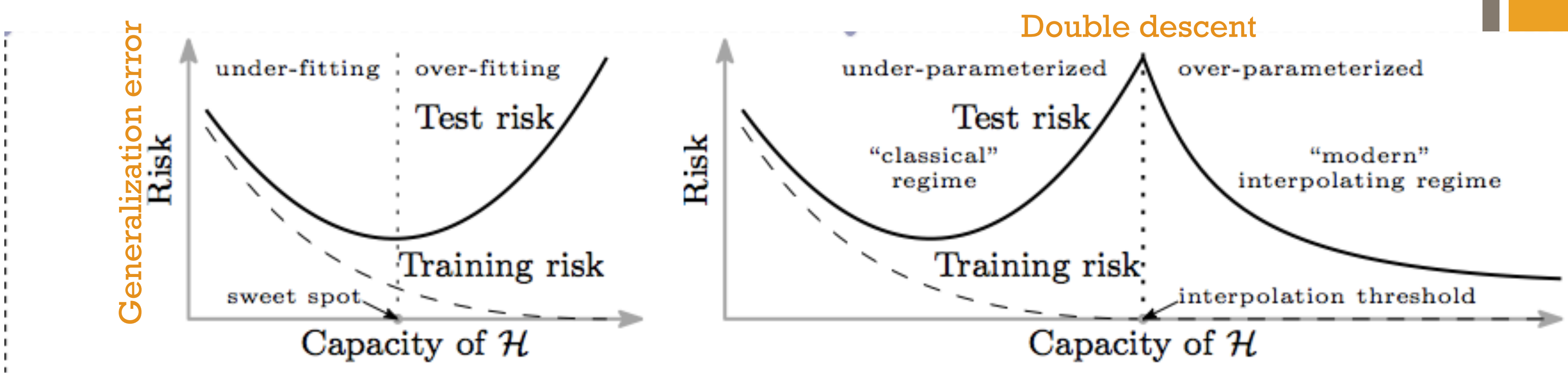
Double Descent

STAT 535+LPL2019

Beyond the Bias-Variance trade-off

Marina Meila
University of Washington

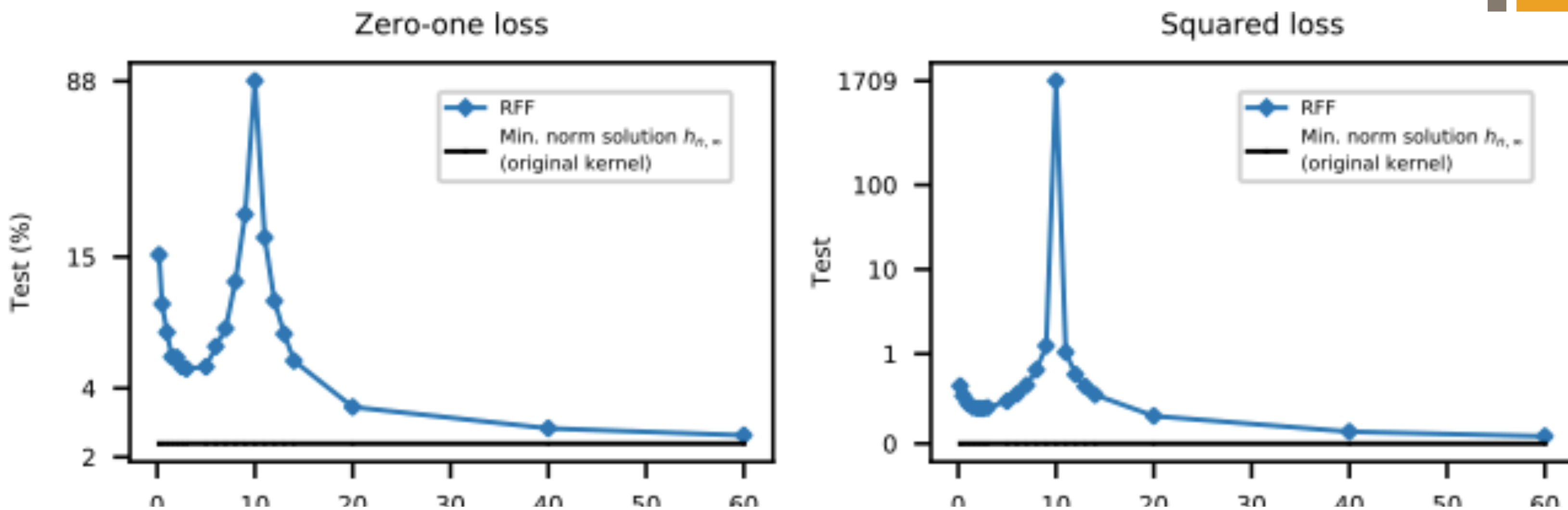
What is observed



Belkin, Hsu, Ma, Mandal 2018

- Classical regime p < N
- Modern/Deep Learning/High dimensional regime N > n
  - Think N fixed, p increases, gamma=p/N
  - Training error = 0 (interpolation)
  - Test error decreases with p (or gamma)

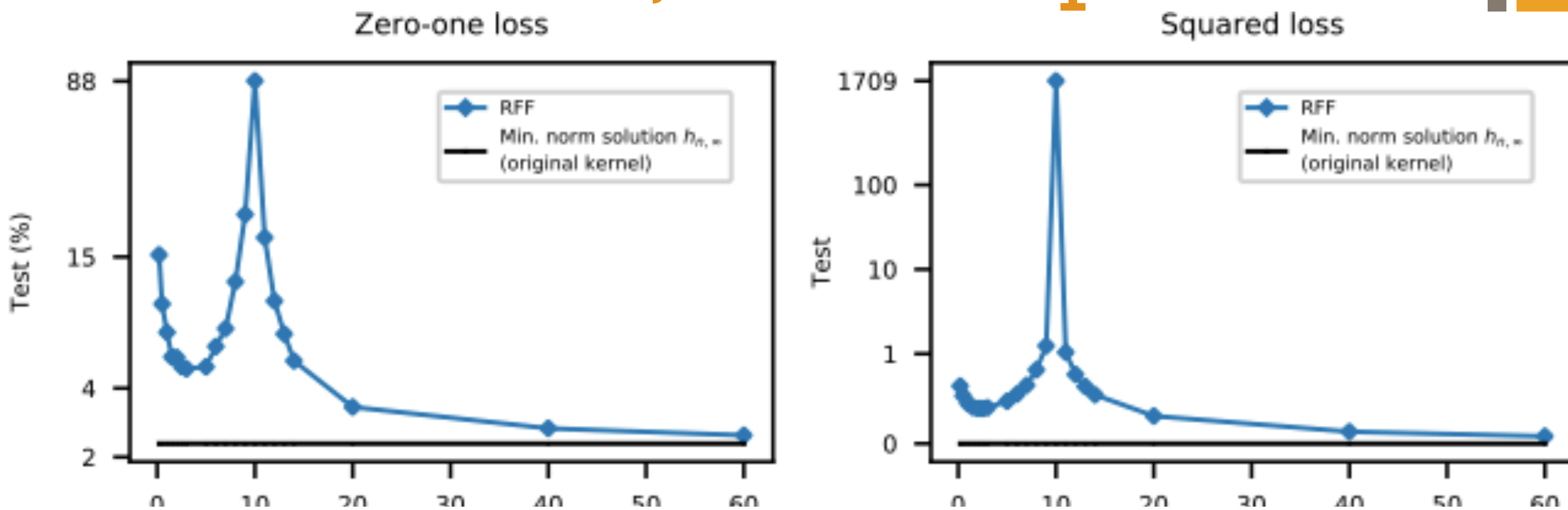
What is observed



Belkin, Hsu, Ma, Mandal 2018

- Double descent curves for the generalization error
  - Random Fourier Features (RFF)
  - ReLU 2 layer networks (with random first layer weights)
  - Random Forests, l2-Adaboost
  - Linear regression
- With and without noise

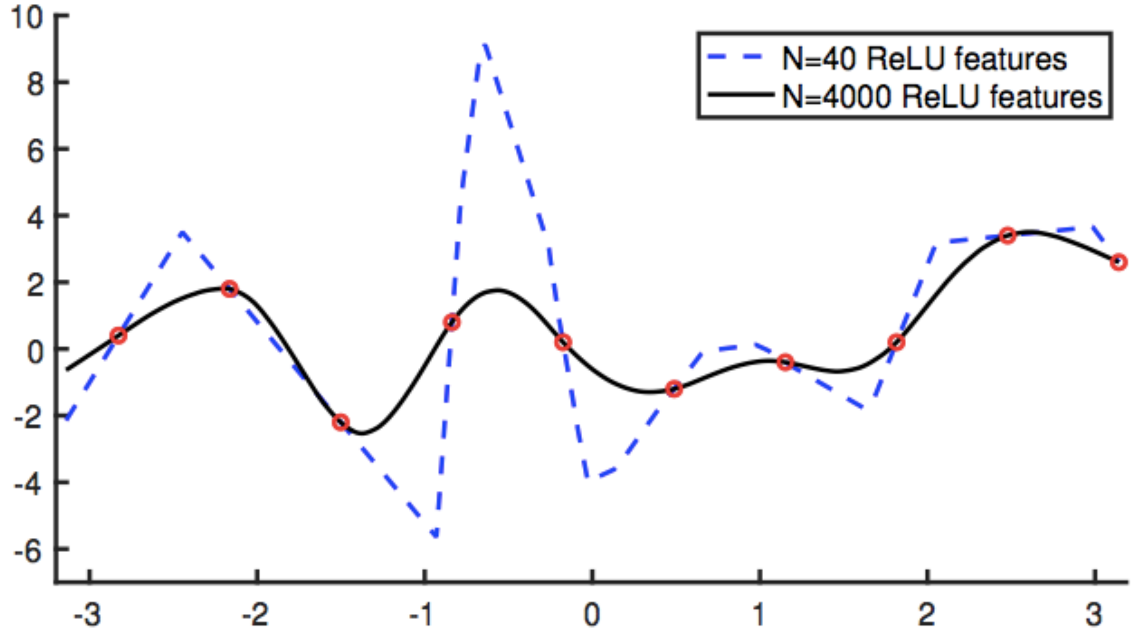
Double descent, the case p > N



Belkin, Hsu, Ma, Mandal 2018

- Model y = <phi(x), beta >
- Large N (cover a compact data domain)
- Features random
- Min-norm solution beta*

Main intuition [Belkin et al.]



- The target function h^* is (mostly) smooth
  - i.e. $||h^*||_{RKHS}$ is small
- $p > N$, no noise, hence h_p interpolates data
- Train to minimize $||h_p||$ subject to 0 training error
- Then $||h_p||$ will decrease with p!

Random Fourier Features (RFF)

Random Fourier features. We first consider a popular class of non-linear parametric models called *Random Fourier Features* (*RFF*) [30], which can be viewed as a class of two-layer neural networks with fixed weights in the first layer. The RFF model family $\mathcal{H}_N$ with N (complex-valued) parameters consists of functions $h\colon \mathbb{R}^d \to \mathbb{C}$ of the form

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

and the vectors $v_1, \ldots, v_N$ are sampled independently from the standard normal distribution in $\mathbb{R}^d$. (We consider $\mathcal{H}_N$ as a class of real-valued functions with $2N$ real-valued parameters by taking real and imaginary parts separately.) Note that $\mathcal{H}_N$ is a randomized function class, but as $N \to \infty$, the function class becomes a closer and closer approximation to the Reproducing Kernel Hilbert Space (RKHS) corresponding to the Gaussian kernel, denoted by $\mathcal{H}_\infty$.

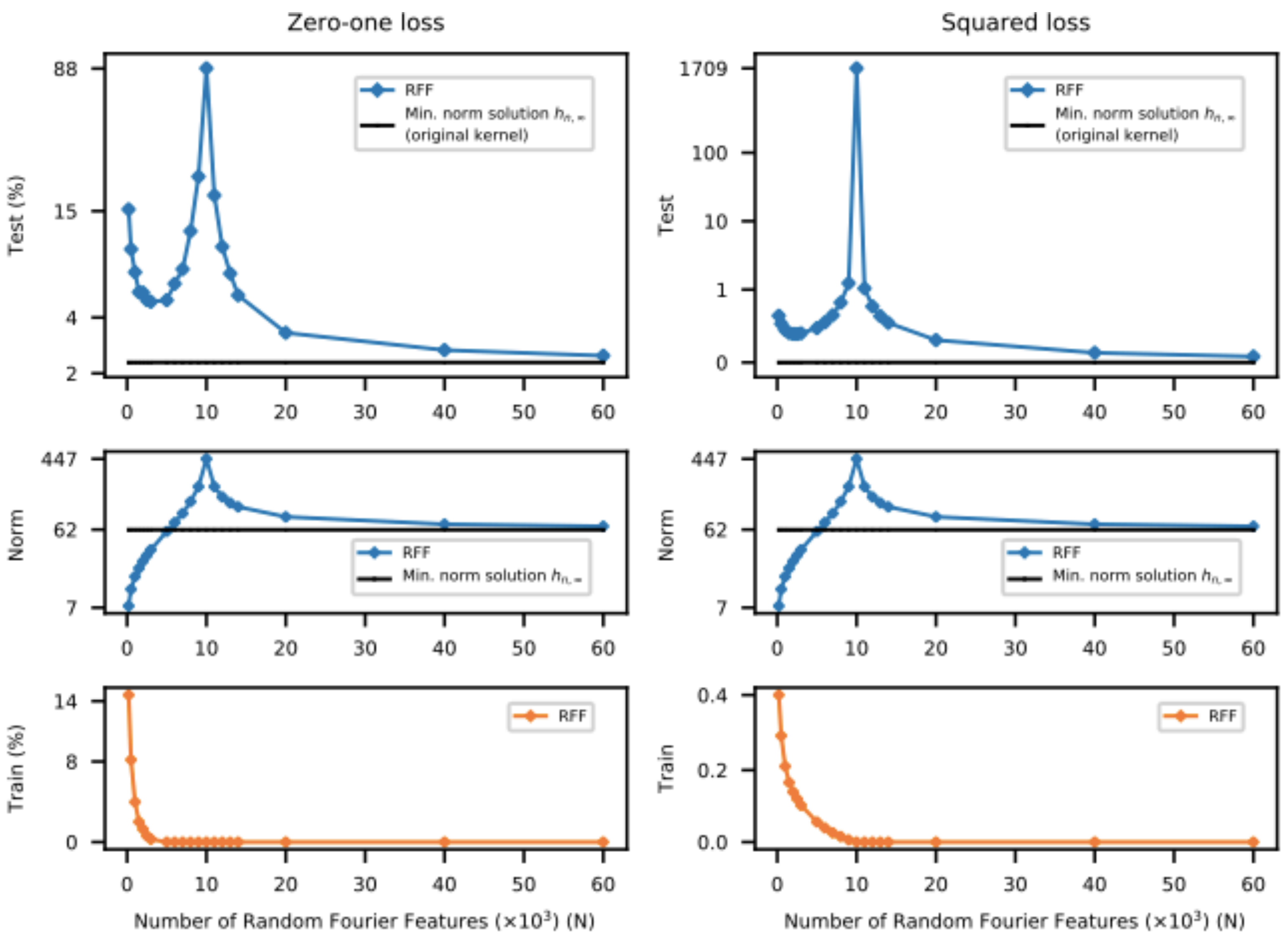
- RFF → $\mathcal{H}_{\text{infinity}}$

Theorem

Theorem 1. *Fix any $h^* \in \mathcal{H}_\infty$. Let $(x_1, y_1), \ldots, (x_n, y_n)$ be independent and identically distributed random variables, where x_i is drawn uniformly at random from a compact cube*[2] *$\Omega \subset \mathbb{R}^d$, and $y_i = h^*(x_i)$ for all i. There exists absolute constants $A, B > 0$ such that, for any interpolating $h \in \mathcal{H}_\infty$ (i.e., $h(x_i) = y_i$ for all i), so that with high probability*

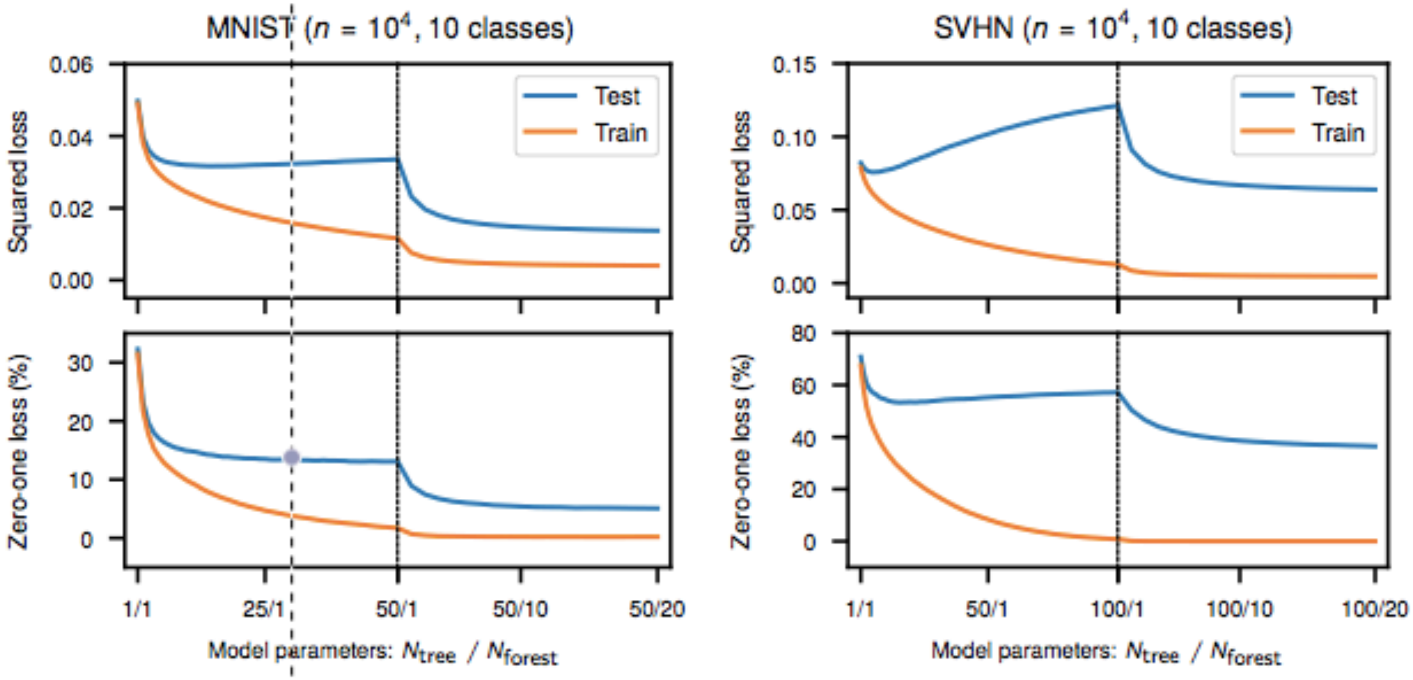
$$\sup_{x \in \Omega} |h(x) - h^*(x)| < A e^{-B(n/\log n)^{1/d}} \left(\|h^*\|_{\mathcal{H}_\infty} + \|h\|_{\mathcal{H}_\infty} \right).$$

RFF



Zero-one loss
Squared loss
Test (%)
Test
RFF
Min. norm solution $h_{n,\infty}$
(original kernel)
Norm
Train (%)
Train
Number of Random Fourier Features ($\times 10^3$) (N)

Boosted decision trees



Linear regression

[Hastie, Montanari, Rosset, Tibshirani 2019]

- Linear, nonlinear features behave the same way
- Model correct, misspecified
- Noise level sigma affects asymptotic error
- and optimal N/n
- Double descent is not regularization

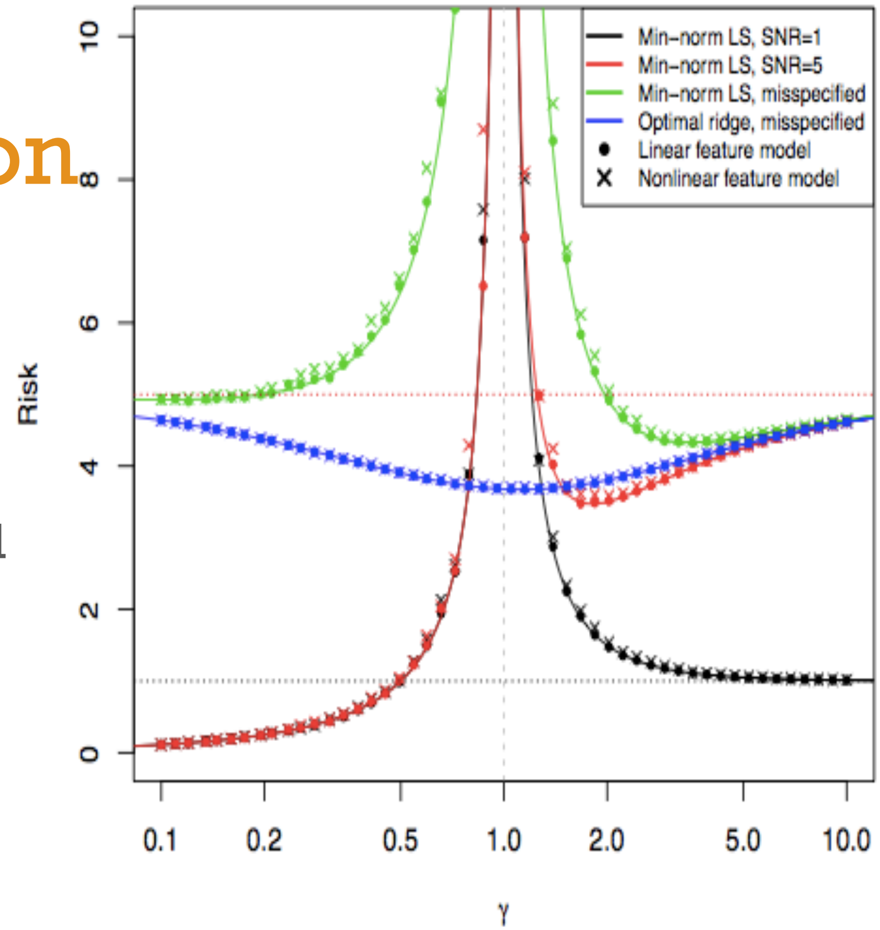
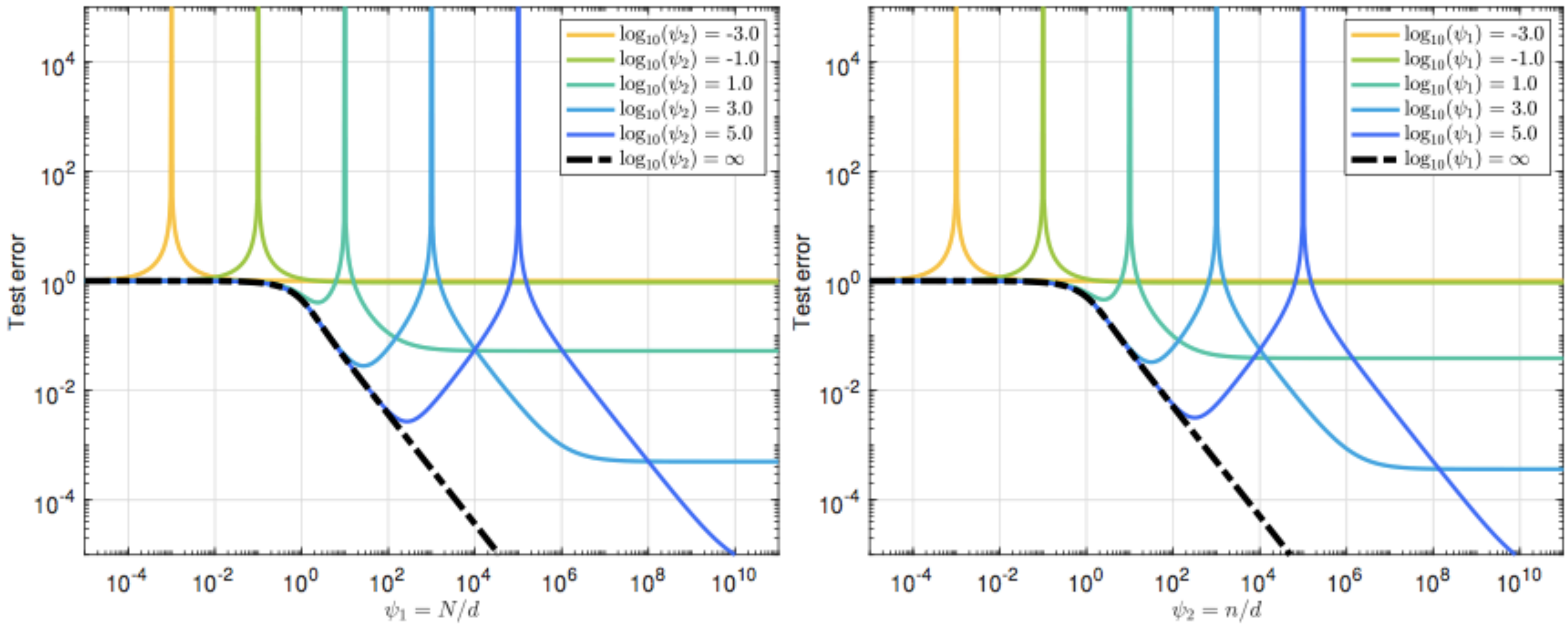


Figure 1: Asymptotic risk curves for the linear feature model, as a function of the limiting aspect ratio γ. The risks for min-norm least squares, when $\mathrm{SNR} = 1$ and $\mathrm{SNR} = 5$, are plotted in black and red, respectively. These two match for $\gamma < 1$ but differ for $\gamma > 1$. The null risks for $\mathrm{SNR} = 1$ and $\mathrm{SNR} = 5$ are marked by the dotted black and red lines, respectively. The risk for the case of a misspecified model (with significant approximation bias, $a = 1.5$ in (13)), when $\mathrm{SNR} = 5$, is plotted in green. Optimally-tuned (equivalently, CV-tuned) ridge regression, in the same misspecified setup, has risk plotted in blue. The points denote finite-sample risks, with $n = 200$, $p = [\gamma n]$, across various values of γ, computed from features X having i.i.d. $N(0, 1)$ entries. Meanwhile, the "x" points mark finite-sample risks for a nonlinear feature model, with $n = 200$, $p = [\gamma n]$, $d = 100$, and $X = \varphi(ZW^T)$, where Z has i.i.d. $N(0, 1)$ entries, W has i.i.d. $N(0, 1/d)$ entries, and $\varphi(t) = a(|t| - b)$ is a "purely nonlinear" activation function, for constants a, b. The theory predicts that this nonlinear risk should converge to the linear risk with p features (regardless of d). The empirical agreement between these two—and the agreement in finite-sample and asymptotic risks—is striking.



- More refined analysis includes noise, non-linearity, data dimension n, ridge regularization lambda [Mei, Montanari 2019]
- When is global minimum in overparametrized regime?
- Enough data N/n > 1
- lambda → 0 (or min-norm LS)
- p >> N
- SNR || beta ||/noise > 1
- Bias, Variance strictly decreasing with p/N to > 0 limit