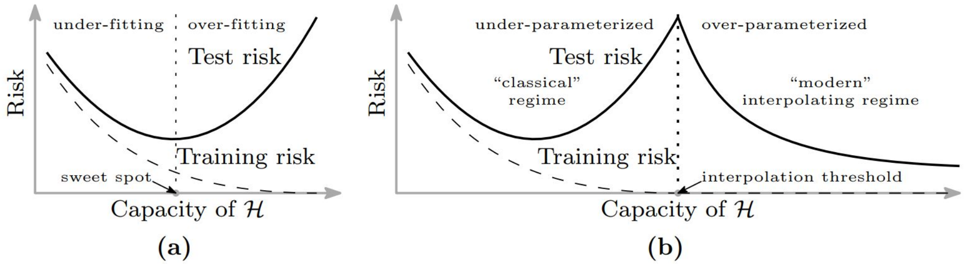
Uniform Convergence of Low-Norm Interpolators in Overparametrized Linear Regression

Joint work with:
DJ Sutherland (TTIC -> UBC)
Nati Srebro (TTIC)

Mini-seminars Day 2, May 28th, 2020

Interpolation learning

- Empirical observations that deep model can generalize reasonably well while interpolating noisy data (zero training error)
- Even more puzzling - double descent (Belkin et al, 2018)



- In high dimensions, there are infinitely many interpolators including solution with arbitrarily bad population risk

Implicit regularization

- Steepest descent for linear regression
  - Gradient descent (wrt l_2 norm) converges to minimal l_2 norm interpolator
  - Coordinate descent (wrt l_1 norm) converges to minimal l_1 norm interpolator
  - Valid even with acceleration and stochasticity
- Matrix factorization - minimal nuclear norm solution
- Logistic regression with separable data - hard margin SVM
- Linear CNN...

What’s known about minimal 2-norm interpolator?

- Benign overfitting in Linear regression (Barlett et al, 2019)
  - nearly tight high probability bound for the excess risk of the minimal norm interpolator, and obtain a necessary and sufficient condition for consistency (converging to Bayes risk)
  - strict restrictions on the spectrum of the Gaussian covariance matrix of x
  - Regime where n = o(p)
- Surprises in High-Dimensional Ridgeless Least Squares Interpolation (Hastie et al, 2019)
  - Random matrix regime, n/p -> constant
  - exact asymptotic risk when n > p and in the isotropic case when n < p; but no consistency
- Exact expressions for double descent and implicit regularization via surrogate random design (Derezinski, Liang and Mahoney 2020)
  - Approximate non-asymptotic expression (for general n, p and covariance)
  - Tools from numerical linear algebra, hard to interpret & does not say anything about consistency

Instead of exploiting the implicit regularization of GD, all existing analyses have been highly specific to the *exact* minimal norm interpolator, relying on tools from random matrix theory or numerical linear algebra.

If the true reason for good generalization in this setting is purely having a small l_2 norm, we should expect **any** interpolator with sufficiently low l_2 norm to achieve low population risk. Therefore, we focus on the quantity

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

Connection to uniform convergence

- Analysis of SVM - scale sensitive & dimension free bound

$$\forall^{\delta}_{S\sim\mathcal{D}^m}, \quad \sup_{w\in\mathbb{R}^d:\,||w||_2\le B} |L_D(w) - L_S(w)| \le 2G\sqrt{\frac{B^2\,\mathbb{E}[||x||^2]\log(2/\delta)}{m}}$$

- Class of low norm interpolators - not quite uniform convergence
- Uniform convergence may be unable to explain generalization in deep learning (Nagarajan and Kolter, 2019)
  - Look at high dimensional linear classification task, SGD on an non-typical loss
  - Show the tightest notion of uniform convergence fails

Contributions

- Tightly characterize the amount of blow-up in risk when not optimizing to the exact minimum norm
- We prove that
  - approximately minimizing the norm (up to constant suboptimality but not a factor of it) can be sufficient for consistency in a setting where minimal norm interpolator is known to be consistent
  - Neither one sided uniform convergence in the Euclidean norm ball, nor any form of two sided uniform convergence is sufficient to explain learning; holds for almost all interpolation method

Assumptions

Consider i.i.d. observations $(x_1, y_1), ..., (x_n, y_n) \sim \mathcal{D}^n$, where the joint distribution $\mathcal{D}$ is given by

(A) $x \in \mathbb{R}^p$ is drawn from $\mathcal{N}(0, \Sigma)$ and $\epsilon \in \mathbb{R}$ is independently drawn from $\mathcal{N}(0, \sigma^2)$.

(B) $y = \langle w^*, x\rangle + \epsilon$ for some $w^* \in \mathbb{R}^p$.

The "junk features" setting further assumes that

(C) $\Sigma = \begin{bmatrix} I_{d_S} & 0_{d_S \times d_J} \\ 0_{d_J \times d_S} & \frac{\lambda_n}{d_J} I_{d_J} \end{bmatrix}$ where $d_S, d_J \in \mathbb{N}$ satisfies $d_S + d_J = p$.
In other words, we can write $x = (x_S, x_J)$, where $x_S \sim \mathcal{N}(0, I_{d_S})$ and $x_J \sim \mathcal{N}(0, \frac{\lambda_n}{d_J} I_{d_J})$.

(D) y only depends on x_S, so the Bayes-optimal predictor is $w^* = (w_S^*, 0_{d_J})$ with $w_S^* \in \mathbb{R}^{d_S}$.

Junk features

- Easy to analyze, captures many desired behaviors of interpolation learning
- Recall that minimal norm interpolator is given by

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

- Ridge regression is

$$\begin{aligned} \hat{w}_\lambda &= \underset{w \in \mathbb{R}^p}{\arg\min} \|Y - X_S w\|^2 + \lambda \|w\|^2 \\ &= (X_S^\intercal X_S + \lambda I_{d_S})^{-1} X_S^\intercal Y = X_S^\intercal (X_S X_S^\intercal + \lambda I_n)^{-1} Y \end{aligned}.$$

Interpolating with appropriately scaled noise = regularization!

Writing $\hat{w}_{MN} = (\hat{w}_{MN,S}, \hat{w}_{MN,J})$, we can easily verify that

- $\hat{w}_{MN,S} = X_S^\intercal (X_S X_S^\intercal + X_J X_J^\intercal)^{-1} Y$, which converges almost surely to the ridge regression estimate with tuning parameter λ_n by the continuous mapping theorem.
- $\hat{w}_{MN,J} = X_J^\intercal (X_S X_S^\intercal + X_J X_J^\intercal)^{-1} Y$. Although the dimension of $\hat{w}_{MN,J}$ goes to ∞ with d_J, if we draw a new $x_J \sim \mathcal{N}(0, \frac{\lambda_n}{d_J} I_{d_J})$, then the strong law of large numbers yields

$$X_J x_J \stackrel{a.s.}{\to} 0_n \quad \text{and so} \quad \langle \hat{w}_{MN,J}, x_J \rangle \stackrel{a.s.}{\to} \langle 0_n, (X_S X_S^\intercal + \lambda I_n)^{-1} Y \rangle = 0.$$

- This is because

$$X_J X_J^T = \lambda_n \frac{Z_J Z_J^T}{d_J} \stackrel{a.s.}{\to} \lambda_n I_n.$$

A few definitions...

- We introduce the minimal risk interpolator to aid our analysis

$$\begin{aligned}\hat{w}_{MR} &= \underset{w \text{ s.t. } Xw=Y}{\arg\min} \; L_{\mathcal{D}}(w) \\ &= \underset{w \text{ s.t. } Xw=Y}{\arg\min} \; (w - w^*)^T \Sigma (w - w^*), \\ &= w^* + \Sigma^{-1} X^T (X \Sigma^{-1} X^T)^{-1} E\end{aligned}$$

- Define the *restricted eigenvalue under interpolation*

Definition 4.1. *Given a covariance matrix* Σ *and design matrix* X *whose columns are i.i.d. draws from* $\mathcal{N}(0, \Sigma)$*, we define the* restricted eigenvalue under interpolation *to be*

$$\kappa_X(\Sigma) = \sup_{\|w\|=1,\; Xw=0} w^\mathsf{T} \Sigma w.$$

I - A general result of uniform consistency

Theorem 4.2. *Fix a sequence (B_n) such that $B_n \geq \|\hat{w}_{MN}\|$ for all n.*

(i) If the minimal norm interpolator is consistent, $L_{\mathcal{D}}(\hat{w}_{MN}) - L_{\mathcal{D}}(w^*) \xrightarrow{a.s.} 0$, then

$$\lim_{n\to\infty} \sup_{\substack{\|w\|\leq B_n \\ L_{\mathbf{S}}(w)=0}} L_{\mathcal{D}}(w) - L_{\mathbf{S}}(w) = L_{\mathcal{D}}(w^*) + \lim_{n\to\infty} \kappa_X(\Sigma) \cdot \left[B_n^2 - \|\hat{w}_{MN}\|^2\right].$$

Thus the class of interpolators with norm less than B_n is uniformly consistent if and only if

$$\lim_{n\to\infty} \kappa_X(\Sigma) \cdot \left[B_n^2 - \|\hat{w}_{MN}\|^2\right] = 0.$$

(ii) It holds that

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

If the minimal risk interpolator is consistent, $L_{\mathcal{D}}(\hat{w}_{MR}) - L_{\mathcal{D}}(w^*) \xrightarrow{a.s.} 0$, then the class of interpolators with norm less than $\|\hat{w}_{MR}\|$ is uniformly consistent if and only if

$$\lim_{n\to\infty} \kappa_X(\Sigma) \cdot \left[\|\hat{w}_{MR}\|^2 - \|\hat{w}_{MN}\|^2\right] = 0.$$

Minimal risk interpolator

Proposition 4.3. *Under Assumptions (**A**) and (**B**), the expected risk of the minimal risk interpolator*

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

- Consistency is equivalent to n = o(p)
- Limitation of interpolation: consider a high dimensional sparse linear regression problem, where n is linear in p; LASSO known to be consistent and minimax optimal, so **any** interpolation method is suboptimal !
- Double descent behavior

$$L_{\mathcal{D}}(\hat{w}_{MR}) \leq L_{\mathcal{D}}(\hat{w}_{MN}) \leq L_{\mathcal{D}}(\hat{w}_{MR}) + 4\kappa_X(\Sigma) \cdot \left[\|\hat{w}_{MR}\|^2 - \|\hat{w}_{MN}\|^2\right]$$

Applying to junk feature setting

Theorem 4.4. *Under Assumptions (**A**) to (**D**)*

(i) Fix a sequence (α_n) *such that* $\lim_{n\to\infty} \alpha_n = \alpha$ *and* $\alpha_n \geq 1$ *for all* n*,*

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

(ii) It holds that

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

Proof Sketch

Proof sketch. This is an application of Theorem 4.2. We can show that with probability one, the following happens:

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

As the first term inside the inverse is converging to I_{d_S} and the second term is vanishingly small, we can expect that $\kappa_X(\Sigma) \approx \frac{\lambda_n}{n}$. Moreover, it can be shown that

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

Consequently, we have

$$\lim_{d_J \to \infty} \mathbb{E} \left[||\hat{w}_{MR}||^2 - \mathbb{E}||\hat{w}_{MN}||^2 \right] \leq ||w_S^*||^2 + \frac{\sigma^2 d_S}{\lambda_n}$$

The desired conclusions follow by plugging in the result from above. $\square$

II - Failure of uniform convergence

- Euclidean norm ball

Theorem 5.1. *Under Assumptions (**A**) to (**D**), the quantity*

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

- Two sided algorithm dependent uniform convergence

Theorem 5.2. *Under Assumptions (**A**) to (**D**), let $\mathcal{A}$ be an algorithm outputting interpolators, $X\mathcal{A}(X, y) = y$, which further satisfies a certain symmetry as well as consistency:*

$$\mathcal{A}\left((X_S, X_J), y\right)_S = \mathcal{A}\left((X_S, -X_J), y\right)_S \quad \textit{and} \quad \lim_{n\to\infty}\lim_{d_J\to\infty} L_{\mathcal{D}}(\mathcal{A}(X, y)) \overset{a.s.}{=} \sigma^2. \tag{2}$$

Then for any $\delta \in (0, \frac{1}{2})$ and set of typical training examples $\mathcal{S}_\delta$ satisfying $\Pr(\mathbf{S} \in \mathcal{S}_\delta) \ge 1 - \delta$, let $\mathcal{W}_\delta = \{\mathcal{A}(X, y) : (X, y) \in \mathcal{S}_\delta\}$ denote the set of typical outputs. Then

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

Notions of uniform convergence

- Algorithm-dependent uniform convergence

Fix $\delta \in (0,1)$ and a learning rule $\mathcal{A}$, find a set of training samples S_δ such that for $S \sim \mathcal{D}^m$

$$\mathbb{P}(S \in S_\delta) \geq 1 - \delta$$

Define the concept class $\mathcal{H}_\delta = \{\mathcal{A}(S) : S \in S_\delta\}$ and the following hold:

$$\sup_{S \in S_\delta} \sup_{h \in H_\delta} |L_D(h) - L_S(h)| \leq \epsilon(m, \delta, \mathcal{D})$$

This would imply

$$\mathbb{P}\Big[L_D(\mathcal{A}(S)) - L_S(\mathcal{A}(S)) \leq \epsilon(m, \delta, \mathcal{D})\Big] \geq 1 - \delta$$

Notions of uniform convergence - continued

- Distribution-dependent uniform convergence

Fix $\delta \in (0,1)$ and a learning rule $\mathcal{A}$, find a hypothesis class $\mathcal{H}_\delta(\mathcal{D})$ such that for $S \sim \mathcal{D}^m$

$$\mathbb{P}(\mathcal{A}(S) \in \mathcal{H}_\delta) \geq 1 - \delta/2$$

and the following hold:

$$\mathbb{P}\left(\sup_{h \in \mathcal{H}_\delta} |L(h) - L(h)| \leq \epsilon(m, \delta, \mathcal{D}) \right) \geq 1 - \delta/2$$

This implies we can find a S_δ and $\mathcal{H}_\delta$ such that $\mathcal{H}_\delta \supseteq \{\mathcal{A}(S) : S \in S_\delta\}$ and

$$\sup_{S \in S_\delta} \sup_{h \in H_\delta} |L(h) - L(h)| \leq \epsilon(m, \delta, \mathcal{D})$$

Contributions

- Tightly characterize the amount of degradation in risk when not optimizing to the exact minimum norm
- We prove that
  - approximately minimizing the norm (up to constant suboptimality but not a factor of it) can be sufficient for consistency in a setting where minimal norm interpolator is known to be consistent
  - Neither one sided uniform convergence in the Euclidean norm ball, nor any form of two sided uniform convergence is sufficient to explain learning; holds for almost every interpolation method
- Through minimal risk interpolator, we show
  - the limitation of interpolation method
  - recover the double descent phenomenon
  - uniform consistency may be a useful technique

Future work - optimistic rate for learning

- Fix a function f such that f(0) = 0, is it possible upper the following?

$$\sup_{w \in \mathcal{W}} L_{\mathcal{D}}(w) - f(L_{\mathbf{S}}(w))$$

For example, Theorem 1 in Srebro, Sridharan, and Tewari [14] considers an f of the form $f(x) = x + g(n, \mathcal{H})\sqrt{x}$ in the general setting where the loss is smooth. The dominant term in the upper bound is $\log^3(n)\, \mathcal{R}_n^2(\mathcal{H})$ multiplied by some numeric constant, where $\mathcal{R}_n(\mathcal{H})$ is the Radamacher complexity of hypothesis class.

- Optimistic rate for learning with a smooth loss (2010)
- The Radamacher bound for linear class is promising:

$$\sqrt{\frac{\|w\|^2\|x\|^2}{n}} \approx \sqrt{\frac{\left(\sigma^2 \frac{n}{\lambda}\right) \cdot \lambda}{n}} = \sqrt{L_{\mathcal{D}}(w^*)}$$

THANK YOU !