On Uniform Convergence and Low-Norm Interpolation Learning

Lijia Zhou, *University of Chicago*
D.J. Sutherland, *TTI-Chicago*
Nathan Srebro, *TTI-Chicago*

NEURAL INFORMATION PROCESSING SYSTEMS

Overview

- The phenomenon of **Interpolation learning** - achieving low population error while training error is exactly zero in a noisy, non-realizable setting, is one of the core mysteries in deep learning
- **Uniform convergence** is the fundamental technique used in learning theory:

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

- In this work, we investigate whether uniform convergence is sufficient to explain the success of **minimal norm interpolator**, the solution found by gradient descent methods, in an underdetermined noisy linear regression model.

Summary of results

We show in our testbed problem that

- 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 **uniform convergence of *zero-error predictors*** in the norm ball is **sufficient** to explain interpolation learning
- moreover, uniform convergence shows that ***near* minimal** norm interpolators can also achieve consistency and it can predict the **exact worse-case error** as norm grows

Setting

High dimension linear regression with “junk” features

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

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

Minimal norm interpolator

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

- Equivalent to ridge regression on signals, hence consistent

Negative result in the norm ball

The generalization gap over even the *smallest* norm ball that contains the minimal norm interpolator **diverges** in the asymptotic regime where consistency is possible

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

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

Negative result in algorithm-dependent sets

- We show that there is **no** algorithm-dependent hypothesis class that we can use to prove consistency

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

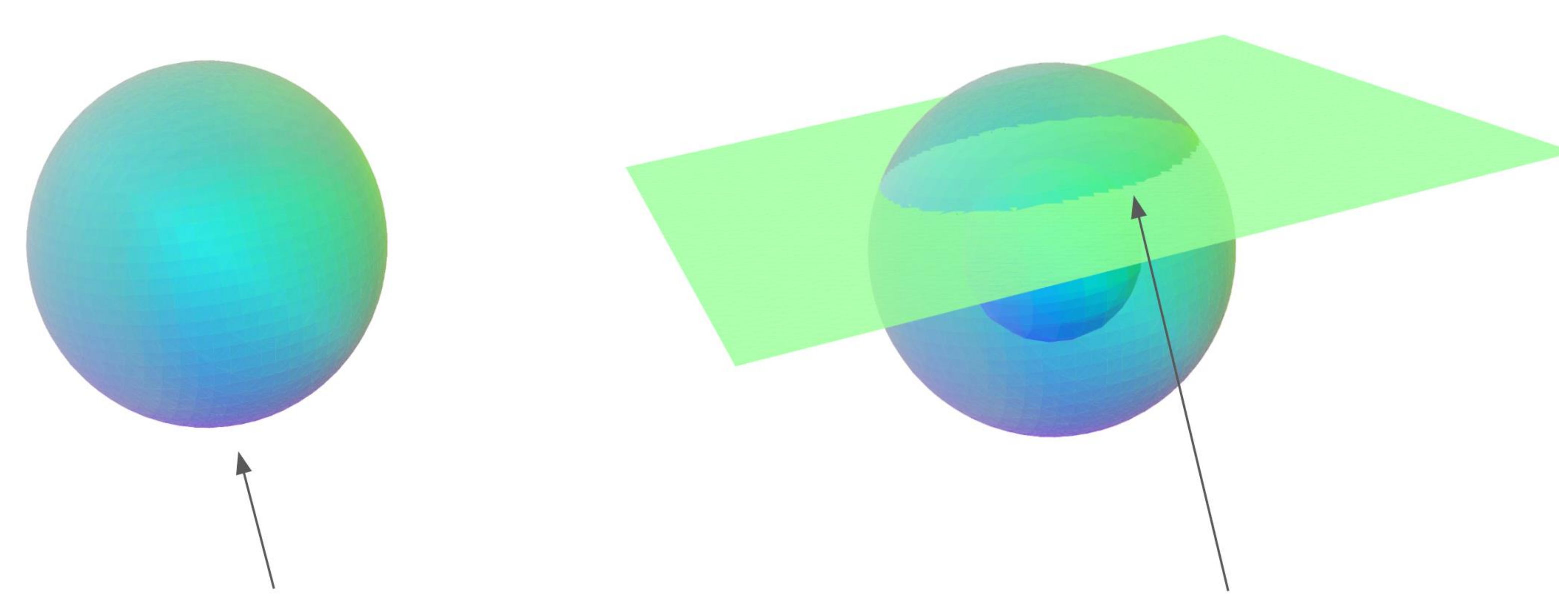
$$\lim_{n\to\infty} \lim_{d_J\to\infty} \sup_{\mathbf{S}\in\mathcal{S}_{n,\delta}} \sup_{\mathbf{w}\in\mathcal{W}_{n,\delta}} |L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w})| \ge 3\sigma^2.$$

- This is the **tightest** notion of uniform convergence if the hypothesis class is not allowed to depend on the training samples.
- Similar result can be found in [Negrea et al. 2020]
- We show this not only for minimal l2-norm interpolation, but **for *all “natural” consistent interpolators*** such as the minimal l1-norm interpolator

$$\mathcal{A}((X_S, X_J), y)_S = \mathcal{A}((X_S, -X_J), y)_S$$

Uniform convergence of *zero-error* predictors

Visualization of hypothesis class



Positive result with “interpolating” UC

- Consider all ***interpolating*** predictors with small norm (α times minimum norm)
- Sup is over intersection of norm ball with (sample-dependent) interpolation hyperplane
- Get exact risk of worst interpolator in ball: α^2 times Bayes risk

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

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

As the minimal norm diverges, the theorem shows that many small but not minimal, norm interpolator can also enjoy consistency.

Proof technique and speculative bound

- Our proofs rely on a novel technique based on **strong duality**, which we think may be broadly applicable.
- Computing the generalization gap over our hypothesis class is equivalent to solving a quadratically constrained quadratic program (**QCQP**).
- The dual is an one dimensional problem, which is much easier to analyze
- A complexity term, which we call **restricted eigenvalue under interpolation**, naturally appears in the derivation of the dual.

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

- Our calculation shows the generalization gap can be characterized as the product of the **squared norm** and **restricted eigenvalue**
- Speculative bound:

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

for some suitable choice of ξ_n

References

Vaishnavh Nagarajan and J. Zico Kolter. “Uniform convergence may be unable to explain generalization in deep learning.” Advances in Neural Information Processing Systems. 2019.

Jeffrey Negrea, Gintare Karolina Dziugaite, and Daniel M. Roy. “In Defense of Uniform Con-vergence: Generalization via derandomization with an application to interpolating predictors” (2019)