On the robustness of overparametrized models

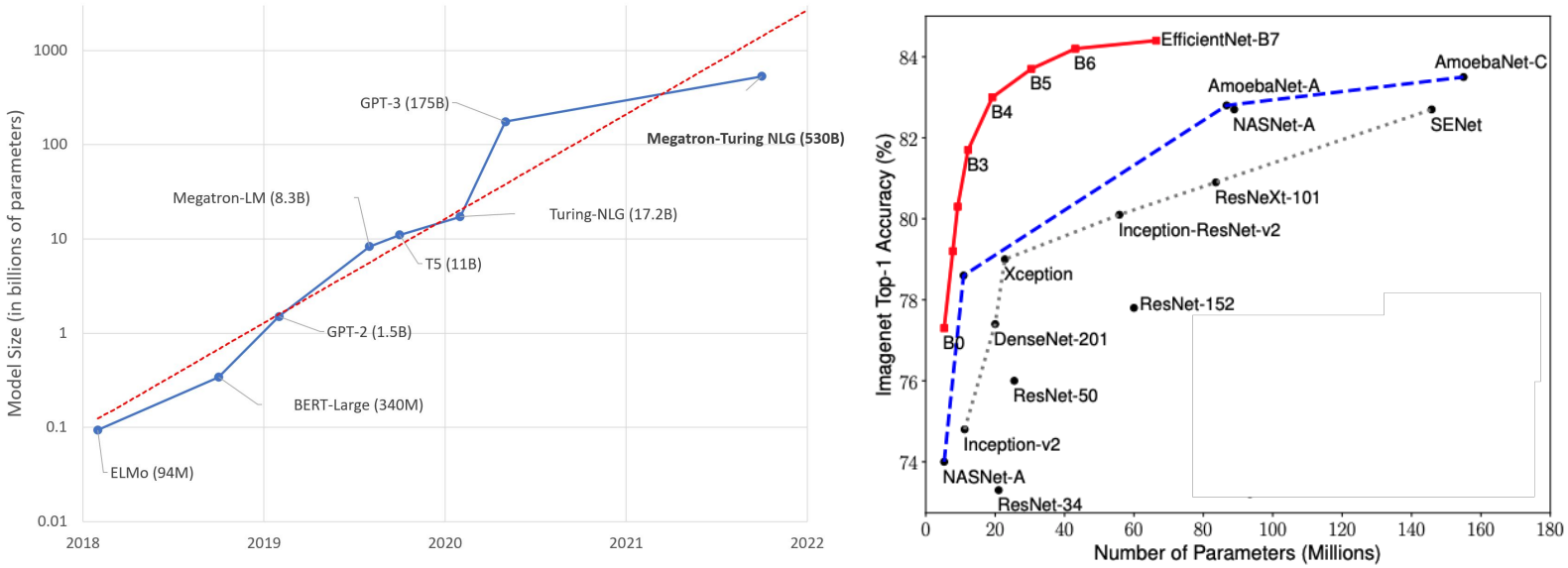
Antônio H. Ribeiro

"To think is to forget a difference, to generalize, to abstract.
In the overly replete world of Funes,
there were nothing but details."
Funes, the Memorious (1942)
Jorge Luis Borges

μ-seminar

November 12th, 2021

Model size in neural networks



(a) Language models

(b) Image classification

Figure: Models number of parameters

Sources: J. Simon (2021) "Large Language Models: A New Moore's Law?". Online (acessed: 2021-11-09). URL: huggingface.co/blog/large-language-models .
M. Tan and Q. V. Le (2019) "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," Proceedings of the 36th International Conference on Machine Learning (ICML). PMLR, vol. 97.

Overparametrized models

The unreasonable effectiveness of overparameterized machine learning models (2021, Fall) — PhD level course (together with Dave and Per).

All material available in: https://github.com/uu-sml/seminars-overparam-ml

Inaugural paper: M. Belkin, D. Hsu, S. Ma, and S. Mandal (2019), "Reconciling modern machine-learning practice and the classical bias–variance trade-off," Proc Natl Acad Sci USA, vol. 116, no. 32, pp. 15849–15854, doi: 10.1073/pnas.1903070116.

PNAS PNAS PNAS PNAS

Reconciling modern machine-learning practice and the classical bias–variance trade-off

Mikhail Belkin[a,b,1], Daniel Hsu[c], Siyuan Ma[a], and Soumik Mandal[a]

[a]Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210; [b]Department of Statistics, The Ohio State University, Columbus, OH 43210; and [c]Computer Science Department and Data Science Institute, Columbia University, New York, NY 10027

Edited by Peter J. Bickel, University of California, Berkeley, CA, and approved July 2, 2019 (received for review February 21, 2019)

Breakthroughs in machine learning are rapidly changing science and society, yet our fundamental understanding of this technology has lagged far behind. Indeed, one of the central tenets of the field, the bias–variance trade-off, appears to be at odds with the observed behavior of methods used in modern machine-learning practice. The bias–variance trade-off implies that a model should balance underfitting and overfitting: Rich enough to express underlying structure in data and simple enough to avoid fitting spurious patterns. However, in modern practice, very rich models such as neural networks are trained to exactly fit (i.e., interpolate) the data. Classically, such models would be considered overfitted, and yet they often obtain high accuracy on test data. This apparent contradiction has raised questions about the mathematical foundations of machine learning and their relevance to practitioners. In this paper, we reconcile the classical understanding and the modern practice within a unified performance curve. This "double-descent" curve subsumes the textbook U-shaped bias–variance trade-off curve by showing how increasing model capacity beyond the point of interpolation results in improved performance. We provide evidence for the existence and ubiquity of double descent for a wide spectrum of models and datasets, and we posit a mechanism for its emergence. This connection between the performance and the structure of machine-learning models delineates the limits of classical analyses and has implications for both the theory and the practice of machine learning.

machine learning | bias–variance trade-off | neural networks

Machine learning has become key to important applications in science, technology, and commerce. The focus of machine learning is on the problem of prediction: Given a sample of training examples $(x_1, y_1), \ldots, (x_n, y_n)$ from $\mathbb{R}^d \times \mathbb{R}$, we learn a predictor $h_n : \mathbb{R}^d \to \mathbb{R}$ that is used to predict the label y of a new point x, unseen in training.

The predictor h_n is commonly chosen from some function class $\mathcal{H}$, such as neural networks with a certain architecture, using empirical risk minimization (ERM) and its variants. In ERM, the predictor is taken to be a function $h \in \mathcal{H}$ that minimizes the empirical (or training) risk $\frac{1}{n}\sum_{i=1}^{n} \ell(h(x_i), y_i)$, where ℓ is a loss function, such as the squared loss $\ell(y', y) = (y' - y)^2$ for regression or 0–1 loss $\ell(y', y) = \mathbb{1}_{\{y' \neq y\}}$ for classification.

The goal of machine learning is to find h_n that performs well on new data, unseen in training. To study performance on new data (known as generalization), we typically assume the training examples are sampled randomly from a probability distribution P over $\mathbb{R}^d \times \mathbb{R}$ and evaluate h_n on a new test example (x, y)

ing data (i.e., have large empirical risk) and hence predict poorly on new data. 2) If $\mathcal{H}$ is too large, the empirical risk minimizer may overfit spurious patterns in the training data, resulting in poor accuracy on new examples (small empirical risk but large true risk).

The classical thinking is concerned with finding the "sweet spot" between underfitting and overfitting. The control of the function class capacity may be explicit, via the choice of $\mathcal{H}$ (e.g., picking the neural network architecture), or it may be implicit, using regularization (e.g., early stopping). When a suitable balance is achieved, the performance of h_n on the training data is said to generalize to the population P. This is summarized in the classical U-shaped risk curve shown in Fig. 1*A* that has been widely used to guide model selection and is even thought to describe aspects of human decision making (3). The textbook corollary of this curve is that "a model with zero training error is overfit to the training data and will typically generalize poorly" (ref. 2, p. 221), a view still widely accepted.

However, practitioners routinely use modern machine-learning methods, such as large neural networks and other nonlinear predictors that have very low or zero training risk. Despite the high function class capacity and near-perfect fit to training data, these predictors often give very accurate predictions on new data. Indeed, this behavior has guided a best practice in deep learning for choosing neural network architectures, specifically that the network should be large enough to permit effortless zero-loss training (called interpolation) of the training data (4). Moreover, in direct challenge to the bias–variance trade-off philosophy, recent empirical evidence indicates that neural

Significance

While breakthroughs in machine learning and artificial intelligence are changing society, our fundamental understanding has lagged behind. It is traditionally believed that fitting models to the training data exactly is to be avoided as it leads to poor performance on unseen data. However, powerful modern classifiers frequently have near-perfect fit in training, a disconnect that spurred recent intensive research and controversy on whether theory provides practical insights. In this work, we show how classical theory and modern practice can be reconciled within a single unified performance curve and propose a mechanism underlying its emergence. We believe this previously unknown pattern connecting the structure and performance of learning architectures will help shape design and understanding of learning algorithms.

STATISTICS

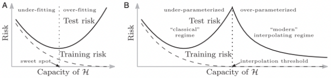


Fig. 1. Curves for training risk (dashed line) and test risk (solid line). (*A*) The classical U-shaped risk curve arising from the bias–variance trade-off. (*B*) The double-descent risk curve, which incorporates the U-shaped risk curve (i.e., the "classical" regime) together with the observed behavior from using high-capacity function classes (i.e., the "modern" interpolating regime), separated by the interpolation threshold. The predictors to the right of the interpolation threshold have zero training risk.

networks and kernel machines trained to interpolate the training data obtain near-optimal test results even when the training data are corrupted with high levels of noise (5, 6).

The main finding of this work is a pattern in how performance on unseen data depends on model capacity and the mechanism underlying its emergence. This dependence, empirically witnessed with important model classes including neural networks and a range of datasets, is summarized in the "double-descent" risk curve shown in Fig. 1*B*. The curve subsumes the classical U-shaped risk curve from Fig. 1*A* by extending it beyond the point of interpolation.

When function class capacity is below the "interpolation threshold," learned predictors exhibit the classical U-shaped curve from Fig. 1*A*. (In this paper, function class capacity is identified with the number of parameters needed to specify a function within the class.) The bottom of the U is achieved at the sweet spot which balances the fit to the training data and the susceptibility to overfitting: To the left of the sweet spot, predictors are underfitted, and immediately to the right, predictors are overfitted. When we increase the function class capacity high enough (e.g., by increasing the number of features or the size of the neural network architecture), the learned predictors achieve (near) perfect fits to the training data—i.e., interpolation. Although the learned predictors obtained at the interpolation threshold typically have high risk, we show that increasing the function class capacity beyond this point leads to decreasing risk, typically going below the risk achieved at the sweet spot in the "classical" regime.

All of the learned predictors to the right of the interpolation threshold fit the training data perfectly and have zero empirical risk. So why should some—in particular, those from richer functions classes—have lower test risk than others? The answer is that the capacity of the function class does not necessarily reflect how well the predictor matches the inductive bias appropriate for the problem at hand. For the learning problems we consider (a range of real-world datasets as well as synthetic data), the inductive bias that seems appropriate is the regular-

apply to regression and also does not predict the second descent beyond the interpolation threshold. Recently, there has been an emerging recognition that certain interpolating predictors (not based on ERM) can indeed be provably statistically optimal or near optimal (11, 12), which is compatible with our empirical observations in the interpolating regime.

In the remainder of this article, we discuss empirical evidence for the double-descent curve and the mechanism for its emergence and conclude with some final observations and parting thoughts.

Neural Networks

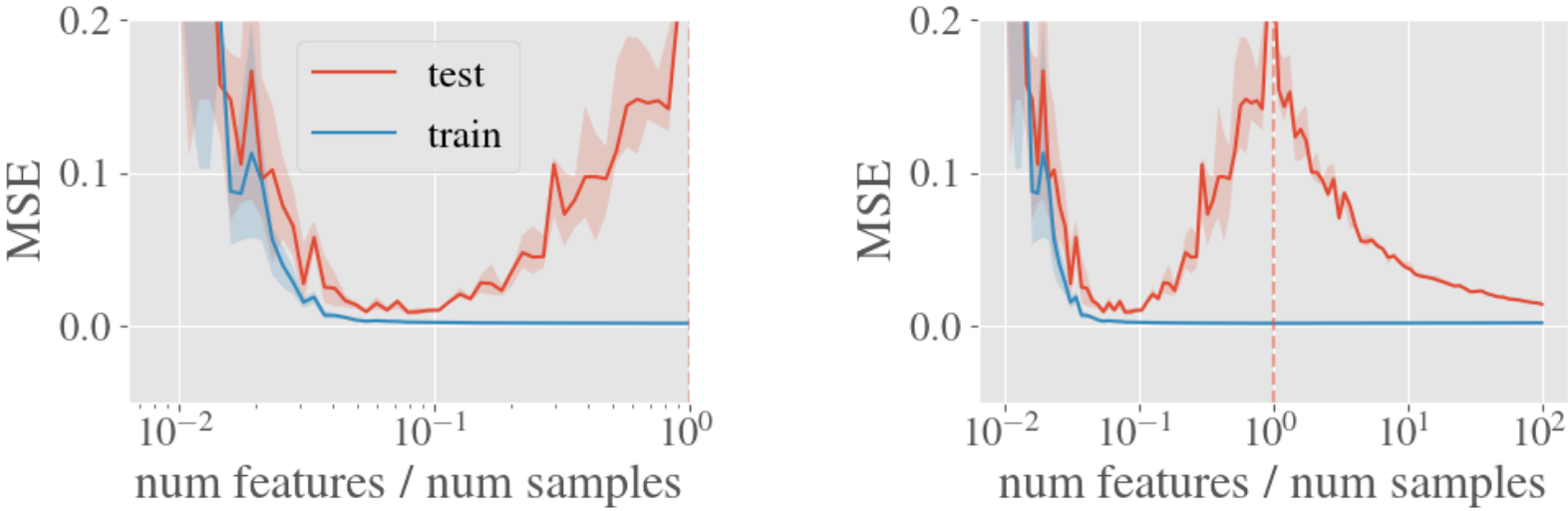
In this section, we discuss the double-descent risk curve in the context of neural networks.

Random Fourier Features. We first consider a popular class of nonlinear parametric models called random Fourier features (RFF) (13), which can be viewed as a class of 2-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 $h : \mathbb{R}^d \to \mathbb{C}$ of the form

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

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 $\mathcal{H}_\infty$. While it is possible to directly use $\mathcal{H}_\infty$ [e.g., as is done with kernel machines (14)], the random classes $\mathcal{H}_N$ are computationally attractive to use when the sample size n is large but the number of parameters N is small compared with n.

Double-descent



(a) U-Shape

(b) Double-descent

Figure: Nonlinear ARX performance in Couple Eletric Drives benchmark.

Antônio H. Ribeiro, Johannes N. Hendriks, Adrian G. Wills, Thomas B. Schön. "Beyond Occam's Razor in System Identification: Double-Descent when Modeling Dynamics". Proceedings of the 19th IFAC Symposium on System Identification (SYSID) - IFAC-PapersOnLine (2021)

Double-descent in linear models

Estimated parameter: using train dataset (x_i, y_i), $i = 1, \cdots, n$:

- Underparametrized:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i} (y_i - x_i{}^{\mathsf{T}} \beta)^2$$

- Overparametrized:

$$\hat{\beta} = \arg\min_{\beta} \|\beta\|_2^2$$
$$\text{subject to } y_i = x_i{}^{\mathsf{T}} \beta$$
$$\text{for every } i$$

Random features: Belking et.al. (2019) generates the features through the nonlinear mapping: $\phi : u_i \mapsto x_i$ obtained from Random Fourier Features.

Benign overfitting

- When the solution in the overparametrized region tends to the optimal one?
- Depend on the eigenvalues of the covariance matrix $\Sigma = \text{Cov}\,[x]$.

P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler, "Benign overfitting in linear regression," Proceedings of the National Academy of Sciences, vol. 117, no. 48, pp. 30063–30070, Apr. 2020, doi: 10.1073/pnas.1907378117.

Example: Latent space model

Data is generated from a lower dimensional subspace: $\dim(z_i) < \dim(x_i)$.

$$x_i = W z_i + \xi_i,$$

$$y_i = \theta^{\mathsf{T}} z_i + \epsilon_i,$$

$W \rightsquigarrow$ Projects z into a higher dimensional space

$\theta \rightsquigarrow$ Output is a linear combination of z, through θ

$\xi \rightsquigarrow$ error in variables

$\epsilon \rightsquigarrow$ additive noise in the output

T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani, "Surprises in High-Dimensional Ridgeless Least Squares Interpolation," arXiv:1903.08560, Nov. 2019.

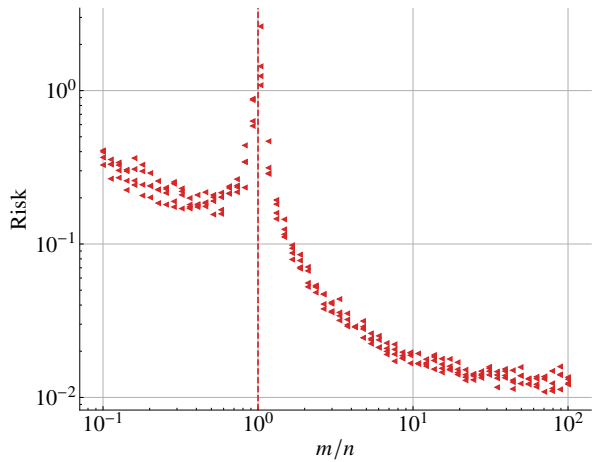


Figure: Model performance vs ratio between features and datapoints

Overparametrized models can generalize effectively when train and test come from the **same** *distribution...*

What when there is a distributional shift?
are they robust?

A.H. Ribeiro, T.B. Schön, "Overparametrized Regression Under L2 Adversarial Attacks". Workshop on the Theory of Overparameterized Machine Learning (TOPML) (2021)
A.H. Ribeiro, T.B. Schön, "On the adversarial robustness of overparametrized linear regression," Work in progress.

Adversarial Attacks

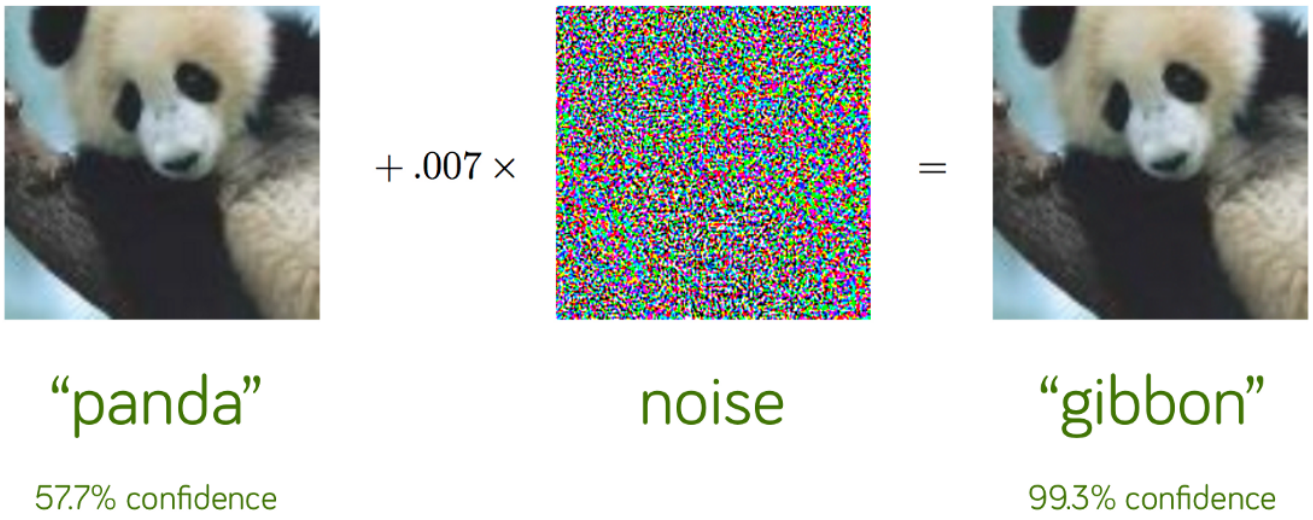


Figure: Illustration of adversarial attack.

Source: I. J. Goodfellow, J. Shlens, C. Szegedy , *“Explaining and Harnessing Adversarial Examples”*, ICLR 2015.

Linear regression under adversarial attacks

Given a data point not seen during training (x, y).

Standard risk:

$$R = E\Big\{(y - x^{\mathsf{T}}\hat{\beta})^2 | \underbrace{x_i, i = 1, \cdots, n}_{\text{training inputs}}\Big\}$$

Adversarial risk:

$$R_p^{\text{adv}} = E\Big\{\max_{\|\Delta x\|_p \leqslant \delta} (y - (x + \Delta x)^{\mathsf{T}}\hat{\beta})^2 | \underbrace{x_i, i = 1, \cdots, n}_{\text{training inputs}}\Big\}$$

$\Delta x \rightsquigarrow$ Adversarially generated disturbance

Can the double-descent be observed in adversarial scenarios?

1. Enlarging the function classes we are able to find models that are smoother.

   M. Belkin, D. Hsu, S. Ma, and S. Mandal. Reconciling modern machine-learning practice and theclassical bias–variance trade-off.Proceedings ofthe National Academy of Sciences, 116(32):15849–15854, Aug. 2019. ISSN 0027-8424, 1091-6490. doi:10.1073/pnas.1903070116

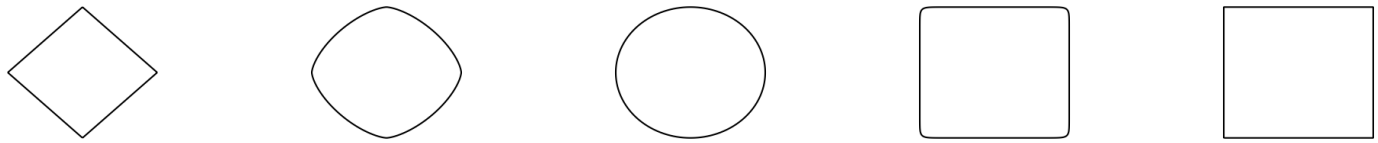
   S. Bubeck and M. Sellke. A Universal Law of Robustness via Isoperimetry. arXiv:2105.12806, June2021

2. Robustness-accuracy tradeoff

   A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards Deep Learning Models Resistantto Adversarial Attacks.Proceedings of the International Conference for Learning Representations (ICLR), 2018.

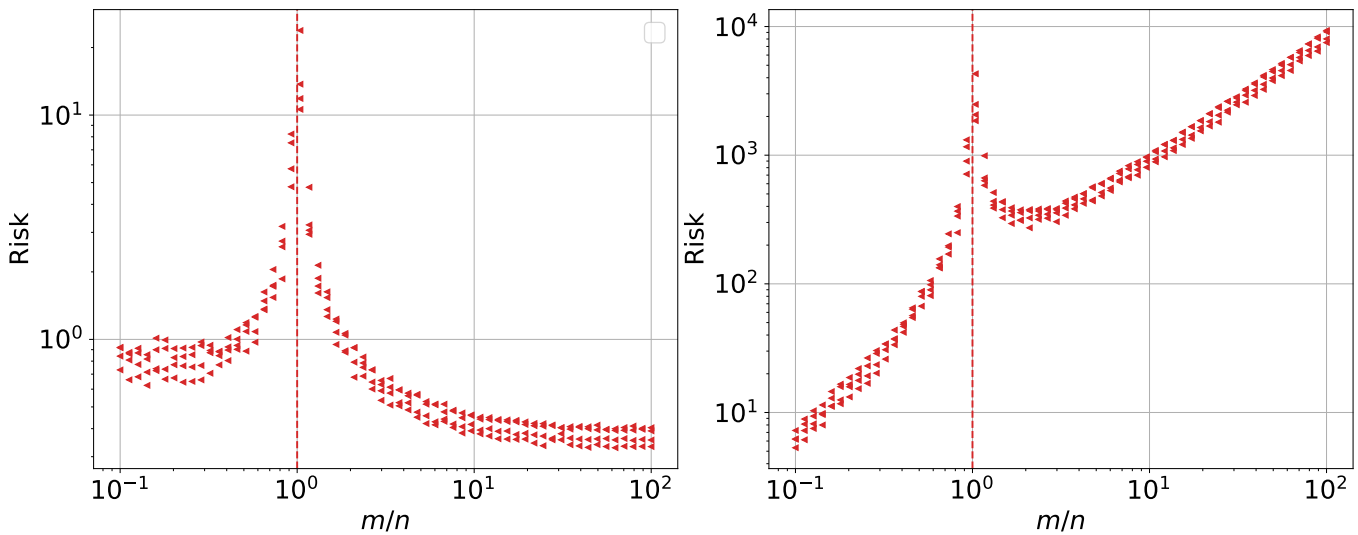
Adversarial attacks

Geometry of the adversarial attack for $\dim(x) = 2$:



(a) ℓ_1 (b) $\ell_{1.5}$ (c) ℓ_2 (d) ℓ_{20} (e) ℓ_∞

Two types of behavior:



(a) Adversarial ℓ_2 risk (b) Adversarial ℓ_∞ risk

Methods

Bounds on the adversarial risk:

$$R + \delta^2 L_q \leqslant R_p^{\text{adv}} \leqslant \left(\sqrt{R} + \delta\sqrt{L_q}\right)^2.$$

$R^{\text{adv}} \rightsquigarrow$ Adversarial risk

$R \rightsquigarrow$ Risk

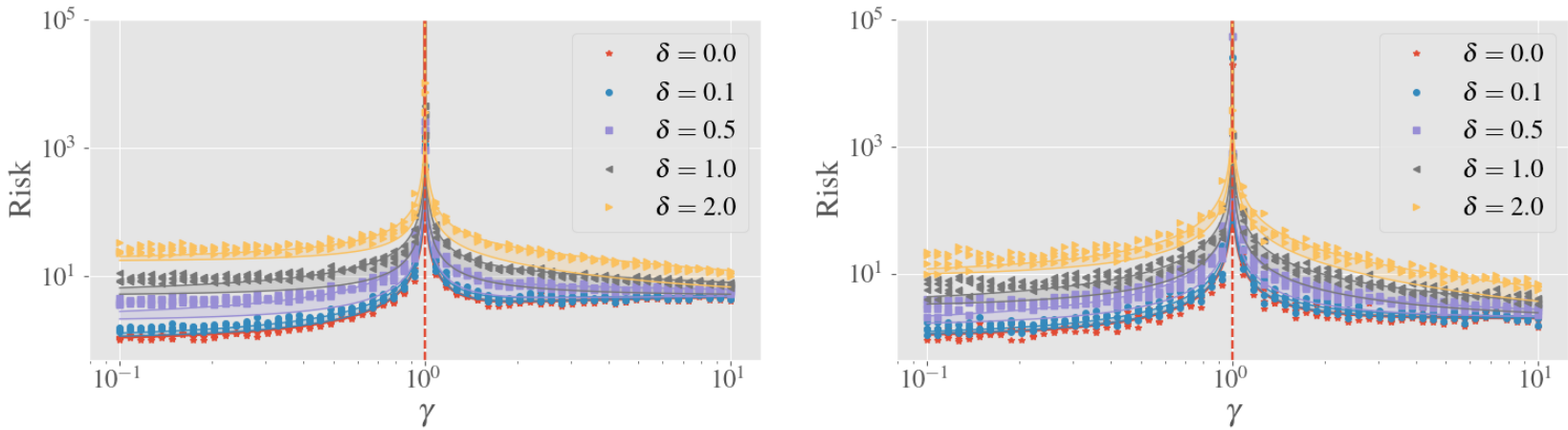
$L_q = E\{\|\hat{\beta}\|_q^2\}$

$\quad \rightarrow$ For an ℓ_p attack $\rightsquigarrow \mathbf{\frac{1}{q} + \frac{1}{p} = 1}$

$\delta \rightsquigarrow$ Adversarial disturbance magnitude

Results (ℓ_2 attacks)

For the ℓ_2-norm we know the asymptotic for all of the above quantities!



(a) Uncorrelated feature (b) Equicorrelated features ($\rho = 0.5$)

Figure: **Adversarial ℓ_2 risk**. The solid line gives the asymptotic risk lower and upper bound computed. The points correspond to the empirical adversarial risk for experiments.

T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani, "Surprises in High-Dimensional Ridgeless Least Squares Interpolation," arXiv:1903.08560, Nov. 2019.

Other ℓ_p attacks

$$\begin{cases} L_2 \leqslant L_q \leqslant (m)^{1-2/p} L_2, & 2 < p \leqslant \infty, \\ (1/m)^{2/p-1} L_2 \leqslant L_q \leqslant L_2, & 1 \leqslant p < 2. \end{cases}$$

$m \rightsquigarrow$ number of features

$L_q = E\{\|\hat{\beta}\|_q^2\}$

For an ℓ_p attack $\rightsquigarrow \mathbf{\frac{1}{q} + \frac{1}{p} = 1}$

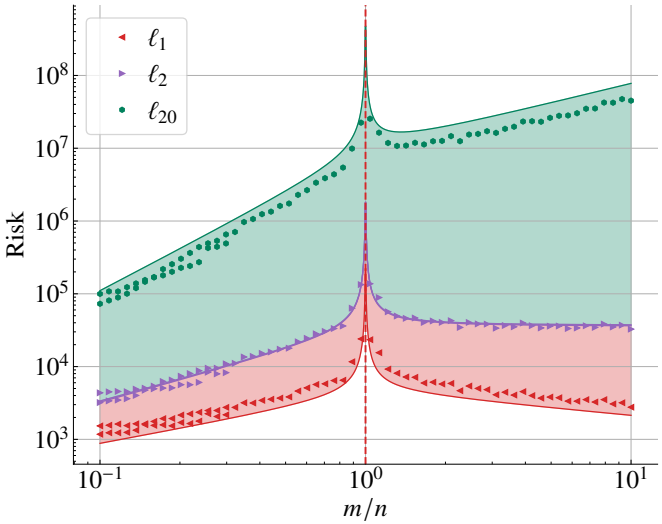


Figure: **Adversarial risk with bounds**.

Conclusion and next steps

- Our framework gives flexibility to experiment diverse scenarios and reuse previous literature.
- We are now focusing on the question when the upper bound in the experiments above is tight.
- It depends on the geometry of the eigenvectors of Σ.

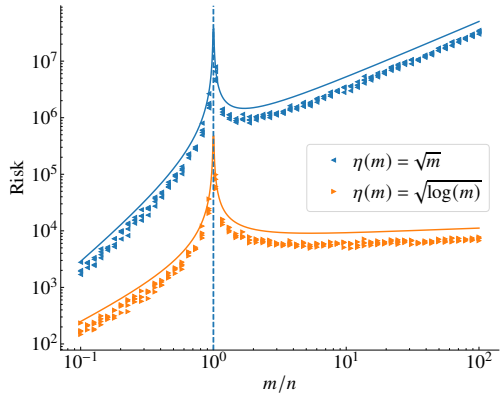


Figure: Example with isotropic features