Learning with Square Loss: Localization through Offset Rademacher Complexity

Tengyuan Liang

Department of Statistics, The Wharton School
University of Pennsylvania

Joint work with Sasha Rakhlin and Karthik Sridharan

$\mathcal{F}$ class of functions on measurable space $(\mathcal{X}, \mathcal{A})$

X and Y jointly distributed $\sim \; P = P_X \times P_{Y|X}$

$(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d. copies

$$\|f - Y\|^2 = \mathbb{E}(f(X) - Y)^2, \quad \|f - Y\|_n^2 = \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Y_i)^2$$

Excess loss with respect to model $\mathcal{F}$:

$$\mathcal{E}(g) = \|g - Y\|^2 - \inf_{f \in \mathcal{F}} \|f - Y\|^2$$

Opt in $\mathcal{F}$:

$$f^* = \operatorname*{argmin}_{f \in \mathcal{F}} \|f - Y\|^2$$

Goals:

- construct $\widehat{\mathsf{f}}$ with small $\mathcal{E}(\widehat{\mathsf{f}})$
- avoid convexity assumption on $\mathcal{F}$
- avoid boundedness of functions and noise (have only weak assumptions on $\mathcal{F}, \mathsf{P}$)

Local Rademacher averages for ERM analysis:

critical radius

$$r^* = \inf \left\{ r > 0 : \quad \mathbb{E} \sup_{f \in \mathcal{F}, \|f - f^*\| \le r} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i (f - f^*)(x_i) \right| \le r^2 \right\}$$

relies on boundedness of functions and Y

tools: Talagrand's inequality for supremum, contraction

First, consider convex $\mathcal{F}$. Then

$$\left\|f^* - Y\right\|_n^2 - \left\|\widehat{f} - Y\right\|_n^2 \geq \left\|\widehat{f} - f^*\right\|_n^2 \qquad (\mathrm{Py})$$

Basic inequality:

$$\begin{aligned}
\mathcal{E}(\widehat{f}) &= \left\|\widehat{f} - Y\right\|^2 - \left\|f^* - Y\right\|^2 \\
&\leq \left\|\widehat{f} - Y\right\|^2 - \left\|f^* - Y\right\|^2 + \left\|f^* - Y\right\|_n^2 - \left\|\widehat{f} - Y\right\|_n^2 - \left\|\widehat{f} - f^*\right\|_n^2 \\
&= 2(P_n - P)[(f^* - Y)(f^* - \widehat{f})] + \left\|f^* - \widehat{f}\right\|^2 - 2\left\|f^* - \widehat{f}\right\|_n^2 \\
&\leq \sup_{f \in \mathcal{F}} \left\{ 2(P_n - P)[(f^* - Y)(f^* - f)] + \left\|f^* - f\right\|^2 - 2\left\|f^* - f\right\|_n^2 \right\}
\end{aligned}$$

Observe: can upper bound by supremum of **negative-mean** process.

Offset Rademacher

Offset Rademacher averages of $\mathcal{G}$ and constant $c \geq 0$ are defined as

$$\widehat{\mathscr{R}}_n^{\text{off}}(\mathcal{G}) = \mathbb{E}_\epsilon \sup_{g \in \mathcal{G}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \epsilon_i g(z_i) - c g^2(z_i) \right\}$$

Empirical Rademacher averages correspond to $c = 0$.

Example

Class of linear functions

$$\mathcal{F} = \left\{ f(x) = \langle w, x \rangle : w \in \mathbb{R}^d \right\}, \quad \Sigma = \sum_{i=1}^{n} x_i x_i^{\mathsf{T}}$$

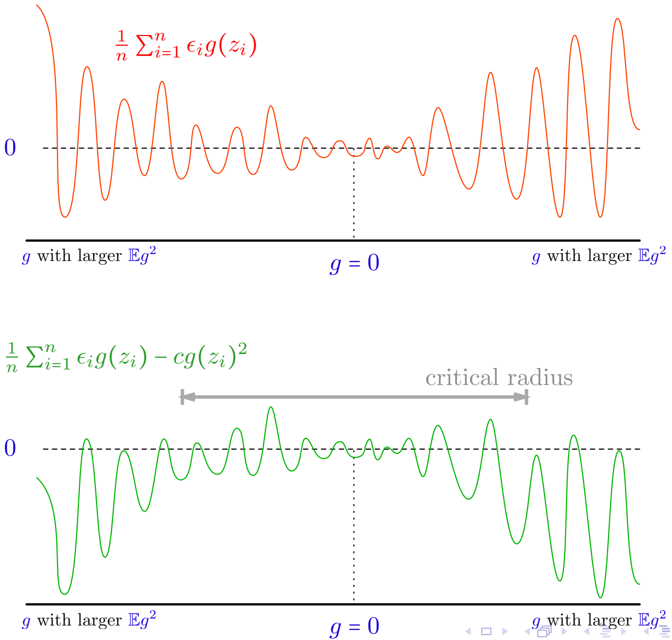
Then offset Rademacher complexity is

$$\mathbb{E}_\epsilon \sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(x_i) - f^2(x_i) \right\} = \frac{1}{n} \mathbb{E}_\epsilon \sup_{w \in \mathbb{R}^d} \left\{ w^{\mathsf{T}} \left(\sum_{i=1}^{n} \epsilon_i x_i \right) - \|w\|_\Sigma^2 \right\}$$

$$= \frac{c}{n} \left\| \sum_{i=1}^{n} \epsilon_i x_i \right\|_{\Sigma^{-1}}^2 \qquad \sim \frac{\sigma^2 d}{n}$$

In contrast, the usual (non-offset) complexity will only give $n^{-1/2}$ rates.

Intuition


$\frac{1}{n}\sum_{i=1}^{n}\epsilon_i g(z_i)$
0
g with larger $\mathbb{E}g^2$
$g = 0$
g with larger $\mathbb{E}g^2$
$\frac{1}{n}\sum_{i=1}^{n}\epsilon_i g(z_i) - cg(z_i)^2$
critical radius
0
g with larger $\mathbb{E}g^2$
$g = 0$
g with larger $\mathbb{E}g^2$

Next: prove (Py) for non-convex classes.

Cannot hope that ERM will work: any selector is suboptimal.

Which $\widehat{f}$ satisfies

$$\|f^* - Y\|_n^2 - \|\widehat{f} - Y\|_n^2 \geq c\,\|\widehat{f} - f^*\|_n^2 \qquad \text{(Py)}$$

with some constant $c > 0$? $\Longrightarrow$ design new algorithm

The Star algorithm

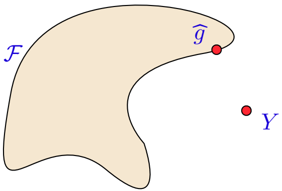
$$\mathrm{star}(\mathcal{F}, g) = \{\lambda g + (1-\lambda) f : f \in \mathcal{F}, \lambda \in [0,1]\}$$

$$\widehat{g} = \underset{f \in \mathcal{F}}{\mathrm{argmin}} \|f - Y\|_n^2, \qquad \widehat{f} = \underset{f \in \mathrm{star}(\mathcal{F}, \widehat{g})}{\mathrm{argmin}} \|f - Y\|_n^2$$

The Star algorithm

$$\mathrm{star}(\mathcal{F}, g) = \{\lambda g + (1-\lambda) f : f \in \mathcal{F}, \lambda \in [0,1]\}$$

$$\widehat{g} = \operatorname*{argmin}_{f \in \mathcal{F}} \|f - Y\|_n^2 , \qquad \widehat{f} = \operatorname*{argmin}_{f \in \mathrm{star}(\mathcal{F}, \widehat{g})} \|f - Y\|_n^2$$

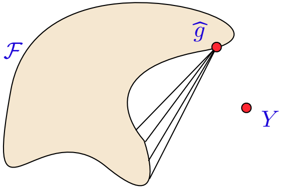


If $\mathcal{F}$ is convex, the Star algorithm coincides with ERM.

The Star algorithm

$$\mathrm{star}(\mathcal{F}, g) = \{\lambda g + (1-\lambda) f : f \in \mathcal{F}, \lambda \in [0,1]\}$$

$$\widehat{g} = \operatorname*{argmin}_{f \in \mathcal{F}} \|f - Y\|_n^2, \qquad \widehat{f} = \operatorname*{argmin}_{f \in \mathrm{star}(\mathcal{F}, \widehat{g})} \|f - Y\|_n^2$$

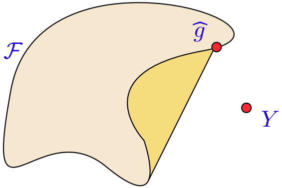


If $\mathcal{F}$ is convex, the Star algorithm coincides with ERM.

The Star algorithm

$$\mathrm{star}(\mathcal{F}, \mathsf{g}) = \{\lambda \mathsf{g} + (1-\lambda)\mathsf{f} : \mathsf{f} \in \mathcal{F}, \lambda \in [0,1]\}$$

$$\widehat{\mathsf{g}} = \operatorname*{argmin}_{\mathsf{f} \in \mathcal{F}} \|\mathsf{f} - \mathsf{Y}\|_n^2 , \qquad \widehat{\mathsf{f}} = \operatorname*{argmin}_{\mathsf{f} \in \mathrm{star}(\mathcal{F}, \widehat{\mathsf{g}})} \|\mathsf{f} - \mathsf{Y}\|_n^2$$

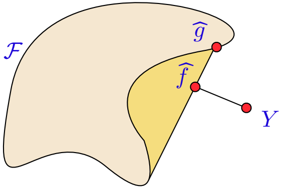


If $\mathcal{F}$ is convex, the Star algorithm coincides with ERM.

The Star algorithm

$\mathrm{star}(\mathcal{F}, \mathrm{g}) = \{\lambda \mathrm{g} + (1-\lambda)\mathrm{f} : \mathrm{f} \in \mathcal{F}, \lambda \in [0,1]\}$

$$\widehat{\mathrm{g}} = \operatorname*{argmin}_{\mathrm{f} \in \mathcal{F}} \|\mathrm{f} - \mathrm{Y}\|_n^2, \qquad \widehat{\mathrm{f}} = \operatorname*{argmin}_{\mathrm{f} \in \mathrm{star}(\mathcal{F}, \widehat{\mathrm{g}})} \|\mathrm{f} - \mathrm{Y}\|_n^2$$



If $\mathcal{F}$ is convex, the Star algorithm coincides with ERM.

Star algorithm was introduced by (Audibert '07) for $\mathcal{F}$ of finite cardinality. He showed it is deviation-optimal for *finite aggregation*.

(Lecué, Mendelson '13): ERM, convex and subgaussian class.

(Rakhlin, Sridharan, Tsybakov '14): 3-step estimator, bounded classes.

Key geometric inequality

Lemma.

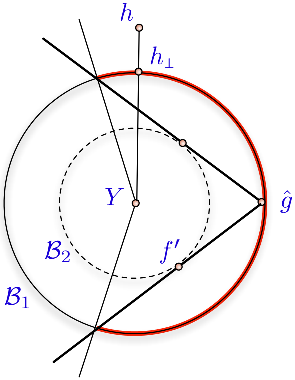
The Star algorithm $\widehat{f}$ satisfies

$$\|h - Y\|_n^2 - \|\widehat{f} - Y\|_n^2 \geq c \cdot \|\widehat{f} - h\|_n^2 \qquad \text{(Py)}$$

for any $h \in \mathcal{F}$ and $c = 1/18$.

If $\mathcal{F}$ is convex, (Py) holds with $c = 1$. If $\mathcal{F}$ is a linear subspace, (Py) holds with equality and $c = 1$.

Proof of key geometric inequality



Corollary.

For $c = 1/18$, the Star estimator satisfies

$$\mathcal{E}(\widehat{f}) \le (P_n - P)[2(f^* - Y)(f^* - \widehat{f})] + \left\|f^* - \widehat{f}\right\|^2 - (1 + c) \cdot \left\|f^* - \widehat{f}\right\|_n^2$$

conditionally on data.

Bounded case: warm up

Lemma.

Define $\mathcal{H} := \mathcal{F} - f^* + \text{star}(\mathcal{F} - \mathcal{F})$. Suppose $K = \sup_f |f|_\infty$, $M = \sup_f |Y - f|_\infty$. Then

$$\mathbb{E}[\mathcal{E}(\hat{f})] \leq c\mathbb{E}\widehat{\mathscr{R}}_n^{\text{off}}(\mathcal{H})$$

Complexity of $\mathcal{H}$ is of same order as that of $\mathcal{F}$.

High probability statement for unbounded functions

Assumption:

Function class $\mathcal{H}$ satisfies the lower isometry bound for $0 < \delta < 1$ and $c = 1/72$ if

$$\mathbb{P}\left(\inf_{h \in \mathcal{H}} \frac{\|h\|_n^2}{\|h\|^2} \geq 1 - c\right) \geq 1 - \delta$$

for all $n \geq n_{LIC}(\mathcal{H}, \delta)$.

(Mendelson 14', 15'): this holds under small ball assumption + norm comparison (e.g. $\|h\|_q \leq L\|h\|_2, 2 < q \leq 4$ for all $h \in \mathcal{H}$). It also holds for subgaussian classes. Holds for heavy-tail.

High probability statement for unbounded functions

Theorem.

$\mathcal{H} := \mathcal{F} - f^* + \mathrm{star}(\mathcal{F} - \mathcal{F})$, $\xi_i = Y_i - f^*(X_i)$. Suppose

$$\sup_{h \in \mathcal{H}} \frac{\mathbb{E}h^4}{(\mathbb{E}h^2)^2} \leq A, \quad \mathbb{E}\xi^4 \leq B. \tag{*}$$

Then

$$\mathbb{P}\left(\mathcal{E}(\widehat{f}) > 4u\right) \leq 4\delta + 4\mathbb{P}\left(\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \xi_i h(X_i) - c \cdot h(X_i)^2 > u\right)$$

for any $u > \frac{c\sqrt{AB}}{n}$, as long as $n > cA \vee n_{\mathrm{LIC}}(\mathcal{H}, \delta)$.

We can remove the moment condition (*) via a probabilistic symmetrization trick by (Panchenko '03).

Critical radius

$$r^* = \inf\left\{r > 0 : \mathbb{P}\left(\sup_{h\in\mathcal{H}\cap rB_2}\left\{\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\xi_i h(X_i) - c\cdot h^2(X_i)\right\} \le r^2\right) \ge 1-\delta\right\}.$$

Lemma.

Assume $\mathcal{H}$ is star-shaped around 0 and lower isometry bound holds. Then with prob. at least $1-2\delta$,

$$\sup_{h\in\mathcal{H}}\left\{\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\xi_i h(X_i) - c\cdot h^2(X_i)\right\} = \sup_{h\in\mathcal{H}\cap r^* B_2}\left\{\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\xi_i h(X_i) - c\cdot h^2(X_i)\right\}$$

$$\le {r^*}^2$$

Example: linear regression

Lemma.

The offset Rademacher is bounded as

$$\mathbb{E}_{\epsilon} \sup_{\beta \in \mathbb{R}^d} \left\{ \frac{1}{n} \sum_{i=1}^{n} 2\epsilon_i \xi_i X_i^T \beta - C\beta^T X_i X_i^T \beta \right\} = \frac{\mathrm{tr}\left(G^{-1}H\right)}{Cn}$$

where $G := \sum_{i=1}^n X_i X_i^T$ and $H = \sum_{i=1}^n \xi_i^2 X_i X_i^T$.

Assuming that conditional moment $\mathbb{E}(\xi^2|X)$ is σ^2, then conditionally on the design, $\mathbb{E}G^{-1}H = \sigma^2 I_d$ and excess loss is order $\frac{\sigma^2 d}{n}$.

Example: finite aggregation

Lemma.

Let $V \subset \mathbb{R}^n$ be a finite set. Then for any $C > 0$,

$$\mathbb{P}_\epsilon \left(\max_{v \in V} \left[\frac{1}{n} \sum_{i=1}^{n} \epsilon_i \xi_i v_i - C v_i^2 \right] \geq M \cdot \frac{\log |V| + \log 1/\delta}{n} \right) \leq \delta,$$

where

$$M := \max_{v \in V} \frac{\sum_{i=1}^n v_i^2 \xi_i^2}{2C \sum_{i=1}^n v_i^2}.$$

Lemma (Chaining).

Let $\mathcal{G}$ be a class of functions from $\mathcal{Z}$ to $\mathbb{R}$. Then for any $z_1, \ldots, z_n \in \mathcal{Z}$

$$\mathbb{E}_\epsilon \sup_{g \in \mathcal{G}} \left[\frac{1}{n} \sum_{t=1}^{n} \epsilon_i g(z_i) - C g(z_i)^2 \right]$$

$$\leq \inf_{\gamma \geq 0, \alpha \in [0, \gamma]} \left\{ \frac{(2/C) \log \mathcal{N}_2(\mathcal{G}, \gamma)}{n} + 4\alpha + \frac{12}{\sqrt{n}} \int_\alpha^\gamma \sqrt{\log \mathcal{N}_2(\mathcal{G}, \delta)} d\delta \right\}$$

where $\mathcal{N}_2(\mathcal{G}, \gamma)$ is an ℓ_2-cover of $\mathcal{G}$ on $(z_1, \ldots, z_n)$ at scale γ (assumed to contain $\mathbf{0}$).

Example: nonparametric function classes

Suppose

$$\log \mathcal{N}_2(\mathcal{F}|_{x_1,\ldots,x_n}, \alpha) \leq \alpha^{-p}$$

Leads to $n^{-\frac{2}{2+p}}$ for $p \in (0,2)$, $n^{-1/p}$ for $p > 2$, and $n^{-1/2}\log(n)$ at $p = 2$.

In bounded case, these were shown in (Rakhlin, Sridharan, Tsybakov '14).

For well-specified models, transition at $p = 2$ does not happen, and the rate remains $n^{-\frac{2}{2+p}}$.

Lower bound

Define worst-case offset Rademacher complexity

$$\mathscr{R}^{\circ}(\mathcal{F}, n) = \sup_{\{x_i\}_{i=1}^{n} \in \mathcal{X}^{\otimes n}} \mathbb{E}_{\epsilon} \sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} 2\epsilon_i f(x_i) - f(x_i)^2 \right\}$$

then the following minimax lower bound on regret holds:

$$\inf_{\hat{g} \in \mathcal{G}} \sup_{P} \left\{ \|\hat{g} - Y\|^2 - \inf_{f \in \mathcal{F}} \|f - Y\|^2 \right\} \geq \mathscr{R}^{\circ}((1+c)n, \mathcal{F}) - \frac{c}{1+c} \mathscr{R}^{\circ}(cn, \mathcal{G}),$$

for any $c > 0$.

Thanks!