A Non-asymptotic Generalization Theory for Over-parameterized Generalized Linear Models

- Dissertation Proposal Presentation by

Lijia Zhou

Department of Statistics
University of Chicago

March 10, 2022

Acknowledgment

This presentation is based on

- work with
  - Frederic Koehler (MIT → Berkeley → Stanford)
  - Danica Sutherland (TTIC → UBC)
  - Nati Srebro (TTIC/UChicago)
- and the ongoing project also with
  - Pragya Sur (Harvard)

Plan

- Motivation
  - why over-parameterization?
  - why uniform convergence?
- Failure of vanilla UC
- Localization
  - UC of interpolators
  - optimistic rates
- extension to GLM by Moreau Envelope
  - model mis-specification
  - different loss
- Future directions

Motivation

Modern statistical applications often involve models
with a lot of parameters, which can exceed the number of samples.

- we will focus on the supervised learning problem and only worry about prediction
- In these situations, the model is usually powerful enough to perfectly fit any noisy labels in the training set
  - As a result, the classical Maximum Likelihood Estimator (MLE) or the Empirical Risk Minimizer (ERM) is no longer well defined
  - Just picking an arbitrary ERM will unlikely lead to a good predictor because of overfitting

Motivation

The textbook solution to solve this high dimensional problem is **regularization**: we can constrain the model complexity in order to achieve a better bias-variance tradeoff

- in the context of linear regression, adding an ℓ_2 penalty leads to ridge regression and ℓ_1 penalty leads to LASSO
- in particular, we need to carefully tune the regularization hyperparameter. For example, a central part of the LASSO theory concerns with setting just the "right" amount of regularization to achieve the minimax rate in certain settings
- regularization is also essential to other machine learning algorithm widely used in practice, ranging from kernel SVM to gradient-boosted trees

Motivation

On the other hand, the recent success of deep learning has challenged how we understand high dimensional statistics:

- state-of-the-art results in computer vision and natural language understanding can be achieved by neural networks trained with little to no regularization
- Bigger models seem to perform better
  - For example, ResNet-152 (2015) contains about 60 million trainable parameters, and there is a language model called GPT-3 (2020) that has over **175 billion** trainable parameters
  - observable even if we randomly flipped some percentage of the training labels
  - the cause for improvement cannot be approximation error because we can already interpolate with a smaller model

Motivation

- Classical ML pipeline:
  feature selection → model fitting → carefully tune the hyperparameter by *cross validation*
- Interpolation learning:
  design the most over-parameterized model that we can successfully optimize → interpolate and we will be happy!

Setting: GLM

The data distribution $\mathcal{D}$ over (x, y) is given by

- $x \sim \mathcal{N}(0, \Sigma)$ with some unknown Σ
- there are unknown weight vectors $w_1^*, ..., w_k^* \in \mathbb{R}^d$, a function $g : \mathbb{R}^{k+1} \to \mathbb{R}$, and a random variable ξ independent of x such that

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

Given a continuous loss function $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$, our goal is to do well on a fresh sample from $\mathcal{D}$ in terms of the population loss

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

This general framework includes, for example, linear regression:

$$y = \langle w^*, x \rangle + \xi, \quad f(\hat{y}, y) = (\hat{y} - y)^2 \tag{0.1}$$

Setting: GLM

Given n i.i.d. sample pairs (x_i, y_i) from $\mathcal{D}$, we can learn a linear model $\hat{w}, \hat{b}$ by minimizing the empirical loss

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

When $d > n$, there can be infinite number of (w, b) such that $\hat{L}_f(w, b) = 0$. However, gradient descent provably find the minimal norm interpolator:

$$\hat{w}, \hat{b} = \underset{(w,b) \text{ s.t. } \hat{L}_f(w,b)=0}{\arg\min} \|w\|_2^2 + b^2$$

Our goal: to understand $L_f(\hat{w}, \hat{b})$ in terms of expectation or high probability bounds

The Case for Uniform Convergence

Note that $(\hat{w}, \hat{b})$ is a model learned from the *random dataset* $\{(x_i, y_i)\}_{i=1}^n$ generated from $\mathcal{D}$.

- In the special case of well-specified linear regression, it is sometimes possible to apply random matrix theory to directly calculate its expected risk
- but there are a number of reasons why we don't want this:
  - **RMT asymptotics:** the Marchenko-Pastur law requires a proportional scaling limit $d/n \to \gamma \in (0, \infty)$ and the data to be white $\Sigma = I_d$
  - **No closed-form:** A direct calculation would be infeasible even for the ℓ_1 norm or beyond a linear regression setting
  - **Model Mis-specification:** the linear model assumption almost never holds exactly in practice
  - **Complexity trade-off:** the trade-off between norm and training error is not explicit, and the implication for low-norm near interpolators and optimally tuned ridge/LASSO is unclear

Uniform Convergence

Classically, by concentration of measure, we have with high probability uniformly over (w, b)

$$|L_f(w,b) - \hat{L}_f(w,b)| \leq \epsilon$$

and so

$$\begin{aligned} L_f(w^*,b^*) \leq L_f(\hat{w},\hat{b}) &\leq \hat{L}_f(\hat{w},\hat{b}) + \epsilon \\ &\leq \hat{L}_f(w^*,b^*) + \epsilon \leq L_f(w^*,b^*) + 2\epsilon \end{aligned}$$

which means $L_f(\hat{w},\hat{b}) \to L_f(w^*,b^*)$ at the same rate as $\epsilon \to 0$. No closed-form expression or strong distribution assumption needed! However, in the interpolation regime

$$L_f(w^*,b^*) \leq \underbrace{\hat{L}_f(\hat{w},\hat{b})}_{=0} + \epsilon$$

Uniform Convergence: the challenges

Can we expect $\epsilon \approx L_f(w^*, b^*)$?

Theorem (informal)

There exists a sequence of distributions $\mathcal{D}_n$ such that

$$\lim_{n\to\infty} \mathbb{E} L(\hat{w}) = L(w^*)$$

but it also holds that

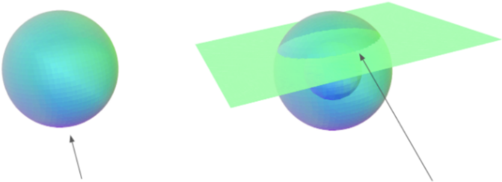
$$\lim_{n\to\infty} \mathbb{E}\left[\sup_{w:\|w\|_2 \leq \|\hat{w}\|_2} |L(w) - \hat{L}(w)| \right] = \infty$$

and for any deterministic hypothesis class $\mathcal{H}$ s.t. $\hat{w} \in \mathcal{H}$ w.p $> 1/2$

$$\lim_{n\to\infty} \mathbb{E}\left[\sup_{w\in\mathcal{H}} |L(w) - \hat{L}(w)| \right] \geq 3 L(w^*).$$

A new hope

Uniform convergence of **interpolators**:



$$\{w : \|w\|_2 \leq B\}$$

$$\{w : \|w\|_2 \leq B, \hat{L}(w) = 0\}$$

Theorem

For the same sequence $\mathcal{D}_n$ and any $\alpha \geq 1$, we have that

$$\lim_{n\to\infty} \mathbb{E}\left[\sup_{\substack{w:\|w\|_2\leq\alpha\|\hat{w}\|_2 \\ \hat{L}(w)=0}} |L(w)-\hat{L}(w)|\right] = \alpha L(w^*)$$

UC of Interpolators

Take-home message: uniform convergence with norm control can work, but we need a "localized" version of uniform convergence that only pays attention to predictors with low training error.

How do we analyze generalization gap restricted to interpolators? In the context of well-specified linear regression, we have

$$L(\hat{w}) \leq \sup_{w \in \mathcal{H}: \hat{L}(w)=0} L(w) = \sup_{w \in \mathcal{H}: Xw = Y} L(w)$$

$$= \sup_{w \in \mathcal{H}} \inf_{\lambda} \underbrace{\langle \lambda, Xw - Y \rangle}_{\text{Gaussian process}} + \underbrace{L(w)}_{\text{deterministic}}$$

We can apply the Gaussian minimax theorem (GMT) !

Gaussian width and Rademacher complexity

Given a covariance Σ, the Gaussian width of a set $\mathcal{K}$ is

$$W_\Sigma(\mathcal{K}) = \mathbb{E}_{x \sim \mathcal{N}(0,\Sigma)} \left[\sup_{w \in \mathcal{K}} |\langle w, x \rangle| \right]$$

and the (average) Rademacher complexity is

$$\mathcal{R}_n(\mathcal{H}) = \mathbb{E}_{\substack{x_1,\ldots,x_n \sim \mathcal{N}(0,\Sigma) \\ s \sim \text{Unif}(\{\pm 1\}^n)}} \left[\sup_{h \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^n s_i h(x_i) \right| \right]$$

Proposition

If we let $\mathcal{H} = \{x \mapsto \langle w, x \rangle : w \in \mathcal{K}\}$, *then*

$$\mathcal{R}_n(\mathcal{H}) = \frac{W_\Sigma(\mathcal{K})}{\sqrt{n}}$$

Generalization bound

In particular, if $\mathcal{K} = \{w : \|w\|_2 \leq B\}$, then

$$W_\Sigma(\mathcal{K}) = \mathbb{E}[B \cdot \|x\|_2] \leq \sqrt{B^2 \cdot \mathbb{E}\|x\|_2^2}$$

Given a covariance matrix Σ, we write $\Sigma = \Sigma_1 \oplus \Sigma_2$ if

1. $\Sigma = \Sigma_1 + \Sigma_2$
2. both Σ_1 and Σ_2 are p.s.d. and their spans are orthogonal

Theorem (informal)

For any covariance splitting $\Sigma = \Sigma_1 \oplus \Sigma_2$ such that $\operatorname{rank}(\Sigma_1) = o(n)$, it holds with high probability that

$$\sup_{w \in \mathcal{K}: \hat{L}(w) = 0} L(w) \leq (1 + o(1)) \cdot \frac{W_{\Sigma_2}(\mathcal{K})^2}{n}$$

Norm bound

It immediately follows that

$$\sup_{w:\|w\|_2\leq B,\hat{L}(w)=0} L(w) \leq (1+o(1)) \cdot \frac{B^2\operatorname{tr}(\Sigma_2)}{n}$$

Theorem (informal)

Under the condition

$$\frac{n}{R(\Sigma_2)} \to 0 \quad \textit{where} \quad R(\Sigma) = \frac{\operatorname{tr}(\Sigma)^2}{\operatorname{tr}(\Sigma^2)}$$

$$\implies \|\hat{w}\|_2 \leq \|w^*\|_2 + (1+o(1)) \cdot \sigma\sqrt{\frac{n}{\operatorname{tr}(\Sigma_2)}} \quad \textit{w.h.p.}$$

Plugging in, we have $L(\hat{w}) \leq \sigma^2 + o(1)$ given that

$$\frac{\operatorname{rank}(\Sigma_1)}{n} \to 0, \quad \frac{n}{R(\Sigma_2)} \to 0, \quad \|w^*\|_2\sqrt{\frac{\operatorname{tr}(\Sigma_2)}{n}} \to 0$$

Examples

- we show **uniform** consistency of low ℓ_2-norm interpolators under the benign overfitting conditions (known to be tight) using uniform convergence and norm control
- the precise convergence rate depend on the eigenvalues of Σ
- Prototypical example: junk feature

$$\Sigma = \begin{pmatrix} I_{d_S} & 0 \\ 0 & \frac{\lambda}{d_J} I_{d_J} \end{pmatrix} \implies \Sigma_1 = \begin{pmatrix} I_{d_S} & 0 \\ 0 & 0 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} 0 & 0 \\ 0 & \frac{\lambda}{d_J} I_{d_J} \end{pmatrix}$$

then $\text{rank}(\Sigma_1) = d_S$, $\text{tr}(\Sigma_2) = \lambda$ and

$$R(\Sigma_2) = \frac{\text{tr}(\Sigma_2)^2}{\text{tr}(\Sigma_2^2)} = \frac{\lambda^2}{\frac{\lambda^2}{d_J^2} d_J} = d_J$$

- we have consistency if $d_S/n \to 0$, $\|w^*\|_2 \sqrt{\lambda/n} \to 0$ and interestingly, <u>over-parameterization</u>: $n/d_J \to 0$

Examples

- more examples of ℓ_2 benign overfitting can be found in Bartlett et al. 2020
- we also extend this to ℓ_1 norm and derive the analogous ℓ_1 benign overfitting conditions. Prototypical example:

$$\Sigma = \begin{pmatrix} I_{d_S} & 0 \\ 0 & \frac{\lambda}{\log d_J} I_{d_J} \end{pmatrix} \implies \mathbb{E}\|x\|_\infty \approx \sqrt{\lambda},\ R_1(\Sigma_2) \approx \log d_J$$

- we have consistency if $d_S/n \to 0$, $\|w^*\|_1 \sqrt{\lambda/n} \to 0$ and **super-exponential** over-parameterization: $n/\log d_J \to 0$
- More recently, Wang et al. 2022 shows that when $d = n^\alpha$ and w^* is sparse, the minimal ℓ_1 norm interpolator can be consistent even though $\Sigma = I_d$. But the convergence rate is $O(1/\log n)$ and cannot be improved.

Extension: optimistic rate

What about near-interpolators?

Theorem (informal)

For any covariance splitting $\Sigma = \Sigma_1 \oplus \Sigma_2$ *such that* $\operatorname{rank}(\Sigma_1) = o(n)$, *w.h.p. uniformly over all* $w \in \mathcal{K}$

$$L(w) \leq (1+o(1)) \cdot \left(\sqrt{\hat{L}(w)} + \frac{W_{\Sigma_2}(\mathcal{K})}{\sqrt{n}} \right)^2$$

In the following slides, we show applications for the consistency of tuned ridge and LASSO, but we also have sharp bounds for:

- OLS with fixed d, with a fast rate of n^{-1}
- OLS with $d/n \to \gamma \in (0, \infty)$ and isotropic data
- LASSO in the isotropic setting

even though we do not necessarily have consistency in the last two

Ridge and LASSO

We can consider optimally-tuned ridge or LASSO:

$$\hat{w}_{\text{ridge}} = \mathop{\arg\min}_{w:\hat{L}(w)\leq\hat{L}(w^*)} \|w\|_2, \quad \hat{w}_{\text{LASSO}} = \mathop{\arg\min}_{w:\hat{L}(w)\leq\hat{L}(w^*)} \|w\|_1$$

then we get a norm bound for free by definition: $\|\hat{w}_{\text{ridge}}\|_2 \leq \|w^*\|_2$ and $\|\hat{w}_{\text{LASSO}}\|_1 \leq \|w^*\|_1$. As a result, ridge is consistent if

$$\frac{\text{rank}(\Sigma_1)}{n} \to 0, \quad \|w^*\|_2 \sqrt{\frac{\text{tr}(\Sigma_2)}{n}} \to 0$$

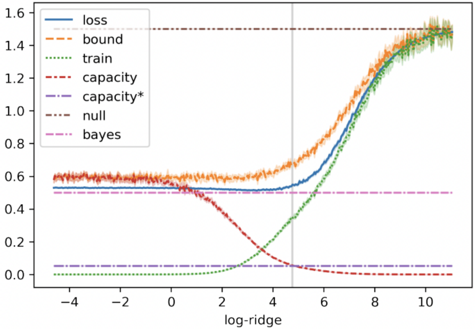
and LASSO is consistent if

$$\|w^*\|_1 \sqrt{\frac{\log d}{n}} \to 0.$$

We can get the slow rate of LASSO this way, and with a slightly more complicated argument, we can get the fast rate of LASSO with the *S-compatibility condition*.

Flatness of regularization path

- Even though the consistency condition for regularized regression is strictly weaker than the benign overfitting condition, optimistic rate shows that, under the benign overfitting condition, as long as the regularization parameter is small enough to get $\|w\|_2 \geq \|w^*\|_2$, we will have consistency.
- Experiment:



General GLM - Moreau Envelope

What about mis-specified settings? linear classification? different losses?

Moreau envelope of f with parameter λ is defined as

$$f_\lambda(\hat{y}, y) = \inf_u f(u, y) + \lambda(u - \hat{y})^2$$

and can be viewed as a smooth approximation to f.

Theorem (informal)

For any covariance splitting $\Sigma = \Sigma_1 \oplus \Sigma_2$ such that Σ_1 spans $w_1^*, ..., w_k^*$ and $\text{rank}(\Sigma_1) = o(n)$, w.h.p. uniformly over all $w \in \mathcal{K}, b \in \mathbb{R}$ and $\lambda \in \mathbb{R}^+$

$$L_{f_\lambda}(w, b) \leq \hat{L}_f(w, b) + \lambda \cdot \frac{W_{\Sigma_2}(\mathcal{K})^2}{n} + o(1)$$

Examples

We can relate f_λ with f and then optimizing over λ:

- square loss:

$$f_\lambda(\hat{y}, y) = \inf_u (u - y)^2 + \lambda(u - \hat{y})^2 = \frac{\lambda}{1+\lambda} f(\hat{y}, y)$$

- same for squared hinge loss!
- M-Lipschitz loss:

$$0 \leq f - f_\lambda \leq \frac{M^2}{4\lambda}$$

- H-smooth and non-negative loss: there exists $\tilde{f}$ s.t. $f = \tilde{f}_{H/2}$ and we can show that

$$\hat{L}_f(w, b) = 0 \implies \hat{L}_{\tilde{f}}(w, b) = 0$$

Implications

- For the square loss/squared hinge loss

$$L_f(w,b) \le \inf_{\lambda>0} \frac{1+\lambda}{\lambda}\left(\hat{L}_f(w,b) + \lambda \cdot \frac{W_{\Sigma_2}(\mathcal{K})^2}{n}\right) + o(1)$$
$$= \left(\sqrt{\hat{L}_f(w,b)} + \frac{W_{\Sigma_2}(\mathcal{K})}{\sqrt{n}}\right)^2 + o(1)$$

- For Lipschitz loss

$$L_f(w,b) \le \inf_{\lambda>0} \hat{L}_f(w,b) + \lambda \cdot \frac{W_{\Sigma_2}(\mathcal{K})^2}{n} + \frac{M^2}{4\lambda} + o(1)$$
$$= \hat{L}_f(w,b) + M\sqrt{\frac{W_{\Sigma_2}(\mathcal{K})^2}{n}} + o(1)$$

- For smooth and non-negative loss

$$\sup_{w\in\mathcal{K},b\in\mathbb{R}:\ \hat{L}_f(w,b)=0} L_f(w,b) \le \frac{H}{2} \cdot \frac{W_{\Sigma_2}(\mathcal{K})^2}{n} + o(1)$$

Summary

- Vanilla UC is not enough for interpolation
  - localization is important
  - it leads us to consider the most extreme form of localization: UC of interpolators
- A novel application of GMT allows us to
  - establish the consistency of minimal ℓ_2 norm interpolator
  - prove new results for basis pursuit
- we extend this to an optimistic-rate type result that
  - provides learning guarantee for any norm and training error
  - recovers classical results for ridge and LASSO
  - show the flatness of regularization path if there is benign overfitting
- finally, we use Moreau envelope to generalize to arbitrary GLM that includes
  - linear regression with the square loss
  - linear classification with the squared hinge loss
  - any Lipschitz loss
  - interpolation learning with a smooth loss

Future directions

- What if the distribution of x is not Gaussian?
  - a related question: what about kernels? The feature map $x \mapsto \phi(x)$ can destroy normality
- tighter norm bounds in the presence of low dimensional structure (e.g. sparsity)
- What about multi-class classification?
- What about non-linear models?
  - for example, two-layer neural networks

Reference

- **On Uniform Convergence and Low-Norm Interpolation Learning**
  - published at NeurIPS 2020 (Spotlight)
  - joint with Danica and Nati.
- **Uniform Convergence of Interpolators: Gaussian Width, Norm Bounds, and Benign Overfitting**
  - published at NeurIPS 2021 (Oral)
  - joint with Frederic, Danica and Nati.
- **Optimistic Rates: A Unifying Theory for Interpolation Learning and Regularization in Linear Regression**
  - Submitted. Under review.
  - joint with Frederic, Danica and Nati.