A Non-Asymptotic Moreau Envelope Theory for High-Dimensional Generalized Linear Models

Lijia Zhou (UChicago), **Frederic Koehler** (Stanford), **Pragya Sur** (Harvard), **Danica J. Sutherland** (UBC + Amii), **Nathan Srebro** (TTI-Chicago)

NEURAL INFORMATION PROCESSING SYSTEMS

Overview

- We study the generalization error for linear models in high dimensions (such as minimal norm interpolation) and introduce a new approach to uniform convergence
- We show that the generalization gap can be controlled by a quantity similar to the Rademacher complexity, and we use our theory to establish consistency and sharp non-asymptotic guarantees even in overparameterized & interpolation settings

Setting

- We consider a supervised learning setting with $x \sim \mathcal{N}(0,\Sigma)$ and the distribution of y only depends on x through $\eta_i = \langle w_i^*, x\rangle$ for $i = 1,\dots,k$. For example,

$$1.\ y = \langle w^*, x\rangle + \xi$$

$$2.\ \Pr(y=1) = \text{sigmoid}(\langle w^*, x\rangle)$$

$$3.\ y = \langle w_1^*, x\rangle\langle w_2^*, x\rangle + \langle w_3^*, x\rangle^2 \xi$$

- The test and training error associated with a continuous loss function is denoted as

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

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

- WLOG, assume $\Sigma^{1/2}w_1^*, \dots, \Sigma^{1/2}w_k^*$ are orthogonal and define a projection matrix

$$Q = I - \sum_{i=1}^{k} w_i^*(w_i^*)^T\Sigma$$

and the Moreau envelope

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

Main result

Let C be a continuous function such that w.h.p over $x \sim \mathcal{N}(0,\Sigma)$, it holds uniformly over $w \in \mathbb{R}^d$

$$\langle Qx, w\rangle \le C(w)$$

then under some mild conditions, w.h.p. it holds over all $(w,b) \in \mathbb{R}^{d+1}, \lambda \in \mathbb{R}^+$

$$L_{f_\lambda}(w,b) \le \left(1 + \tilde{O}\left(\sqrt{\frac{k}{n}}\right)\right)\left(\hat{L}_f(w,b) + \lambda\frac{C(w)^2}{n}\right)$$

Applications

- For the square loss $(y-\hat{y})^2$ and squared hinge loss $(1-y\hat{y})_+^2$, we have

$$f_\lambda(\hat{y}, y) = \frac{\lambda}{1+\lambda} f(\hat{y}, y)$$

and so plugging in the main result, we get

$$L_f(w,b) \le (1 + o(1))\left(\sqrt{\hat{L}_f(w,b)} + \sqrt{\frac{C(w)^2}{n}}\right)^2$$

- If f is M-Lipschitz, then $0 \le f - f_\lambda \le M^2/4\lambda$, and plugging in

$$L_f(w,b) \le (1 + o(1))\left(\hat{L}_f(w,b) + M\sqrt{\frac{C(w)^2}{n}}\right)$$

- If f is nonnegative and H-smooth, then we can represent $f = \tilde{f}_{H/2}$ and $\hat{L}_f = 0 \implies \hat{L}_{\tilde{f}} = 0$, and so uniformly over all (w,b) such that $\hat{L}_f(w,b) = 0$

$$L_f(w,b) \le (1 + o(1))\,\frac{H}{2}\cdot\frac{C(w)^2}{n}$$

Norm-based generalization

- Denote $\Sigma^\perp = Q\Sigma Q^T$, we can pick C by

$$\langle Qx, w\rangle \le \|Qx\|_2\|w\|_2 \le \left(\sqrt{\text{Tr}(\Sigma^\perp)} + O(\|\Sigma^\perp\|_{op}^{1/2})\right)\|w\|_2$$

- Recall the definition of effective ranks:

$$r(\Sigma) = \frac{\text{Tr}(\Sigma)}{\|\Sigma\|_{op}}, \quad R(\Sigma) = \frac{\text{Tr}(\Sigma)^2}{\text{Tr}(\Sigma^2)}$$

and we show that

$$\|\hat{w}\|_2^2 \le \|w^\sharp\|_2^2 + \left(1 + O\left(\frac{n}{R(\Sigma^\perp)}\right)\right)\frac{nL_f(w^\sharp, b^\sharp)}{\text{Tr}(\Sigma^\perp)}$$

- Therefore, we have benign overfitting given that

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

in both linear regression and classification settings, regardless of model mis-specification