On Uniform Convergence and Low-Norm Interpolation Learning

Lijia Zhou

UChicago

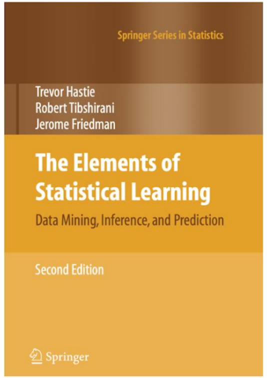
Joint work with:

D.J. Sutherland
TTI-Chicago -> UBC

Nati Srebro
TTI-Chicago

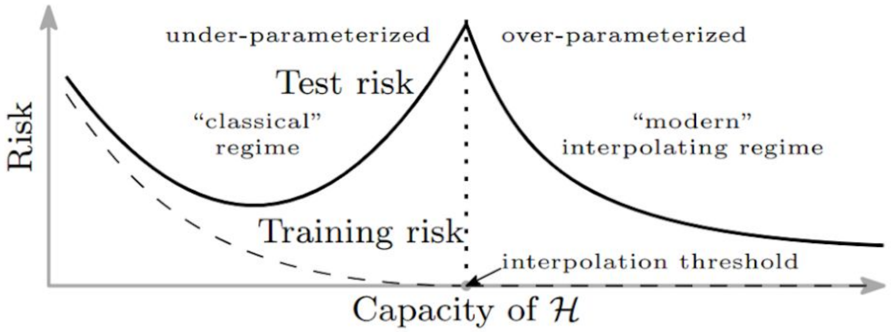
Interpolation learning

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
- Related to “double descent” (Belkin et al, 2018)



Interpolation learning

- Recent works of interpolation learning are not based on uniform convergence. Can interpolation learning be explained by uniform convergence?

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

- Want the left hand side to converge to the Bayes optimal risk
- Uniform convergence may be unable to explain generalization in deep learning (Nagarajan and Kolter, 2019)

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, training error converges to Bayes risk and the generalization gap vanishes
- **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!

Can we show consistency of **interpolators** in noisy settings with **uniform convergence**?

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

Answer: For fixed $\mathcal{F}$, **NO**.

But **YES** if $\mathcal{F}$ *only contains interpolating predictors*!

Our testbed problem

- a specific high dimensional linear regression problem with “junk” features

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

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

- Low norm interpolation learning: minimal l2 norm interpolator

$$\hat{w}_{MN} = \underset{w \in \mathbb{R}^p \text{ s.t. } Xw = Y}{\arg\min} \|w\|_2^2 = X^\mathsf{T}(XX^\mathsf{T})^{-1}Y.$$

- We are only going to worry about consistency in expectation

$$\mathbb{E}[L_\mathcal{D}(\hat{f}) - L_\mathcal{D}(f^*)] \to 0$$

Negative results

- l2 norm ball

Theorem: If $\lambda_n = o(n)$,

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

Negative results

- l2 norm ball

Theorem: If $\lambda_n = o(n)$,

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

- what about other hypothesis classes?

Theorem (à la [Nagarajan/Kolter, NeurIPS 2019]):
For each $\delta \in (0, \frac{1}{2})$, let $\mathbf{Pr}\left(\mathbf{S} \in \mathcal{S}_{n,\delta}\right) \geq 1 - \delta$,
$\hat{\mathbf{w}}$ a *natural* consistent interpolator,
and $\mathcal{W}_{n,\delta} = \{\hat{\mathbf{w}}(\mathbf{S}) : \mathbf{S} \in \mathcal{S}_{n,\delta}\}$. Then, almost surely,

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

Uniform convergence may be unable to explain generalization in deep learning

Vaishnavh Nagarajan
Department of Computer Science
Carnegie Mellon University
Pittsburgh, PA
vaishnavh@cs.cmu.edu

J. Zico Kolter
Department of Computer Science
Carnegie Mellon University &
Bosch Center for Artificial Intelligence
Pittsburgh, PA
zkolter@cs.cmu.edu

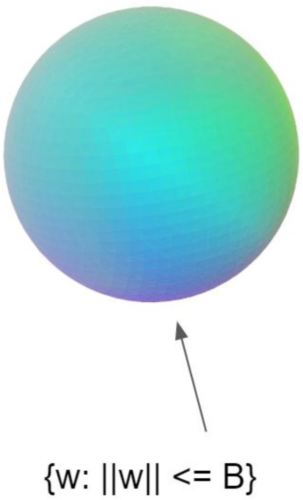
Positive results

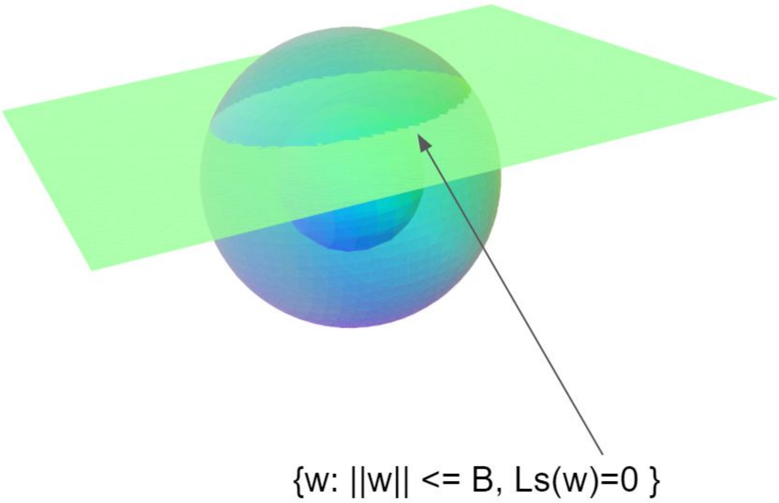
- Uniform convergence of *zero-error predictor*

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

Positive results

- Uniform convergence of *zero-error predictor* $$\sup_{\|\mathbf{w}\| \le B,\ L_{\mathbf{S}}(\mathbf{w})=0} |L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w})|$$
- Visualization of the hypothesis class:



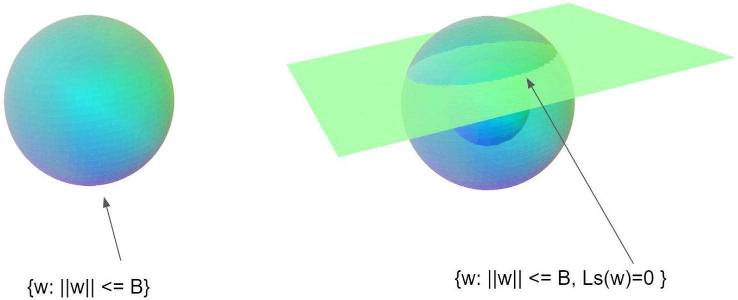


Positive results

- Uniform convergence of *zero-error predictor*

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

- Visualization of the hypothesis class:



- Intersection between norm ball and interpolation hyperplane

Theorem: If $\lambda_n = o(n)$,

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

Some **low-norm non-interpolators** don't generalize

Some **high-norm interpolators** don't generalize

All **low-norm interpolators** generalize

The combination is vital!

Speculative bound

- This result would be implied by a general result like

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

  with an appropriate choice of complexity measure ξ_n

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

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

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

- Issue: hidden factor on $\frac{B^2 \xi_n}{n}$ of $c \le 200{,}000 \log^3(n)$

Key observation for proofs

Can change variables in $\sup_{\mathbf{w}:\|\mathbf{w}\|\le B,\, L_{\mathbf{S}}(\mathbf{w})=0} L_{\mathcal{D}}(\mathbf{w})$ to

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

- The columns of F form an orthonormal basis for ker(X), where X is the design matrix
- $\hat{\mathbf{w}}$ is any interpolator, i.e. Xw = Y
- This is a Quadratically Constrained Quadratic Program (QCQP)
- Strong duality holds for QCQP with single constraint without any assumption on Sigma

Tools

- Decompose generation gap = risk of surrogate interpolator + its gap to worst interpolator

- 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)

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

Two general results

- Picking the surrogate to be minimal risk interpolator

get without **any** distributional assumptions that

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

(amount of missed energy) · (available norm)

- Picking the surrogate to be 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

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
- Analyzing generalization gap via duality may be broadly applicable