Agnostic Interpolation Learning Beyond Linear Regression

Lijia Zhou

Department of Statistics, University of Chicago

Joint work with Frederic Koehler (Stanford), Danica Sutherland (UBC), Pragya Sur (Harvard), Zhen Dai (UChicago), Jamie Simon (UC Berkeley), Gal Vardi (TTIC/Hebrew University), Nati Srebro (TTIC)

May 23, 2023

THE UNIVERSITY OF CHICAGO

Introduction

Interpolation Learning: it is possible for a high-dimensional model to interpolate **noisy** training labels, while generalizing well to unseen test data.

- Many prior works focus on the setting of *linear regression* with a *well-specified* model:

  $$y = \langle w^*, x \rangle + \xi$$

  where ξ is independent of x and $\mathbb{E}[\xi] = 0, E[\xi^2] = \sigma^2$.

- It is shown that the minimal ℓ_2 norm interpolant is consistent, i.e. test error converges in probability to the Bayes error σ^2, under some conditions on the covariance matrix Σ (e.g., Bartlett et al. 2020).

This talk...

- Beyond Linear Regression
  - uniform convergence
  - applications: max-margin classification, phase retrieval, ReLU regression, low-rank matrix sensing
- Agnostic Learning:
  - can the minimal norm interpolant achieve the best error attainable by **any linear predictor**?
  - can the minimal norm interpolant achieve the best error attainable by **any regularized estimator**?

Generalized Linear Model (GLM)

We receive i.i.d. sample pairs (x_i, y_i) from some data distribution $\mathcal{D}$ over $\mathbb{R}^d \times \mathcal{Y}$.

Fix any loss function $f : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$, we can fit a linear model $\hat{w}$ by minimizing the empirical loss $\hat{L}_f$:

$$\hat{L}_f(w) = \frac{1}{n} \sum_{i=1}^{n} f(\langle w, x_i \rangle, y_i),$$

with the goal of achieving small population loss L_f:

$$L_f(w) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[f(\langle w, x \rangle, y)].$$

Uniform Convergence (the old approach)

Decompose the test error

$$L_f(w) \le \underbrace{\hat{L}_f(w)}_{\text{training error}} + \underbrace{\sup_{w \in \mathcal{K}} |L_f(w) - \hat{L}_f(w)|}_{\text{Generalization gap}}$$

If f is M-Lipschitz: for any $y \in \mathcal{Y}$ and $\hat{y}_1, \hat{y}_2 \in \mathbb{R}$

$$|f(\hat{y}_1, y) - f(\hat{y}_2, y)| \le M|\hat{y}_1 - \hat{y}_2|,$$

then we can bound the generalization gap by the Rademacher complexity $\mathcal{R}_n$:

$$\sup_{w \in \mathcal{K}} |L_f(w) - \hat{L}_f(w)| \le 2 \cdot M\mathcal{R}_n$$

Uniform Convergence (the new approach)

If $\sqrt{f}$ is $\sqrt{H}$-Lipschitz (e.g., the square loss), then we can show

$$\sup_{w \in \mathcal{K}} \left| \sqrt{L_f(w)} - \sqrt{\hat{L}_f(w)} \right| \leq \sqrt{H \mathcal{R}_n^2}.$$

Specializing to interpolants $\hat{L}_f(\hat{w}) = 0$, we obtain

$$\sqrt{L_f(\hat{w})} \leq \sqrt{\hat{L}_f(\hat{w})} + \sqrt{H \mathcal{R}_n^2} \implies L_f(\hat{w}) \leq H \mathcal{R}_n^2.$$

For the class of norm constrained linear predictors
$\mathcal{K} = \{w \in \mathbb{R}^d : \|w\| \leq B\}$ with an arbitrary norm $\| \cdot \|$, we have

$$\mathcal{R}_n \leq \frac{B \cdot \mathbb{E}\|x\|_*}{\sqrt{n}}$$

Disclaimer!

For technical reasons, we need to assume that x is **Gaussian**, but x can have arbitrary mean and covariance. We also assume the condition distribution of y depends on x through $W^T x$ for some $W \in \mathbb{R}^{d \times k}$ where $k = o(n)$. For example,

1. $\mathcal{Y} = \mathbb{R}$ and $y = \langle w^*, x \rangle + \xi$
2. $\mathcal{Y} = \mathbb{R}$ and

$$y = \underbrace{\langle w^*, x \rangle}_{\text{linear signal}} + \underbrace{|x_1| \cdot \cos x_2}_{\text{non-linear term}} + \underbrace{x_3 \cdot \xi}_{\text{heteroscedasticity}}$$

3. $\mathcal{Y} = \{-1, 1\}$ and

$$\Pr(y = 1) = \text{sigmoid}(\langle w^*, x \rangle)$$

Application 1: Linear Regression

To upper bound

$$\min_{\substack{w \in \mathbb{R}^d: \\ \forall i \in [n], \langle w, x_i \rangle = y_i}} \|w\|_2$$

we consider $w = w^\sharp + w^\perp$, where $w^\sharp$ is the linear predictor with the **least population error** and

$$w^\perp = \arg\min_{\substack{w \in \mathbb{R}^d: \\ \forall i \in [n], \langle w, x_i \rangle = y_i - \langle w^\sharp, x_i \rangle}} \|w\|_2$$

The intuition behind the norm calculation is that if the effective ranks (Bartlett et al. 2020) are high, then x_i are approximately orthogonal and we can choose

$$w^\perp \approx \sum_{i=1}^{n} \left[\frac{y - \langle w^\sharp, x_i \rangle}{\|x_i\|^2} \right] x_i$$

Application 1: Linear Regression

and so the norm is

$$\|w^{\perp}\|_2^2 \leq (1 + o(1)) \frac{n \cdot \mathbb{E}[(y - \langle w^\sharp, x\rangle)^2]}{\mathbb{E}\|x\|_2^2},$$

and plugging into the bound $L_f(\hat{w}) \leq \frac{\|\hat{w}\|_2^2 \mathbb{E}\|x\|_2^2}{n}$, given that the norm of $w^\sharp$ is not too large, we show that

$$L_f(\hat{w}) \leq (1 + o(1))\, \mathbb{E}[(y - \langle w^\sharp, x\rangle)^2].$$

This calculation

- makes almost no assumption on the form of y
- allows us to replace $w^\sharp$ with any other linear predictor
- can provide finite-sample convergence rate

Application 2: Max-margin Classification

To upper bound

$$\min_{\substack{w \in \mathbb{R}^d: \\ \forall i \in [n], \langle w, x_i \rangle y_i \geq 1}} \|w\|_2$$

we also consider $w = w^\sharp + w^\perp$ where

$$w^\perp = \arg\min_{\substack{w \in \mathbb{R}^d: \\ \forall i \in [n], \langle w, x_i \rangle = y_i (1 - y_i \langle w^\sharp, x_i \rangle)_+}} \|w\|_2$$

then the same argument shows

$$\|w^\perp\|_2^2 \leq (1 + o(1)) \frac{n \cdot \mathbb{E}[(1 - y\langle w^\sharp, x \rangle)_+^2]}{\mathbb{E}\|x\|_2^2},$$

and so the max-margin solution is consistent with respect to the squared hinge loss $f(\hat{y}, y) = (1 - \hat{y}y)_+^2$, which is 1 square-root Lipschitz!

Application 3: Phase Retrieval

To upper bound

$$\min_{\substack{w\in\mathbb{R}^d: \\ \forall i\in[n], \langle w, x_i\rangle^2 = y_i^2}} \|w\|_2$$

we also consider $w = w^\sharp + w^\perp$. Let $I = \{i \in [n] : \langle w^\sharp, x_i\rangle \geq 0\}$, then we should let

$$w^\perp = \operatorname*{arg\,min}_{\substack{w\in\mathbb{R}^d: \\ \forall i\in I, \langle w, x_i\rangle = y_i - |\langle w^\sharp, x_i\rangle| \\ \forall i\notin I, \langle w, x_i\rangle = |\langle w^\sharp, x_i\rangle| - y_i}} \|w\|_2.$$

and so the minimal norm solution in phase retrieval is consistent with respect to $f(\hat{y}, y) = (|\hat{y}| - y)^2$, which is also 1 square-root Lipschitz!

Application 4: ReLU Regression

Let $\sigma(\hat{y}) := \max\{\hat{y}, 0\}$ be the ReLU activation. To upper bound

$$\min_{\substack{w \in \mathbb{R}^d: \\ \forall i \in [n], \sigma(\langle w, x_i \rangle) = y_i}} \|w\|_2$$

we also consider $w = w^\sharp + w^\perp$. This time, we let $I = \{i \in [n] : y_i > 0\}$ and we pick

$$w^\perp = \arg\min_{\substack{w \in \mathbb{R}^d: \\ \forall i \in I, \langle w, x_i \rangle = y_i - \langle w^\sharp, x_i \rangle \\ \forall i \notin I, \langle w, x_i \rangle = -\sigma(\langle w^\sharp, x_i \rangle)}} \|w\|_2$$

and the consistent loss in this case is

$$f(\hat{y}, y) = \begin{cases} (\hat{y} - y)^2 & \text{if} \quad y > 0 \\ \sigma(\hat{y})^2 & \text{if} \quad y = 0 \end{cases}$$

which is again 1 square-root Lipschitz!

The General Strategy

To compute the minimal norm required to interpolate:

- consider predictors of the form $w = w^\sharp + w^\perp$
- fix any $w^\sharp$, figure out the constraints on $\langle w^\perp, x_i \rangle$
- square the constraints to find the *correct* loss f to use
- chances are f is **square-root Lipschitz**

Apply the uniform convergence guarantee with square-root Lipschitz loss, and we are done!

Application 5: Low-rank Matrix Sensing

Consider the minimal *nuclear* norm solution:

$$\hat{X} = \underset{\substack{X \in \mathbb{R}^{d_1 \times d_2}: \\ \forall i \in [n], \langle A_i, X \rangle = y_i}}{\arg\min} \|X\|_*$$

Assume that the entries of A_i are i.i.d. standard Gaussian and $y_i = \langle A_i, X^* \rangle + \xi$ with $\xi \sim \mathcal{N}(0, \sigma^2)$ and X^* has rank r. Then we can compute the minimal norm to show

$$\frac{\|\hat{X} - X^*\|_F^2}{\|X^*\|_F^2} \lesssim \frac{r(d_1 + d_2)}{n} + \sqrt{\frac{r(d_1 + d_2)}{n}} \frac{\sigma}{\|X^*\|_F} + \left(\sqrt{\frac{d_1}{d_2}} + \frac{n}{d_1 d_2} \right) \frac{\sigma^2}{\|X^*\|_F^2}.$$

In particular, overfitting is benign if (i) $r(d_1 + d_2) = o(n)$, (ii) $d_1 d_2 = \omega(n)$, and (iii) $d_1/d_2 \to \{0, \infty\}$. This can happen for example when $r = \Theta(1), d_1 = \Theta(n^{1/2}), d_2 = \Theta(n^{2/3})$.

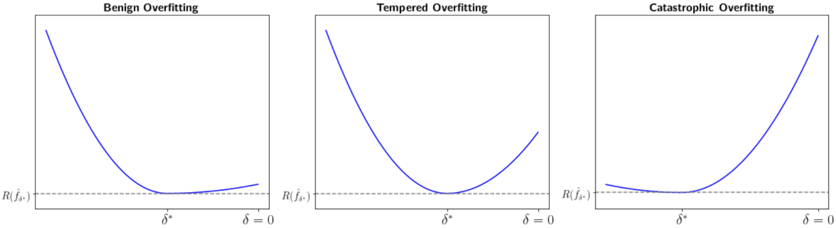
Cost of Overfitting in KRR

We consider kernel ridge regression:

$$\hat{f}_\delta = \arg\min_{f \in \mathcal{H}} \hat{R}(f) + \frac{\delta}{n}\|f\|_{\mathcal{H}}^2.$$

Given any data distribution $\mathcal{D}$ over $\mathcal{X} \times \mathbb{R}$ and sample size $n \in \mathbb{N}$, we define the **cost of overfitting** as:

$$C(\mathcal{D}, n) := \frac{R(\hat{f}_0)}{\inf_{\delta \geq 0} R(\hat{f}_\delta)}.$$



Spectrum of the Kernel

Given the marginal distribution of x, we can find the Mercer's decomposition:

$$K(x, x') = \sum_i \lambda_i \phi_i(x)\phi(x')$$

where $\mathbb{E}_x[\phi_i(x)\phi_j(x)] = \delta_{ij}$. The effective ranks of a sequence of eigenvalues $\{\lambda_i\}_{i=1}^{\infty}$ in descending order are defined as

$$r_k = \frac{\sum_{i>k} \lambda_i}{\lambda_{k+1}} \qquad \text{and} \qquad R_k := \frac{\left(\sum_{i>k} \lambda_i\right)^2}{\sum_{i>k} \lambda_i^2}.$$

Tightest Bound on $C(\mathcal{D}, n)$

Using the non-rigorous result from Simon et al. 2021, we show that there is a quantity $\mathcal{E}_0$, which only depends on n and the spectrum of the kernel $\{\lambda_i\}$, such that

$$C(\mathcal{D}, n) \leq \mathcal{E}_0.$$

Moreover, for all marginal distribution of x and sample size n, there exists $P(y|x)$ such that $C(\mathcal{D}, n) = \mathcal{E}_0$. In well-specified settings, $C(\mathcal{D}, n)/\mathcal{E}_0 \to 1$.

Benign Overfitting

For any $n \in \mathbb{N}$, let k_n be the first integer $k < n$ such that $n \leq k + r_k$. If no such k_n exists, we simplify let $k_n = n$. Then $\mathcal{E}_0 \to 1$ if and only if

$$\lim_{n\to\infty} \frac{k_n}{n} = 0 \quad \text{and} \quad \lim_{n\to\infty} \frac{n}{R_{k_n}} = 0.$$

The above result is agnostic to the distribution of y and allows the spectrum to change with n. An agnostic view on interpolation learning:

- as long as the benign overfitting conditions hold, *no matter how hard it is to learn the target*, the interpolating ridgeless solution is as asymptotically good as the optimally balanced predictor

Benign, Tempered, or Catastrophic

Suppose that the spectrum $\{\lambda_i\}$ is fixed as n increases and contains infinitely many non-zero eigenvalues.

- If $\lim_{k\to\infty} k/r_k = 0$, then overfitting is benign: $\lim_{n\to\infty} \mathcal{E}_0 = 1$.
- If $\lim_{k\to\infty} k/r_k \in (0, \infty)$, then overfitting is tempered: $\lim_{n\to\infty} \mathcal{E}_0 \in (1, \infty)$.
- If $\lim_{k\to\infty} k/r_k = \infty$, then overfitting is catastrophic: $\lim_{n\to\infty} \mathcal{E}_0 = \infty$.

Moreover, when overfitting is tempered, the cost of overfitting can be bounded by

$$\mathcal{E}_0 \lesssim 1 + \frac{k}{r_k}$$

Future Directions

- Rigorous version of the cost of overfitting
  - beyond the setting of ridge regression
- Gaussian universality
  - we have a simple counterexample (motivated by Shamir 2022) for linear regression where we can prove that we only have uniform convergence for the *weighted* square loss
  - not only uniform convergence fails, the consistency result with respect to the square loss also fails
- Extension to neural networks