Uniform Convergence of Interpolators: Gaussian Width, Norm Bounds, and Benign Overfitting

Frederic Koehler
MIT -> Simons Institute

Lijia Zhou
UChicago

Danica J. Sutherland
UBC

Nati Srebro
TTI-Chicago

Interpolation Learning and Double Descent

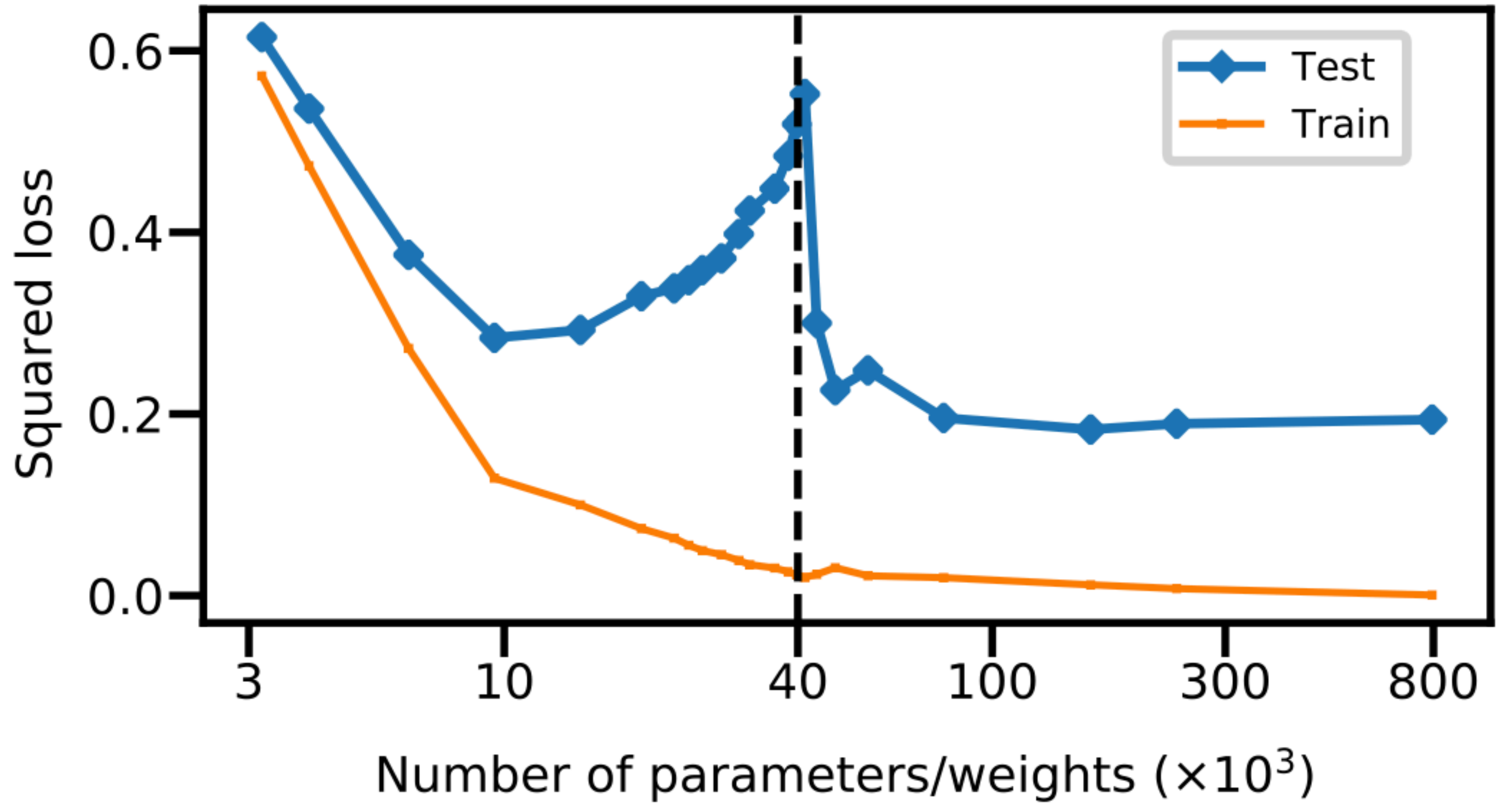


Figure 4: **Double descent risk curve for fully connected neural network on MNIST.**

[Belkin et al, 2019]

The Theoretical Testbed

[Hastie et al, 2019], [Bartlett et al, 2019], [Belkin et al, 2020], [Negrea et al, 2020], [Chinot-Lerasle, 2020], [Ju et al, 2020], [Muthkumar et al, 2020], [Zhou et al, 2020], [Tsigler-Bartlett, 2020], [Bartlett-Long, 2020], [Chinot et al, 2021], …

Gaussian Linear Regression Model

- $X_i \sim N(0,\Sigma)$ iid are the rows of matrix $X : n \times d$
- $Y_i = \langle X_i, w^* \rangle + N(0,\sigma^2)$ and w^* **unknown**
- Goal: given (X,Y), minimize **test error** $L(w) = E[(Y_0 - \langle X_0, w \rangle)^2]$ on fresh sample (X_0, Y_0)

- **Interpolation:** when **training error** $\hat{L}(w) = \frac{1}{n}\|Y - Xw\|_2^2 = 0$
- **Benign overfitting:** $\hat{w} = \arg\min_{\hat{L}(w)=0} \|w\|_2$ is **consistent** in many cases: $L(\hat{w}) \to \sigma^2$

[Bartlett et al, 2019]

Failure of uniform convergence?

- Conventional method for bounding test error:

$$L(w) \leq \hat{L}(w) + \sup_{w \in \mathscr{K}} \left| L(w) - \hat{L}(w) \right|$$

Test error Train Error Generalization Gap

- $\mathscr{K}$ is a class of "simple" hypotheses containing w. Ex. $\mathscr{K} = \{w : \|w\|_2 \leq B\}$
- Unfortunately, this **does not work** in our setting! [Negrea et al, 2020], [Zhou et al, 2020], [Bartlett-Long 2020]
    - "Generalization gap" term is larger than $L(\hat{w}) - \hat{L}(\hat{w}) = \sigma^2$.

Generalization theory for interpolation?

What theoretical analyses do we have?

- ~~VC-dimension/Rademacher complexity/covering/margin bounds.~~
  - ~~Cannot deal with interpolated classifiers when Bayes risk is non-zero.~~
  - ~~Generalization gap cannot be bound when empirical risk is zero.~~
- ~~Regularization-type analyses (Tikhonov, early stopping/SGD, etc.)~~
  - ~~Diverge as $\lambda \to 0$ for fixed n.~~
- ~~Algorithmic stability.~~
  - ~~Does not apply when empirical risk is zero, expected risk nonzero.~~

$$L_{\mathcal{D}}(\hat{f}) \leq L_S(\hat{f}) + \text{bound}$$

WYSIWYG bounds:

training loss $= 0$

$\not\approx$

expected loss

- Classical smoothing methods (i.e., Nadaraya-Watson).

Oracle bounds

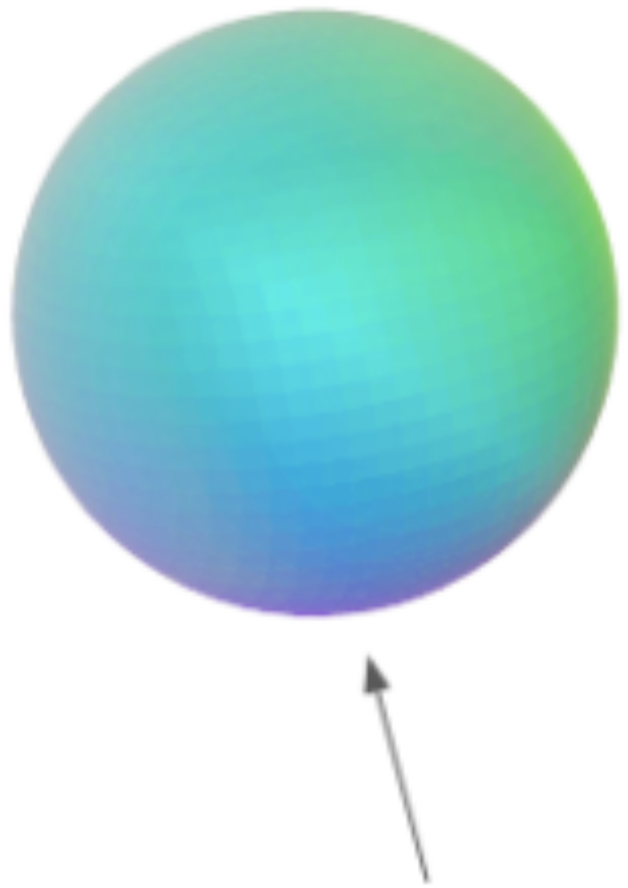
Misha Belkin
Simons Institute
July 2019

A common sentiment: classical learning theory may not be able to explain modern ML & interpolation learning, uniform convergence is obsolete. See also [Neyshabur et al, 2015], [Zhang et al, 2017], [Nagarajan-Kolter, 2019], [Bartlett-Long, 2020], [Belkin, 2021] …

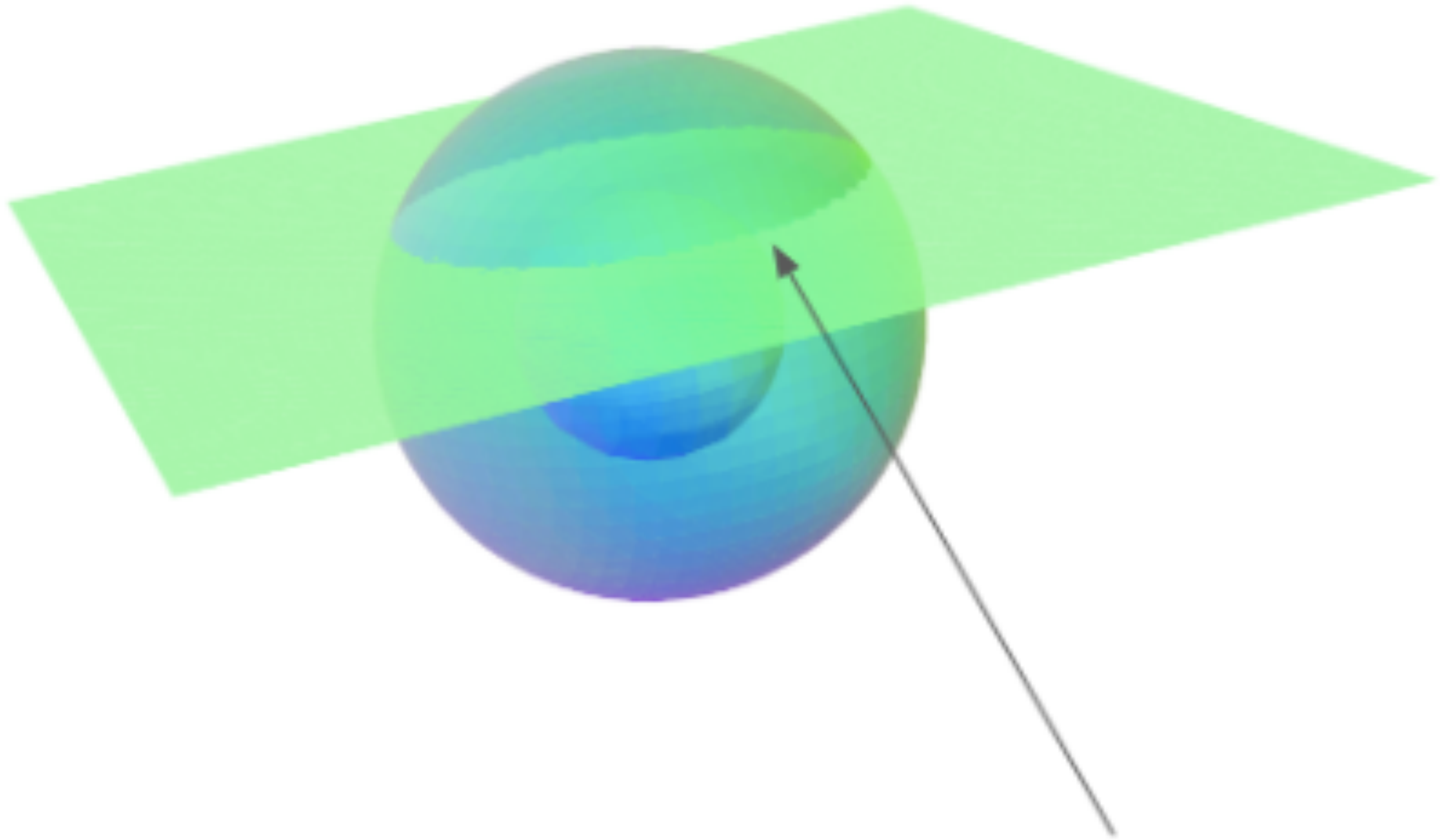
Uniform convergence of interpolators

- Worst-case error among all **interpolators** with low complexity.

$$\boxed{L(\hat{w}) \leq \sup_{w \in \mathcal{K}, \hat{L}(w)=0} L(w)}$$



$\{w : \|w\|_2 \leq B\}$



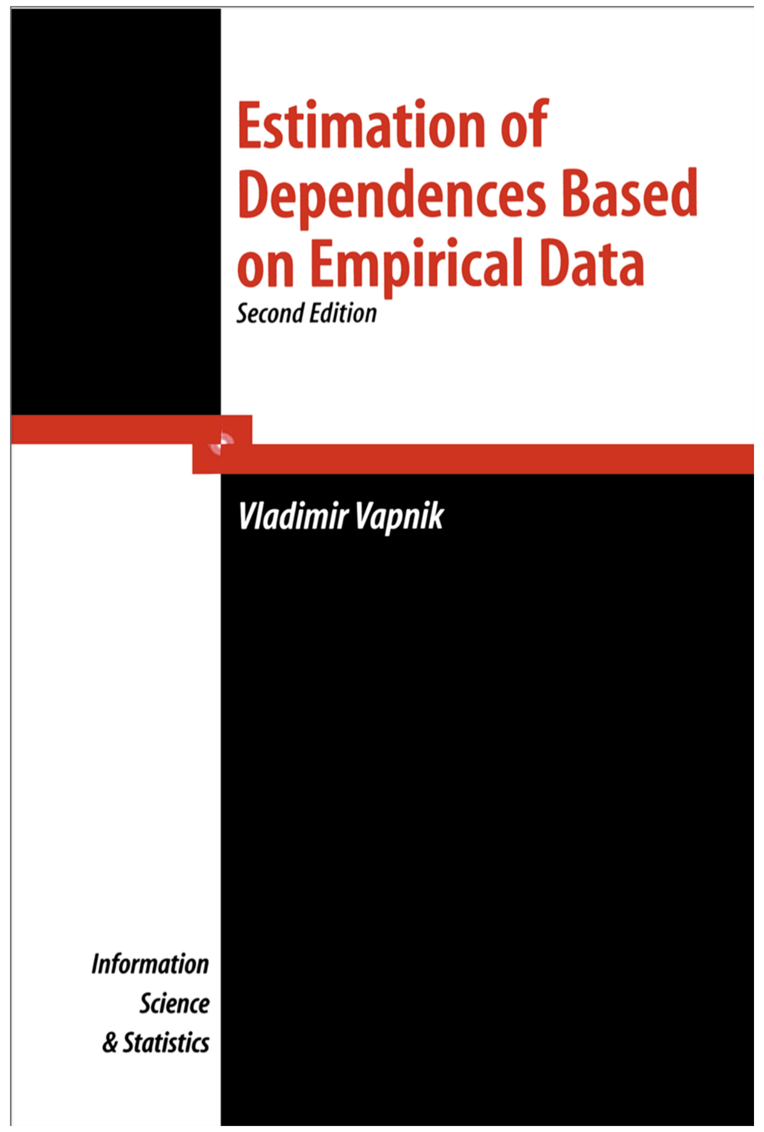
$\{w : \|w\|_2 \leq B, \hat{L}(w) = 0\}$

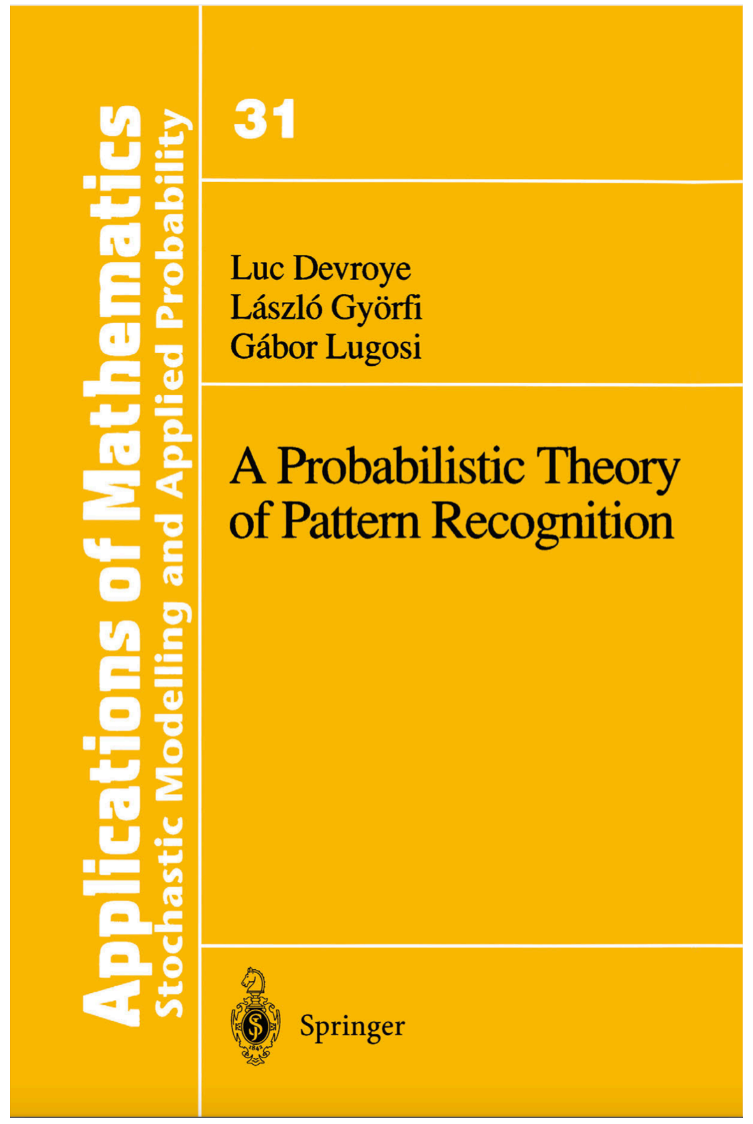
Uniform convergence of interpolators has been used in the noiseless setting since at least [Vapnik '82]. Below: [Devroye et al '96]

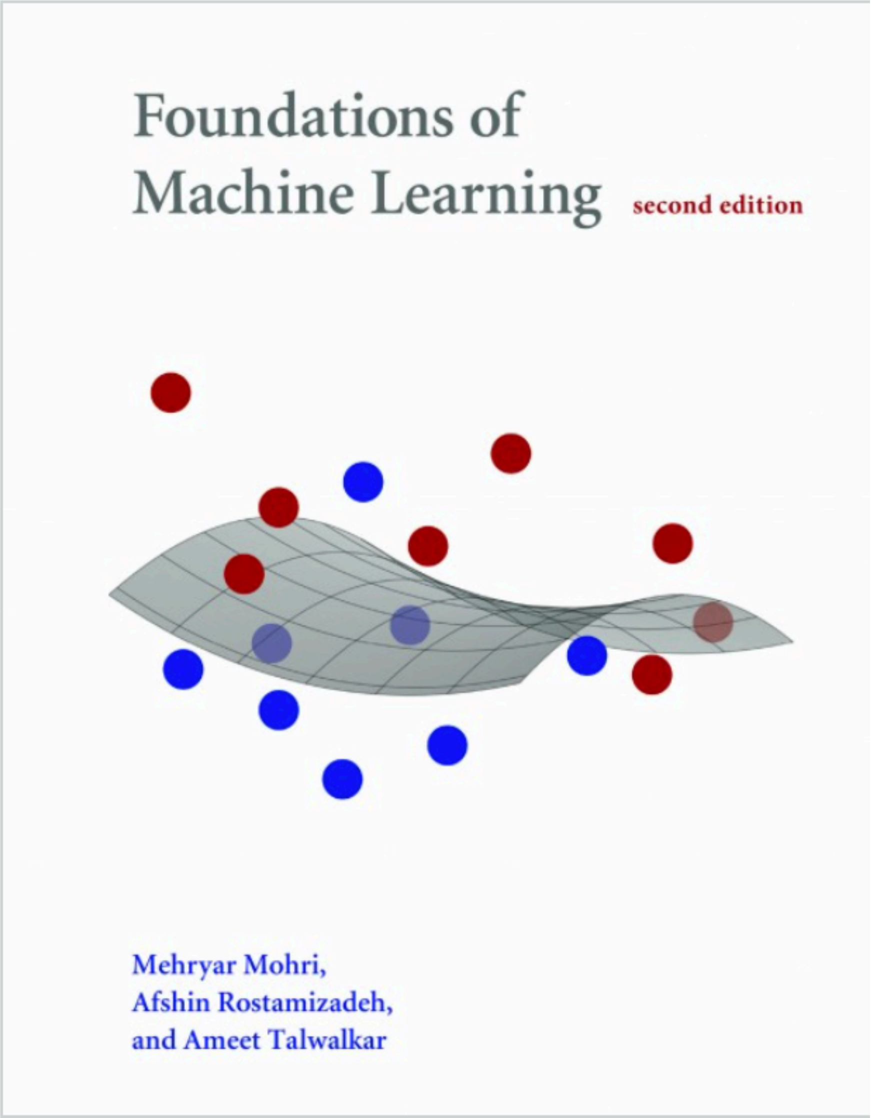
based on the random permutation argument developed in the original proof of the Vapnik-Chervonenkis inequality (1971).

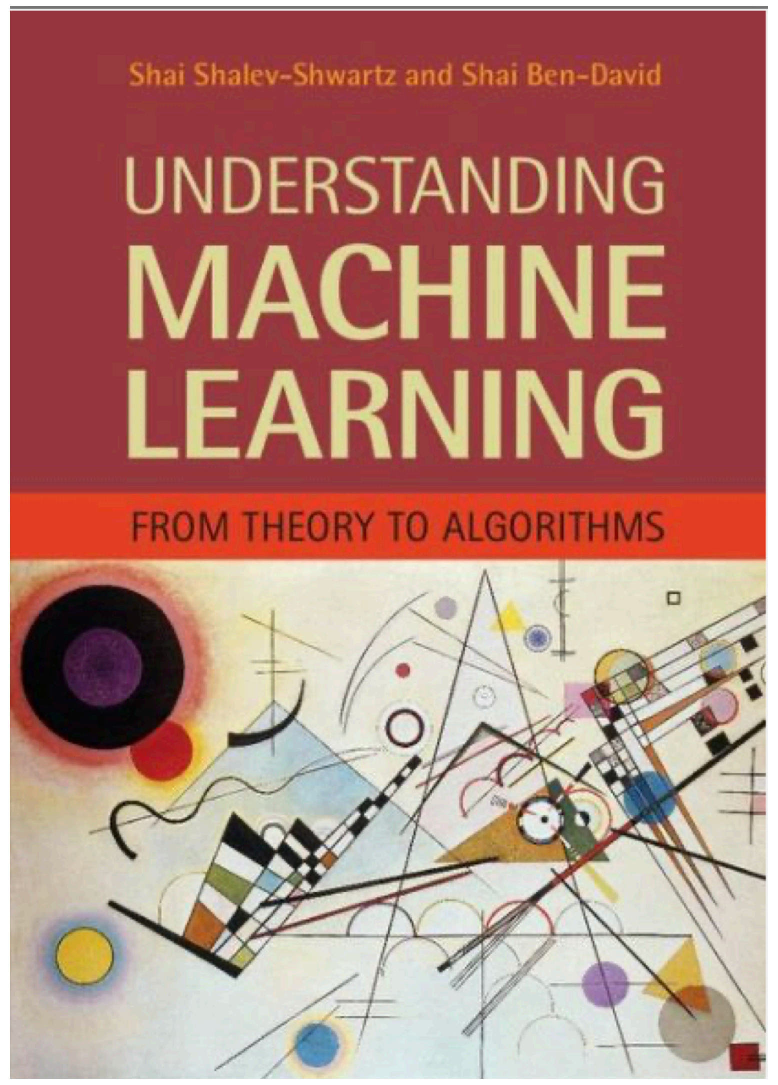
PROOF. For $n\epsilon \leq 2$, the inequality is clearly true. So, we assume that $n\epsilon > 2$. First observe that since $\inf_{\phi \in \mathcal{C}} L(\phi) = 0$, $\widehat{L}_n(\phi_n^*) = 0$ with probability one. It is easily seen that

$$L(\phi_n^*) \leq \sup_{\phi : \widehat{L}_n(\phi) = 0} |L(\phi) - \widehat{L}_n(\phi)|.$$









Conjecture [Zhou et al, 2020]:

$$\sup_{\|w\|_2 \le B, \hat{L}(w)=0} L(w) \le \frac{B^2 \mathbb{E}\|x\|^2}{n} + o(1)$$

- controlling the generalization error reduces to calculating the least amount of norm required to perfectly fit the data

- [Zhou et al, 2020]: In prototypical "junk features" model, proved conjecture and used to explain benign overfitting in this model.

In this paper we prove the conjecture for arbitrary Gaussian data as a special case of a more **general uniform convergence result** in terms of **Gaussian width**. Based on this, we recover the benign overfitting conditions of [Bartlett et al, 2019], and generalize them to arbitrary norms such as ℓ_1.

Main generalization bound

- **Gaussian width:** natural measure of "complexity" of a set, long used in generalization theory (e.g. [Bartlett-Mendelson, 2002])

$$\boxed{W(\mathcal{K}) = \mathbb{E}_{H\sim N(0,I_d)}\left[\sup_{w\in\mathcal{K}} |\langle H, w\rangle|\right]}$$

- Theorem (informal): for any covariance matrix $\Sigma = \mathbb{E}[xx^T]$, for any splitting $\Sigma = \Sigma_1 \oplus \Sigma_2$ such that $\mathrm{rank}(\Sigma_1) = o(n)$, it holds with high probability that

$$\boxed{\sup_{w\in\mathcal{K},\hat{L}(w)=0} L(w) \le (1+o(1)) \cdot \frac{W(\Sigma_2^{1/2}\mathcal{K})^2}{n}}$$

ℓ_2 norm ball: $\mathscr{K} = \{w : \|w\|_2 \leq B\}$

$$\boxed{\sup_{\|w\| \leq B, \hat{L}(w)=0} L(w) \leq (1 + o(1)) \frac{B^2 \mathbb{E}\|x\|^2}{n}}$$

because $W(\Sigma_2^{1/2}\mathscr{K}) = B \cdot \mathbb{E}_{H \sim N(0,I_d)} \|\Sigma_2^{1/2} H\| \leq \sqrt{B^2 \mathbb{E}\|x\|^2}$

- Confirms the prediction from [Zhou et al, 2020]
- Recovers the benign overfitting conditions of [Bartlett et al, 2019]

because we can prove $\|\hat{w}\|^2 \leq (1 + o(1)) \dfrac{\sigma^2 n}{\mathbb{E}_{x \sim N(0,\Sigma_2)} \|x\|^2}$

A new application: ℓ_1 norm ball

- What about regularizers besides ℓ_2?
  - ℓ_1 norm is key to LASSO, Adaboost, compressed sensing…
  - Not so easy to analyze (no closed form)! Is it consistent? [Ju et al, 2020].
- **Theorem (this work):** Minimum ℓ_1 norm interpolator (basis pursuit) is consistent in junk features model (small number of signal features, large number of small irrelevant "junk features"). Follows from general 'benign overfitting' conditions.

junk features: $\Sigma = \begin{bmatrix} I_{d_S} & 0 \\ 0 & \alpha I_{d_J} \end{bmatrix}, \qquad d_J \to \infty, \alpha \to 0$

Summary

- In linear regression, we showed via uniform convergence of interpolators that the norm, and more generally Gaussian width, controls generalization error of interpolators and explains benign overfitting.
  - Forthcoming work: extension to near-interpolators via “optimistic rates” theory
- **Why do we care about uniform convergence?**
  - unify classical statistical learning theory with modern practice in ML
  - can extend to settings where a direct analysis is difficult (ex: ℓ_1 interpolation) and highlight the “key” to good generalization (ex: low norm)
  - implicit regularization + uniform convergence can be a principled method to study more general overparameterized models, e.g. deep networks

Thanks for listening!