A Statistical Learning Theory for High Dimensional Interpolants and Regularized Estimators

Lijia Zhou

Department of Statistics
University of Chicago

Based on joint works with Frederic Koehler (Stanford), Danica Sutherland (UBC), Pragya Sur (Harvard), Nati Srebro (TTIC/UChicago)

Feb 15, 2023

Introduction

- Modern machine learning models often contain more parameters than the number of training samples.
- In these situations, **overfitting** is a very important concern. Understanding how high dimensional predictors can generalize is a fundamental problem in statistical learning theory.
- Classical wisdom of ML: simpler model generalize better.
- Instead of looking at the number of parameters, we can use norm of the predictor as a complexity measure (e.g. Rademacher complexity $\mathcal{R}_n$).
- Bias-variance tradeoff can be precisely quantified by **uniform convergence bounds**.

Classical Uniform Convergence Bounds

If we have a M-Lipschitz loss function, then it is well known that

$$L(w) \leq \hat{L}(w) + 2M\mathcal{R}_n.$$

Alternatively, if the loss is non-negative and H-smooth, then we have

$$L(w) \leq \hat{L}(w) + \tilde{O}\left(H\mathcal{R}_n^2 + \sqrt{\hat{L}(w) \cdot H\mathcal{R}_n^2}\right)$$

which is also known as "optimistic rate."
These bounds work well for conventional high dimensional problems because we can usually set $\hat{L}(w) \approx L^*$ (Bayes error) and $\mathcal{R}_n \approx 0$ with a careful choice of regularization parameter.
Uniform convergence provides an easy and general approach to establish consistency!

Benign Overfitting

However, the same uniform convergence result cannot be used to understand the benign overfitting phenemenon (achieving consistency $L(w) \to L^*$ while interpolating noisy training data $\hat{L}(w) = 0$).
This is because plugging in $\hat{L}(w) = 0$, the bounds become

$$L(w) \leq 2M\mathcal{R}_n \quad \text{or} \quad L(w) \leq \tilde{O}\left(H\mathcal{R}_n^2\right)$$

and so $\mathcal{R}_n$ is at least $\Omega(1)$. Therefore, the exact multiplicative factor $2M$ or $\tilde{O}(H)$ need to be as tight as possible in order to establish consistency.
Question: what is the tightest possible multiplicative factor? Can it be used to show benign overfitting?

Preview of Positive Results

For M-Lipschitz loss, we have (ignoring lower order terms)

$$L(w) \leq \hat{L}(w) + M\mathcal{R}_n$$

and for non-negative and $\sqrt{H/2}$ square-root Lipschitz loss (any non-negative H smooth function satisfies this), we have

$$L(w) \leq \hat{L}(w) + \frac{1}{2}H\mathcal{R}_n^2 + \sqrt{2}\sqrt{\hat{L}(w) \cdot H\mathcal{R}_n^2}$$
$$\iff \sqrt{L(w)} \leq \sqrt{\hat{L}(w)} + \sqrt{H/2} \cdot \mathcal{R}_n$$

Benign overfitting: $H = 2$ for the square loss/squared hinge loss, and if the minimal ℓ_2 norm interpolant is consistent, then $\mathcal{R}_n^2 \approx L^*$ and uniform convergence can recover consistency.

Gaussian Multi-Index Model

The data $\{(x_i, y_i)\}_{i=1}^n$ are *independent and identically distributed* (i.i.d.) and follow some data distribution $\mathcal{D}$ given by

(A) d-dimensional Gaussian features with arbitrary mean and covariance: $x \sim \mathcal{N}(\mu, \Sigma)$

(B) a generic multi-index model: there exist $w_1^*, ..., w_k^* \in \mathbb{R}^d$, a random variable $\xi \sim \mathcal{D}_\xi$ independent of x (not necessarily Gaussian), and an unknown link function $g : \mathbb{R}^{k+1} \to \mathcal{Y}$ such that

$$\eta_i = \langle w_i^*, x \rangle, \quad y = g(\eta_1, ..., \eta_k, \xi).$$

Remark. Without loss of generality, we can assume $\Sigma^{1/2} w_i^*$ are orthonormal and define $Q = I - \sum_{i=1}^k w_i^* (w_i^*)^T \Sigma$

Generalized Linear Objective

Fix a continuous loss $f : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$, define the empirical loss and population loss as

$$\hat{L}(w) = \frac{1}{n} \sum_{i=1}^{n} f(\langle w, x_i \rangle, y_i), \quad L(w) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[f(\langle w, x \rangle, y)]$$

(C) there exists $\tau > 0$ such that uniformly over all $w \in \mathbb{R}^d$

$$\frac{\mathbb{E}_{(x,y)\sim\mathcal{D}}[f(\langle w, x \rangle, y)^4]^{1/4}}{\mathbb{E}_{(x,y)\sim\mathcal{D}}[f(\langle w, x \rangle, y)]} \leq \tau$$

(D) the class of functions on $\mathbb{R}^k \times \mathbb{R}$ defined by

$$\{(x, y) \to \mathbb{1}\{f(\langle w, x \rangle, y) > t\} : w \in \mathbb{R}^k, t \in \mathbb{R}\}$$

has VC dimension at most h.

Uniform Convergence

Theorem

Under the assumptions (A), (B), (C), (D), let $C_\delta : \mathbb{R}^d \to [0, \infty]$ *be any continuous function such that with probability at least* $1 - \delta/4$ *over* $x \sim \mathcal{N}(0, \Sigma)$*, it holds uniformly over all* $w \in \mathbb{R}^d$ *that* $\langle Qw, x \rangle \leq C_\delta(w)$*. If* f *is M-Lipschitz w.r.t. the first argument for any* $y \in \mathcal{Y}$*, then with probability at least* $1 - \delta$*, it holds uniformly over all* $w \in \mathbb{R}^d$

$$(1-\epsilon)L(w) \leq \hat{L}(w) + M\sqrt{C_\delta(w)^2/n}$$

Alternatively, if $\sqrt{f}$ *is* $\sqrt{H/2}$ *Lipschitz, then the following holds*

$$(1-\epsilon)L(w) \leq \left(\sqrt{\hat{L}(w)} + \sqrt{H/2 \cdot C_\delta(w)^2/n}\right)^2$$

where $\epsilon = \tilde{O}(\tau\sqrt{h/n})$.

Complexity function C_δ

For any norm $\| \cdot \|$, we have

$$\langle Qw, x \rangle = \langle Qw, Q^T x \rangle \leq \|Qw\| \cdot \|Q^T x\|_*$$

and so we can let $\sqrt{C_\delta(w)^2/n} \approx \frac{\|Qw\| \cdot \mathbb{E}\|Q^T x\|_*}{\sqrt{n}}$ which can be viewed as the Rademacher complexity $\mathcal{R}_n$. However, this is better than using $\mathcal{R}_n$ because

- we don't need to fix a hypothesis class first and our generalization bound can be tightly applied to predictors with different norms simultaneously
- we can choose to work with either $\|w\|$ or $\|Qw\|$, and $\|Qw\|$ can be significantly smaller if the learned $\hat{w}$ is concentrated in a few fixed directions
- the data norm $\mathbb{E}\|Q^T x\|_*$ can also be significantly smaller (e.g. spiked covariance settings)

The cost of using Q is paid in $\epsilon = \tilde{O}(\tau\sqrt{h/n})$ because usually $h = O(k)$.

Benign Overfitting

Let's specialize to the square loss or the squared hinge loss where $H = 2$ and fix $\|\cdot\|$ to be the Euclidean norm. Consider the minimal ℓ_2 norm interpolant for regresion

$$\min_{w \in \mathbb{R}^d} \|w\|_2 \quad \text{s.t. } \langle w, x_i \rangle = y_i, \ \forall i \in [n]$$

or hard-margin SVM for classification

$$\min_{w \in \mathbb{R}^d} \|w\|_2 \quad \text{s.t. } y_i \langle w, x_i \rangle \geq 1, \ \forall i \in [n].$$

In both cases, we have $\hat{L}(\hat{w}) = 0$.

Benign Overfitting (continued)

Then our uniform convergence guarantee implies that

$$(1-\epsilon)L(\hat{w}) \leq \frac{\|\hat{w}\|_2^2 \operatorname{tr}(\Sigma^\perp)}{n}$$

where $\Sigma^\perp = Q^T \Sigma Q$. Indeed, if $w^\sharp$ is the Bayes optimal predictor and we define the effective rank $R(\Sigma) = \frac{\operatorname{tr}(\Sigma)^2}{\operatorname{tr}(\Sigma^2)}$, then we can show the norm bound

$$\|\hat{w}\|_2^2 \leq \|w^\sharp\|_2^2 + \left(1 + O\left(\frac{n}{R(\Sigma^\perp)}\right)\right) \frac{nL^*}{\operatorname{tr}(\Sigma^\perp)}$$

and so plugging in yields

$$(1-\epsilon)L(\hat{w}) \leq \frac{\|w^\sharp\|_2^2 \operatorname{tr}(\Sigma^\perp)}{n} + \left(1 + O\left(\frac{n}{R(\Sigma^\perp)}\right)\right) L^*$$

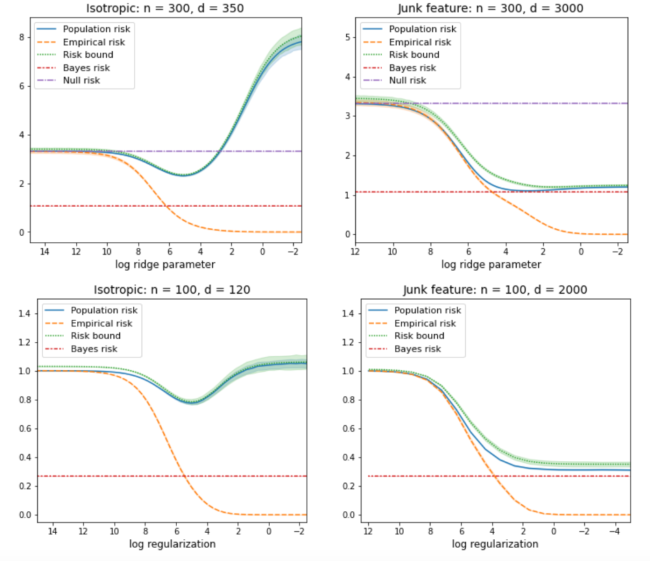
Benign Overfitting (continued)

Finally, we see that $L(\hat{w}) \to L^*$ in probability if the benign overfitting condition holds

$$\frac{k}{n} \to 0, \quad \frac{\|w^\sharp\|_2^2 \operatorname{tr}(\Sigma^\perp)}{n}, \quad \frac{n}{R(\Sigma^\perp)} \to 0$$

This set of conditions is known to be sufficient and (almost) necessary. In fact, using the full optimistic rate result, we can show that any ridge or soft-margin SVM solution with training error smaller than $\hat{L}(w^\sharp)$ is consistent.
Moreover, this result allows the distribution $\mathcal{D}$ to be misspecified by a linear model.

Experiments



Extensions

- High probability version of precise asymptotics in proportional scaling regime
- LASSO and minimal-ℓ_p norm interpolant
- application to non-convex learning problems
  - phase retrieval: $f(\langle \hat{w}, x\rangle, y) = (|\langle \hat{w}, x\rangle| - y)^2$
  - ReLU regression: $f(\langle \hat{w}, x\rangle, y) = (\sigma(\langle \hat{w}, x\rangle) - y)^2$ where $\sigma(\hat{y}) = \max(0, \hat{y})$.
  - any other 1-Lipschitz activation function (e.g. Sigmoid, tanh)
  - classification loss $f(\langle \hat{w}, x\rangle, y) = (1 - \sigma(\langle \hat{w}, x\rangle)y)_+^2$
- two-layer neural network with N hidden units
  - weights are shared in the bottom layer, but each hidden unit is allowed to have a separate bias term
  - predictors of the form $h(x) = \sum_{i=1}^N a_i \sigma(\langle w, x_i\rangle - b_i)$
- Moreau envelope theory

Reference

- **On Uniform Convergence and Low-Norm Interpolation Learning**, *NeurIPS 2020 (Spotlight)*
- **Uniform Convergence of Interpolators: Gaussian Width, Norm Bounds, and Benign Overfitting**, *NeurIPS 2021 (Oral)*
- **Optimistic Rates: A Unifying Theory for Interpolation Learning and Regularization in Linear Regression**, *Under Review.*
- **A Non-Asymptotic Moreau Envelope Theory for High-Dimensional Generalized Linear Models**, *NeurIPS 2022.*
- **Learning with Square Root Lipschitz Losses**, *Preprint soon.*