On Uniform Convergence and Low-Norm Interpolation Learning

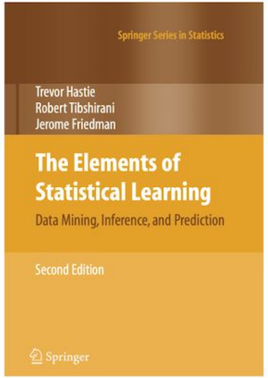
Lijia Zhou
UChicago

based on arXiv:2006.05942 (NeurIPS 2020)
with

Danica J. Sutherland
UBC

Nati Srebro
TTI-Chicago

Classical wisdom: “a model with zero training error is overfit to the training data and will typically generalize poorly”



- Interpolation Learning: achieving **low population error** while **training error is exactly zero** in a **noisy**, non-realizable setting

Low norm interpolation learning

- Implicit regularization in linear regression
  - Square loss objective $L_{\mathbf{S}}(w) = \frac{1}{n}\|Y - Xw\|^2$
  - When initialized at the origin, gradient descent finds minimal norm interpolator $\hat{w}_{MN} = \underset{w \in \mathbb{R}^p \text{ s.t. } Xw = Y}{\arg\min} \|w\|_2^2 = X^\top (XX^\top)^{-1} Y.$

- *Benign overfitting in linear regression* [Bartlett et al, 2019]
  - very nice results that tightly characterize the excess risk of $\hat{w}_{MN}$
  - but its analysis does not leverage the **minimal norm** aspect of it
  - is low norm really the key to good generalization?

Uniform convergence

- Pick a collection of hypotheses from which the learning rule outputs with high probability, and then show that the maximal generalization gap over this hypothesis class is small with high probability

$$\underbrace{L_{\mathcal{D}}(\hat{f})}_{>0} \leq \underbrace{L_{\mathbf{S}}(\hat{f})}_{0} + \sup_{f \in \mathcal{F}} |L_{\mathcal{D}}(f) - L_{\mathbf{S}}(f)|$$

Why uniform convergence?

- If uniform convergence (or some version of it) works
  - combined with implicit regularization can be a unified and principled method to study more complex overparameterized model
  - Is a naïve application enough? If not, what kind of modification is necessary?
  - the phenomenon of interpolation learning is “robust” - has practical implications
  - the techniques from a uniform-convergence type analysis may generalize to other interpolators

Why uniform convergence?

- If there is no way to make uniform convergence work even in this simple setting
  - Maybe it’s time to wholly abandon uniform convergence
  - Bad news for implicit regularization
    - Why try to find an implicit regularizer if the analysis has to depend crucially on the specific algebraic structure?

Challenge: getting the tight constant!

$$\underbrace{L_{\mathcal{D}}(\hat{f})}_{>0} \leq \underbrace{L_{\mathbf{S}}(\hat{f})}_{0} + \sup_{f \in \mathcal{F}} |L_{\mathcal{D}}(f) - L_{\mathbf{S}}(f)|$$

- In low dimensional settings, the generalization gap vanishes and the training error converges to Bayes risk
- **OK to have a constant factor** in the upper bound of generalization gap
- In high dimensional interpolation settings, the first term is zero so the generalization gap needs to converge *exactly* to the Bayes risk!

Negative results - I

- *Uniform convergence may be unable to explain generalization in deep learning* [Nagarajan and Kolter, 2019]
    - If we can identify a hypothesis class $\mathcal{H}$ from which the learning rule outputs with high probability, and the generalization gap over $\mathcal{H}$ is small with high probability, then there **exists** a collection of training sets $\mathcal{S}_\delta$ and if we consider only the outputs of learning rule $\mathcal{H}_\delta := \bigcup_{S \in \mathcal{S}_\delta} \{h_S\}$, it holds that the uniform generalization gap $\sup_{S \in \mathcal{S}_\delta} \sup_{h \in \mathcal{H}_\delta} \left| \mathcal{L}_\mathcal{D}(h) - \hat{\mathcal{L}}_S(h) \right|$ is small

Negative results - I

- *Uniform convergence may be unable to explain generalization in deep learning* [Nagarajan and Kolter, 2019]
  - Failure of algorithm-dependent uniform convergence
    - **Any** collection of typical training sets $\mathcal{S}_\delta$ has large uniform generalization gap, but the actual predictor found by gradient descent has small generalization gap (the quantity $\mathcal{L}_\mathcal{D}(h_S) - \hat{\mathcal{L}}_S(h_S)$ is small)
    - the empirical risk does not have to be evaluated on the training set that the algorithm uses to learn: $\sup_{S \in \mathcal{S}_\delta} \sup_{h \in \mathcal{H}_\delta} \left| \mathcal{L}_\mathcal{D}(h) - \hat{\mathcal{L}}_S(h) \right|$
  - limitation
    - bound has to consider **two sided** difference

Negative results - II

- *In defense of uniform convergence: generalization via derandomization with an application to interpolating predictors* [Negrea, Dziugaite and Roy, 2020]
  - For any sequence of $\mathcal{S}_{\frac{1}{3},n}$, as sample size tends to infinity, the expectation of the uniform generalization gap over the outputs of learning rule $\mathcal{H}_{\frac{1}{3},n}$ is at least 1.5 * the Bayes risk
  - two-sided uniform convergence is not sufficient to explain consistency of interpolation and standard symmetrization techniques cannot be directly applied to interpolation learning

Negative results - III

- *Failures of model-dependent generalization bounds for least-norm interpolation* [Bartlett and Long, 2021]
  - the *excess risk* of the learned minimal norm interpolator
  - any bound $R_P(h) - R_P^* \leq \epsilon(h, n, \delta)$ that
    - only depend on the learned hypothesis, sample size and confidence
    - satisfies certain anti-monotonicity condition in n
    - holds **uniformly** for all unit scale sub-gaussian distribution with high probability

  then there is a sequence of distribution on which the minimal norm interpolator is consistent, but for most n, the bound is bounded away from zero with constant probability

Negative results - III

- *Failures of model-dependent generalization bounds for least-norm interpolation* [Bartlett and Long, 2021]
    - allow distribution dependence only through the learned predictor; the only property that we know about the population is unit sub-Gaussianity, so it cannot capture bounds that adapt to
        - **noise levels** in the problem
        - the **empirical risk** of the learned predictor

Our setting

	“signal”, d_S	“junk”, $d_J \to \infty$
$\mathbf{x}$	$\mathbf{x}_S \sim \mathcal{N}\left(\mathbf{0}_{d_S}, \mathbf{I}_{d_S}\right)$	$\mathbf{x}_J \sim \mathcal{N}\left(\mathbf{0}_{d_J}, \frac{\lambda_n}{d_J}\mathbf{I}_{d_J}\right)$
$\mathbf{w}^*$	$\mathbf{w}_S^*$	$\mathbf{0}$

$$y = \underbrace{\langle \mathbf{x}, \mathbf{w}^* \rangle}_{\langle \mathbf{x}_S, \mathbf{w}_S{}^* \rangle} + \mathcal{N}(0, \sigma^2)$$

- Minimal norm interpolator is consistent [Bartlett et al, 2019]
- The prediction of minimal norm interpolator on new samples is asymptotically equivalent to ridge regression using only the signal part
  - New “junk” is asymptotically almost sure orthogonal to the old “junk”
  - Signal part converge to ridge regression estimate

Our setting

	"signal", d_S	"junk", $d_J \to \infty$
$\mathbf{x}$	$\mathbf{x}_S \sim \mathcal{N}\left(\mathbf{0}_{d_S}, \mathbf{I}_{d_S}\right)$	$\mathbf{x}_J \sim \mathcal{N}\left(\mathbf{0}_{d_J}, \frac{\lambda_n}{d_J}\mathbf{I}_{d_J}\right)$
$\mathbf{w}^*$	$\mathbf{w}_S^*$	$\mathbf{0}$

$$y = \underbrace{\langle \mathbf{x}, \mathbf{w}^* \rangle}_{\langle \mathbf{x}_S, \mathbf{w}_S{}^* \rangle} + \mathcal{N}(0, \sigma^2)$$

- The prediction of minimal norm interpolator on new samples is asymptotically equivalent to ridge regression using only the signal part
  - as long as the bias introduced by regularization is negligible, ridge regression estimate is consistent
  - interchanging limit and expectation yields consistency

Our negative results

- could we have discovered consistency via uniform convergence?
  - Rademacher bounds assume Lipschitz loss, which does not hold for square loss on unbounded domain
- **NO!**
  - generalization gap over even the smallest norm ball that contains the minimal norm interpolator diverges:

<u>Theorem:</u> If $\lambda_n = o(n)$,

$$\lim_{n\to\infty} \lim_{d_J\to\infty} \mathbb{E}\left[\sup_{\|\mathbf{w}\| \le \|\hat{\mathbf{w}}_{MN}\|} |L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w})| \right] = \infty.$$

Proof sketch

- decompose generalization gap as

$$L_{\mathcal{D}}(w) - L_{\mathbf{S}}(w) = \left[L_{\mathcal{D}}(w^*) - \frac{\|E\|^2}{n}\right] + (w - w^*)^{\top}(\Sigma - \hat{\Sigma})(w - w^*) + 2\left\langle w - w^*, \frac{X^{\top}E}{n}\right\rangle$$

- from above, obtain the lower bound

$$\sup_{\|w\| \leq \|\hat{w}_{MN}\|} |L_{\mathcal{D}}(w) - L_{\mathbf{S}}(w)| \geq \|\Sigma - \hat{\Sigma}\| \cdot (\|\hat{w}_{MN}\| - \|w^*\|)^2 - \left|L_{\mathcal{D}}(w^*) - \frac{\|E\|^2}{n}\right|$$

$$\Theta\left(\sqrt{\frac{\lambda_n}{n}}\right) \qquad \Theta\left(\frac{n}{\lambda_n}\right)$$

Koltchinskii/Lounici, Bernoulli 2017

Beyond norm balls and minimal 2-norm interpolator

- there is no fixed hypothesis class that we can choose to prove consistency & holds for all natural interpolator $\mathcal{A}((X_S, X_J), y)_S = \mathcal{A}((X_S, -X_J), y)_S$

Theorem (à la [Nagarajan/Kolter, NeurIPS 2019]):

For each $\delta \in (0, \frac{1}{2})$, let $\mathbf{Pr}(\mathbf{S} \in \mathcal{S}_{n,\delta}) \geq 1 - \delta$,

$\hat{\mathbf{w}}$ a *natural* consistent interpolator,

and $\mathcal{W}_{n,\delta} = \{\hat{\mathbf{w}}(\mathbf{S}) : \mathbf{S} \in \mathcal{S}_{n,\delta}\}$. Then, almost surely,

$$\lim_{n\to\infty} \lim_{d_J\to\infty} \sup_{\mathbf{S}\in\mathcal{S}_{n,\delta}} \sup_{\mathbf{w}\in\mathcal{W}_{n,\delta}} |L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w})| \geq 3\sigma^2 .$$

Proof sketch

- For each $\mathbf{S} = (X, Y) \in \mathcal{S}_{n,\delta}$
  - consider $\tilde{\mathbf{S}} = ((X_S, -X_J), Y)$ and $\tilde{w} = \mathcal{A}(\tilde{\mathbf{S}})$
  - when there is no signal part, have $-X\tilde{w} = Y \implies X\tilde{w} = -Y$

    so we have $L_{\mathbf{S}}(\tilde{w}) = \frac{1}{n}\|(-Y) - Y\|^2 = \frac{4}{n}\|Y\|^2 \overset{a.s.}{=} 4\sigma^2$.

  - The general case can be handled by an orthogonal projection

Lessons from these negative results

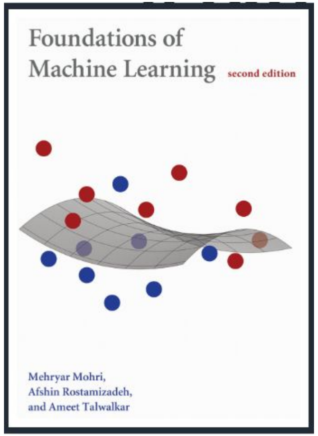
- In order to get consistency, we need to consider
  - one sided uniform convergence, or
  - some “localized” version of uniform convergence that doesn’t pay attention to cases with high empirical risk
- this phenomenon seem to extend beyond linear regression and the minimal norm interpolator
- Small norm is not sufficient for generalization generally
  - but is it sufficient in the context of interpolation?
  - uniform convergence of **zero-error** *predictor*

$$\sup_{\|\mathbf{w}\|\leq B,\ L_{\mathbf{S}}(\mathbf{w})=0} |L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w})|$$

Uniform convergence of zero-error predictor

- Is this uniform convergence?

$$\sup_{\|\mathbf{w}\| \le B,\ L_{\mathbf{S}}(\mathbf{w})=0} |L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w})|$$



In the example of axis-aligned rectangles that we examined, the hypothesis h_S returned by the algorithm was always *consistent*, that is, it admitted no error on the training sample S. In this section, we present a general sample complexity bound, or equivalently, a generalization bound, for consistent hypotheses, in the case where the cardinality $|H|$ of the hypothesis set is finite. Since we consider consistent hypotheses, we will assume that the target concept c is in H.

Theorem 2.1 **Learning bounds — finite H, consistent case**
Let H be a finite set of functions mapping from $\mathcal{X}$ to $\mathcal{Y}$. Let $\mathcal{A}$ be an algorithm that for any target concept $c \in H$ and i.i.d. sample S returns a consistent hypothesis h_S: $\widehat{R}(h_S) = 0$. Then, for any $\epsilon, \delta > 0$, the inequality $\Pr_{S \sim \mathcal{D}^m}[R(h_S) \le \epsilon] \ge 1 - \delta$ holds if

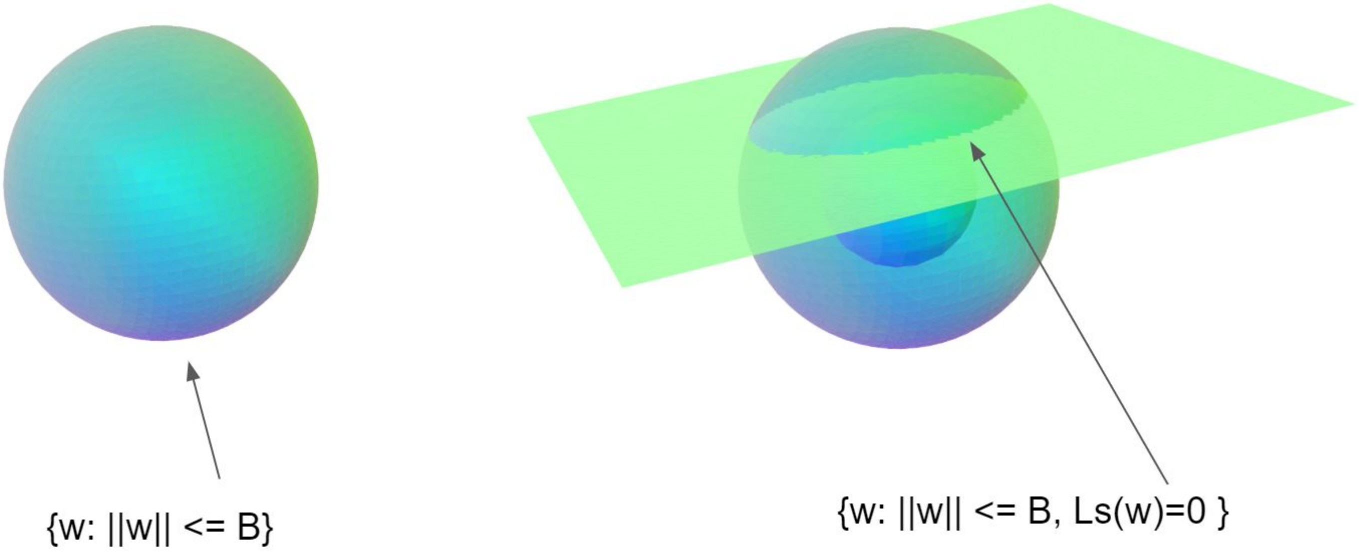
$$m \ge \frac{1}{\epsilon}\Big(\log|H| + \log\frac{1}{\delta}\Big). \tag{2.8}$$

This sample complexity result admits the following equivalent statement as a generalization bound: for any $\epsilon, \delta > 0$, with probability at least $1 - \delta$,

$$R(h_S) \le \frac{1}{m}\Big(\log|H| + \log\frac{1}{\delta}\Big). \tag{2.9}$$

Proof Fix $\epsilon > 0$. We do not know which consistent hypothesis $h_S \in H$ is selected by the algorithm $\mathcal{A}$. This hypothesis further depends on the training sample S. Therefore, we need to give a *uniform convergence bound*, that is, a bound that holds for the set of all consistent hypotheses, which a fortiori includes h_S. Thus,

Visualization of “interpolating” hypothesis class



How to analyze this generalization gap?

- By a change of variable, the generalization gap equals

$$L_{\mathcal{D}}(\mathbf{w}^*) + \sup_{\mathbf{z}:\|\hat{\mathbf{w}}+\mathbf{F}\mathbf{z}\|^2 \le B^2} (\hat{\mathbf{w}} + \mathbf{F}\mathbf{z} - w^*)^\top \mathbf{\Sigma} (\hat{\mathbf{w}} + \mathbf{F}\mathbf{z} - w^*)$$

- $\hat{\mathbf{w}}$ is any interpolator, i.e. Xw = Y
  the columns of F form an orthonormal basis for ker(X)
- expanding the quadratic term, we can decompose
  - generation gap = risk of surrogate interpolator + gap to worst interpolator
- the gap is formulated as a Quadratically Constrained Quadratic Program (QCQP)
  - can be analyzed easily by its dual
  - strong duality holds for QCQP with single constraint without convexity assumption

Some definitions

- Restricted eigenvalue under interpolation

$$\kappa_{\mathbf{X}}(\mathbf{\Sigma}) = \sup_{\|\mathbf{w}\|=1,\ \mathbf{X}\mathbf{w}=\mathbf{0}} \mathbf{w}^{\top}\mathbf{\Sigma}\mathbf{w}$$

- Minimal risk interpolator
  - best interpolator possible, but cannot be computed in practice
  - useful for theoretical analysis to show lower bound & upper bound

$$\hat{\mathbf{w}}_{MR} = \operatorname*{argmin}_{\mathbf{w}:\mathbf{X}\mathbf{w}=\mathbf{y}} L_{\mathcal{D}}(\mathbf{w})$$

Two general results

- Strategy: decompose generation gap as risk of a surrogate interpolator + gap to worst interpolator
    - with minimal risk interpolator

$$\sup_{\substack{\|\mathbf{w}\| \leq \|\hat{\mathbf{w}}_{MR}\| \\ L_{\mathbf{S}}(\mathbf{w})=0}} L_{\mathcal{D}}(\mathbf{w}) = L_{\mathcal{D}}(\hat{\mathbf{w}}_{MR}) + \underbrace{\beta}_{1 \leq \beta \leq 4} \kappa_X(\Sigma) \left[\|\hat{\mathbf{w}}_{MR}\|^2 - \|\hat{\mathbf{w}}_{MN}\|^2 \right]$$

    - with minimal norm interpolator

$$\sup_{\substack{\|\mathbf{w}\| \leq \alpha\|\hat{\mathbf{w}}_{MN}\| \\ L_{\mathbf{S}}(\mathbf{w})=0}} L_{\mathcal{D}}(\mathbf{w}) = L_{\mathcal{D}}(\hat{\mathbf{w}}_{MN}) + (\alpha^2 - 1)\, \kappa_X(\Sigma)\, \|\hat{\mathbf{w}}_{MN}\|^2 + R_n$$

$R_n \to 0$ if $\hat{\mathbf{w}}_{MN}$ is consistent

In our case...

- norm calculation

$$\lim_{d_J \to \infty} \mathbb{E}\|\hat{w}_{MR}\|^2 = \|w^*\|^2 + \frac{\sigma^2 n}{\lambda_n}$$

$$\lim_{d_J \to \infty} \mathbb{E}\|\hat{w}_{MN}\|^2 = \|w^*\|^2 + \sigma^2 \frac{n - d_S}{\lambda_n} + \beta_n \left(\frac{\sigma^2 d_S - \lambda_n \|w_S^*\|^2}{n} \right)$$

- restricted eigenvalue

$$\lim_{d_J \to \infty} \kappa_X(\Sigma) = \frac{\lambda_n}{n} \left\| \left[\frac{X_S^\intercal X_S}{n} + \frac{\lambda_n}{n} I_{d_S} \right]^{-1} \right\|$$

- Consistency of minimal risk interpolator

$$\mathbb{E}\, L_{\mathcal{D}}(\hat{w}_{MR}) = \frac{p-1}{p-1-n} L_{\mathcal{D}}(w^*)$$

Plugging in...

$$1 \leq \beta \leq 4$$

$$\sup_{\substack{\|\mathbf{w}\| \leq \|\hat{\mathbf{w}}_{MR}\| \\ L_{\mathbf{S}}(\mathbf{w})=0}} L_{\mathcal{D}}(\mathbf{w}) = L_{\mathcal{D}}(\hat{\mathbf{w}}_{MR}) + \beta\,\kappa_X(\Sigma)\left[\|\hat{\mathbf{w}}_{MR}\|^2 - \|\hat{\mathbf{w}}_{MN}\|^2\right]$$

$$L_{\mathcal{D}}(w^*) \qquad \frac{\lambda_n}{n} \qquad \frac{\sigma^2 d_S}{\lambda_n}$$

- can conclude

$$\lim_{n\to\infty}\lim_{d_J\to\infty}\mathbb{E}\left[\sup_{\|w\|\leq\|\hat{w}_{MR}\|,\, L_{\mathbf{S}}(w)=0} L_{\mathcal{D}}(w) - L_{\mathbf{S}}(w)\right] = L_{\mathcal{D}}(w^*)$$

Plugging in...

$$L_{\mathcal{D}}(w^*) \qquad \frac{\lambda_n}{n} \qquad \sigma^2 \frac{n - d_S}{\lambda_n}$$

$$\sup_{\substack{\|\mathbf{w}\| \leq \alpha \|\hat{\mathbf{w}}_{MN}\| \\ L_{\mathbf{S}}(\mathbf{w})=0}} L_{\mathcal{D}}(\mathbf{w}) = L_{\mathcal{D}}(\hat{\mathbf{w}}_{MN}) + (\alpha^2 - 1)\,\kappa_X(\Sigma)\,\|\hat{\mathbf{w}}_{MN}\|^2 + R_n$$

$R_n \to 0$ if $\hat{\mathbf{w}}_{MN}$ is consistent

- can conclude

$$\lim_{n\to\infty} \lim_{d_J\to\infty} \mathbb{E}\left[\sup_{\|w\| \leq \alpha_n \|\hat{w}_{MN}\|,\ L_{\mathbf{S}}(w)=0} L_{\mathcal{D}}(w) - L_{\mathbf{S}}(w) \right] = \alpha^2 L_{\mathcal{D}}(w^*)$$

Some observations...

- Speculative bound

$$\sup_{\|w\|\leq B,\, L_{\mathbf{S}}(w)=0} L_{\mathcal{D}}(w) - L_{\mathbf{S}}(w) \leq \frac{1}{n} B^2 \xi_n + o_P(1)$$

- Rademacher complexity

$$\mathfrak{R}_n(\mathcal{W}_B) = \mathbb{E}_{\mathbf{S}}\, \mathbb{E}_{\sigma\sim\mathrm{Unif}(\pm 1)^n} \sup_{w:\|w\|\leq B} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \langle w, x^{(i)}\rangle \leq \sqrt{\frac{1}{n} B^2\ \mathbb{E}\|x\|^2}$$

$$\lim_{d_J\to\infty} \frac{(\mathbb{E}\|\hat{w}_{MN}\|^2)(\mathbb{E}\|x\|^2)}{n} = \sigma^2 + o(1)$$

Optimistic rate

- Risk dependent bound for smooth loss

Applying [Srebro/Sridharan/Tewari 2010]: for all $\|\mathbf{w}\| \leq B$,

ξ_n: high-prob bound on $\max_{i=1,...,n} \|\mathbf{x}_i\|^2$

$$L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w}) \leq \tilde{\mathcal{O}}_P\left(\frac{B^2\xi_n}{n} + \sqrt{L_{\mathbf{S}}(\mathbf{w})\frac{B^2\xi_n}{n}}\right)$$

- Issue: hidden factor on $\frac{B^2\xi_n}{n}$ of $c \leq 200,000 \log^3(n)$
- If we can get c = 1, it would imply speculative bound and can quantify how much population risk degrade if we don't optimize to exact zero error

Summary

- Uniformly bounding the difference between empirical and population errors cannot show any learning in the norm ball
- Uniform convergence over any set, even one depending on the exact algorithm and distribution, cannot show consistency
- but we show that an "interpolating" uniform convergence bound does
  - show low norm is sufficient for interpolation learning in our testbed problem; near minimal norm interpolator can also achieve consistency!
  - predict exact worst-case error as norm grows
- when applying uniform convergence in the context of interpolation learning, need to consider optimistic-rate, or risk dependent type of bound