On Uniform Convergence and Low-Norm Interpolation Learning

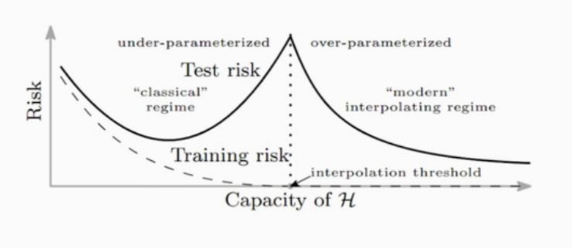
Jinwon Park

January 17, 2022

Seoul National University

Introduction

- Classical belief
  - 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)



Notation

- iid observations $(x_1, y_1), \cdots, (x_n, y_n) \sim \mathcal{D}^n$ where $\mathcal{D}$ is given by,
  - $x \in \mathbb{R}^p$ is drawn from $\mathcal{N}(0, \Sigma)$ with $\Sigma \succ 0$ and $\epsilon \in \mathbb{R}$ from $\mathcal{N}(0, \sigma^2)$
  - There is some fixed $\omega^* \in \mathbb{R}^p$ such that $y = \langle \omega^*, x \rangle + \epsilon$
- consider a "junk features" setting, where x decomposes into "signal" and "junk" components
  - let $\Sigma = \begin{bmatrix} I_{d_s} & 0_{d_S \times d_J} \\ 0_{d_J \times d_S} & \frac{\lambda_n}{d_J} I_{d_J} \end{bmatrix}$ where $d_S + d_J = p$ and $\lambda_n > 0$
  - In other words, $x = (x_S, x_J)$, where $x_S \sim \mathcal{N}(0, I_{d_S})$ and $x_J \sim \mathcal{N}(0, \frac{\lambda_n}{d_J} I_{d_J})$
  - Further the label depends only on x_S : $\omega^* = (\omega_S^*, 0_{d_J})$ with $\omega_S^* \in \mathbb{R}^{d_S}$
- The population risk and empirical risk are,

$$L_{\mathcal{D}}(\omega) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[(y - \langle \omega, x \rangle)^2] = L_{\mathcal{D}}(\omega^*) + \|\omega - \omega^*\|_{\Sigma}^2$$

$$L_S(\omega) = \frac{1}{n}\|Y - X\omega\|^2 = L_S(\omega^*) + \|\omega - \omega^*\|_{\hat{\Sigma}}^2 - \frac{2}{n}\langle X^T E, \omega - \omega^* \rangle$$

Interpolation Learning

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

$$L_{\mathcal{D}}(\hat{f}) \leq L_S(\hat{f}) + \sup_{f \in \mathcal{F}} |L_{\mathcal{D}}(f) - L_S(f)|$$

- 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)

Interpolation Learning

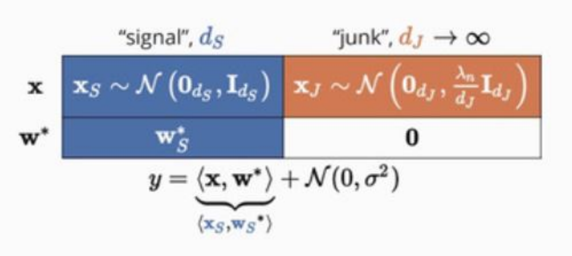
- In low dimensional settings, training error converges to Bayes risk and the generalization gap vanishes
- In high dimensional interpolation settings, the first term is zero so the generalization gap needs to converge exactly to the Bayes risk!
- Can we show consistency of interpolators in noisy settings with uniform convergence?

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



- Low norm interpolation learning: minimal l2 norm interpolator

$$\hat{\omega}_{MN} = \underset{\omega \in \mathbb{R}^p \text{ s.t. } X\omega = Y}{\operatorname{argmin}} \|\omega\|_2^2 = X^T(XX^T)^{-1}Y$$

- The paper only cares about consistency in expectation

$$\mathbb{E}[L_{\mathcal{D}}(\hat{\omega}_{MN}) - L_D(\omega^*)] \to 0$$

Negative results

- l2 norm ball

Theorem: If $\lambda = o(n)$

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

- what about other hypothesis classes?

Theorem: Nagarajan, Kolter, NeurIPS 2019[a]

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

$\hat{\omega}$ a natural consistent interpolator, and $\mathcal{W}_{n,\delta} = \{\hat{\omega}(S) : S \in \mathcal{S}_{n,\delta}\}$

Then, almost surely,

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

[a] Uniform convergence may be unable to explain generalization in deep learning

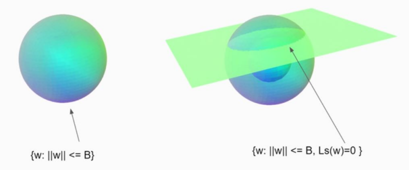
[a] Uniform convergence may be unable to explain generalization in deep learning

Positive results

- Uniform convergence of zero-error predictor

$$\sup_{\|\omega\| \leq B, L_S(\omega)=0} |L_{\mathcal{D}}(\omega) - L_S(\omega)|$$

- Visualization of the hypothesis class:



- Intersection between norm ball and interpolation hyperplane

Theorem: if $\lambda_n = o(n)$, fix a sequence $(\alpha_n) \to \alpha$ with each $\alpha_n \geq 1$, then

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

Positive results

Some low-norm non-interpolators do not generalize

Some high-norm interpolators do not generalize

All low-norm interpolators generalize, hence the combination is vital!

Speculative bound

- This result would be implied by a general result like

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

  with an appropriate choice of complexity measure ξ_n[b]
- Optimistic rate:

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

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

[b] ξ_n: high-prob bound on $\max_{i=1,\ldots,n} \|x_i\|^2$

[c] Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. "Optimistic Rates for Learning with a Smooth Loss" (2010) arXiv: 1009.3896.

- Decomposes generation gap (=risk) of surrogate interpolator + its gap to worst interpolator
- Restricted eigenvalue under interpolation

$$\kappa_X(\Sigma) = \sup_{\|\omega\|=1, X\omega=0} \omega^T \Sigma \omega$$

- Minimal risk interpolator (best interpolator possible, but cannot be computed in practice)

$$\hat{\omega}_{MR} = \underset{\omega: X\omega=y}{\operatorname{argmin}} L_{\mathcal{D}}(\omega)$$

Two general results

- Picking the surrogate to be minimal risk interpolator

get without any distributional assumptions that

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

- Picking the surrogate to be minimal norm interpolator

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

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 the paper shows that an "interpolating" uniform convergence bound does;
  - show low norm is sufficient for interpolation learning in testbed problem; near minimal norm interpolator an also achieve consistency
  - predict exact worst-case error as norm grows
- Analyzing generalization gap via duality may be broadly applicable