Local Rademacher Averages and Empirical Minimization

Peter Bartlett
Division of Computer Science and Department of Statistics
UC Berkeley

Joint work with
Olivier Bousquet, Shahar Mendelson and Petra Philips.

slides at http://www.cs.berkeley.edu/∼bartlett/talks

Motivation: A prediction problem

- i.i.d. $(X, Y), (X_1, Y_1), \ldots, (X_n, Y_n)$ from $\mathcal{X} \times \mathcal{Y}$.
- Use data $(X_1, Y_1), \ldots, (X_n, Y_n)$ to choose $\hat{f} : \mathcal{X} \to \mathcal{Y}$ with small risk,

$$\mathbb{E}\ell(Y, f(X)),$$

  where $\ell : \mathcal{Y}^2 \to \mathbb{R}^+$ is a loss function.
- Empirical risk minimization: choose $\hat{f} \in \mathcal{F}$ to minimize

$$\hat{\mathbb{E}}\ell(Y, f(X)) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f(X_i)).$$

- Question: What is $\mathbb{E}\ell(Y, \hat{f}(X)) - \inf_{f \in \mathcal{F}} \mathbb{E}\ell(Y, f(X))$?

Loss Classes

- Fix a class $\mathcal{G}$ of functions on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. e.g., excess loss class: $g : (x, y) \mapsto \ell(y, f(x)) - \ell(y, f^*(x))$, where $f^* \in \arg\min_{f \in \mathcal{F}} \mathbb{E}\ell(Y, f(X))$.
- Minimizing $\hat{\mathbb{E}}g$ is equivalent to empirical risk minimization over $\mathcal{F}$.
- $\mathbb{E}g$ is excess risk.

Empirical Minimization

From now on, we'll consider:

- i.i.d. $X, X_1, \ldots, X_n$ from $\mathcal{X}$,
- a class $\mathcal{F}$ of $[0, 1]$-valued functions on $\mathcal{X}$ (with $\mathbb{E}f \geq 0$),
- $\hat{f} \in \arg\min_{f\in\mathcal{F}} \hat{\mathbb{E}}f$.

Question: What is $\mathbb{E}\hat{f}$?

Notation

Define: $\xi_n(r_1, r_2) = \mathbb{E} \sup \left\{ \mathbb{E}f - \hat{\mathbb{E}}f : f \in \mathcal{F}, r_1 \le \mathbb{E}f < r_2 \right\},$

$$\xi_n(r) = \mathbb{E} \sup \left\{ \mathbb{E}f - \hat{\mathbb{E}}f : f \in \mathcal{F}, \mathbb{E}f = r \right\}.$$

- Classical results:

$$\mathbb{E}\hat{f} \le \sup\{r > 0 : \xi_n(0,1) - r \ge 0\} + \cdots .$$

  Implied by bounds on Vapnik-Chervonenkis dimension/uniform covering numbers. But conservative (valid for any probability distribution). Also implied by bounds on covering numbers in $L_2(P)$. But not useful when P is unknown.

Risk Bounds

- Global uniform convergence is stronger than necessary: Asymptotic analysis of M-estimators shows that can replace supremum of empirical process with a fixed point of the modulus of continuity of the empirical process. (e.g., van de Geer, 2000)
- Analogous results are known for the finite sample case, of the form

$$\mathbb{E}\hat{f} \leq \sup\{r > 0 : \psi_n(0, r) - r \geq 0\} + \cdots ,$$

where $\psi_n(r_1, r_2) = \mathbb{E} \sup \left\{ \mathbb{E}f - \hat{\mathbb{E}}f : f \in \mathcal{F}, r_1 \leq \mathbb{E}f^2 < r_2 \right\}.$

(Koltchinskii and Panchenko, 2000), (Lugosi and Wegkamp, 2004), (Bartlett, Bousquet and Mendelson, 2004), (Koltchinskii, 2004).

Outline

1. Improvement (L_1 shells versus L_2 balls):

$$\mathbb{E}\hat{f} \leq \sup\{r > 0 : \xi_n(r) - r \geq 0\} + \cdots .$$

2. Estimating the fixed point $\xi_n(r) = r$ from data, using Rademacher averages.

3. An optimal bound:

$$\mathbb{E}\hat{f} = \arg\max_{r>0}(\xi_n(r) - r) \pm \cdots .$$

4. Examples: The improvement can be enormous. But in general, the better bound cannot be estimated from data.

Assumptions

Bounded Each f in $\mathcal{F}$ maps to $[-1, 1]$.

Star-shaped If $f \in \mathcal{F}$ and $0 \leq \alpha \leq 1$, then $\alpha f \in \mathcal{F}$.

Bernstein For some $0 < \beta \leq 1$ and $B \geq 1$, every $f \in \mathcal{F}$ satisfies

$$\mathbb{E}f^2 \leq B\,(\mathbb{E}f)^{\beta}.$$

Examples of Bernstein classes:
- non-negative functions.
- excess loss class from a convex function class and a strictly convex loss.
- excess loss class for low-noise classification.

(For simplicity, suppose $\beta = 1$.)

Isomorphic coordinate projections

Theorem: If $\mathcal{F}$ is bounded, star-shaped, Bernstein, and contains a function with $\mathbb{E}f = 0$, then with probability at least $1 - e^{-x}$, the empirical minimizer satisfies

$$\mathbb{E}\hat{f} \leq \sup\{r > 0 : \xi_n(r) - r/4 \geq 0\} \vee \frac{cx}{n}.$$

Recall: $\xi_n(r) = \mathbb{E}\sup\left\{\mathbb{E}f - \hat{\mathbb{E}}f : f \in \mathcal{F}, \mathbb{E}f = r\right\}.$

Isomorphic coordinate projections: Proof idea

Definition: The coordinate projection $\Pi_{X_1^n} : f \mapsto (f(X_1), \ldots, f(X_n))$ is an **ϵ-isomorphism** for $\mathcal{F}$ if for every $f \in \mathcal{F}$,

$$(1 - \epsilon)\mathbb{E}f \leq \hat{\mathbb{E}}f \leq (1 + \epsilon)\mathbb{E}f.$$

Isomorphic coordinate projections: Proof idea

Theorem: If $r \geq \frac{cx}{n\alpha^2}$, with probability $1 - e^{-x}$,

$\xi_n(r) \leq (1-\alpha)r$

$\implies \Pi_{X_1^n}$ is an ϵ-isomorphism of $\mathcal{F}_r \implies \xi_n(r) < (1+\alpha)r$.

where $\mathcal{F}_r = \{f \in \mathcal{F} : \mathbb{E}f = r\}$.

Proof: Talagrand's functional Bernstein inequality (the Bernstein property controls the variance term).

Theorem: For star-shaped $\mathcal{F}$,
$\Pi_{X_1^n}$ is an ϵ-isomorphism of $\mathcal{F}_r$
$\iff \Pi_{X_1^n}$ is an ϵ-isomorphism of $\{f \in \mathcal{F} : \mathbb{E}f \geq r\}$.

Isomorphic coordinate projections: Proof idea

Combining gives:

Theorem: If
$$r \geq \frac{\xi_n(r)}{2} \vee \frac{cx}{n\alpha^2},$$
then with probability at least $1 - e^{-x}$, every $f \in \mathcal{F}$ satisfies
$$\mathbb{E}f \leq \frac{\hat{\mathbb{E}}f}{2} \vee r.$$

Thus, if some f has $\mathbb{E}f = 0$, the empirical minimizer satisfies $\mathbb{E}\hat{f} \leq r$.

Isomorphic coordinate projections

Theorem: If $\mathcal{F}$ is bounded, star-shaped, Bernstein, and contains a function with $\mathbb{E}f = 0$, then with probability at least $1 - e^{-x}$, the empirical minimizer satisfies

$$\mathbb{E}\hat{f} \leq \sup\{r > 0 : \xi_n(r) - r/4 \geq 0\} \vee \frac{cx}{n}.$$

How can we use data to estimate

$$r^* = \sup\left\{r > 0 : \xi_n(r) \geq \frac{r}{4}\right\}?$$

Estimating the fixed point from data

Definition: For $f \in \mathcal{F}$ and $X_1, \ldots, X_n$, the Rademacher average is

$$R_n f = \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(X_i),$$

where $\sigma_1, \ldots, \sigma_n$ are independent uniform $\{\pm 1\}$ random variables.
Define

$$R_n \mathcal{F} = \sup_{f \in \mathcal{F}} R_n f,$$

and the empirical version,

$$\mathbb{E}_\sigma R_n \mathcal{F} = \mathbb{E} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(X_i) \,\middle|\, X_1, \ldots, X_n \right].$$

Estimating the fixed point from data

We want an upper bound on $r^* = \sup\left\{r > 0 : \xi_n(r) \geq \frac{r}{4}\right\}$.
We compute $\hat{r}^* = \sup\left\{r > 0 : \hat{\xi}_n(r) \geq \frac{r}{4}\right\}$. Justification:

$$\begin{aligned}
\xi_n(r) &\leq 2\mathbb{E}R_n(\mathcal{F}_r) \qquad \text{(symmetrization)}\\
&\leq 4\mathbb{E}_\sigma R_n(\mathcal{F}_r) + \frac{x}{n} \qquad \text{(Talagrand's inequality)}\\
&\leq 4\mathbb{E}_\sigma R_n(\hat{\mathcal{F}}_{r/2,3r/2}) + \frac{r}{c} \quad \text{(w.p. } 1-e^{-x} \text{ if } r \geq r^* \vee cx/n\text{)}\\
&= \hat{\xi}_n(r).
\end{aligned}$$

Here, $\mathcal{F}_r = \{f \in \mathcal{F} : \mathbb{E}f = r\}$ and
$\hat{\mathcal{F}}_{r/2,3r/2} = \{f \in \mathcal{F} : \hat{\mathbb{E}}f \in [r/2, 3r/2]\}$.

Estimating the fixed point from data

Using binary search (with $O(\log n)$ steps), we can compute an estimate $\hat{r}$ that is with high probability larger than r^*:

Theorem: If $\mathcal{F}$ is bounded, star-shaped, Bernstein, and contains a function with $\mathbb{E}f = 0$, then with probability at least $1 - cne^{-x}$, the empirical minimizer satisfies

$$\mathbb{E}\hat{f} \leq \hat{r} \vee \frac{cx}{n}.$$

A near-optimal bound

Roughly:

$$\mathbb{E}\hat{f} = \arg\max_{r>0}(\xi_n(r) - r) \pm \cdots .$$

More precisely: Define the range of near-maximizers of $\xi_n(r) - r$:

$$r_{\epsilon,+} = \sup\left\{0 \leq r \leq 1 : \xi_n(r) - r \geq \sup_s \left(\xi_n(s) - s\right) - \epsilon\right\},$$

$$r_{\epsilon,-} = \inf\left\{0 \leq r \leq 1 : \xi_n(r) - r \geq \sup_s \left(\xi_n(s) - s\right) - \epsilon\right\}.$$

A near-optimal bound

Theorem:

1. With probability at least $1 - e^{-x}$,
$$\mathbb{E}\hat{f} \leq r_{\epsilon,+} \vee \frac{1}{n},$$
2. If $\xi_n(0, c_1/n) < \sup_{s>0} (\xi_n(s) - s) - \epsilon$, then with probability at least $1 - e^{-x}$,
$$\mathbb{E}\hat{f} \geq r_{\epsilon,-},$$

provided
$$\epsilon \geq \left(\frac{c(x + \log n)}{n} \sup_{s>0} (\xi_n(s) - s) \right)^{1/2}.$$

A near-optimal bound: Proof idea

- Split $\mathcal{F}$ into shells of different expectation.
- Define $s = \arg\max_{r>0}(\xi_n(r) - r)$.
- Use concentration to show that there is likely to be a function f in $\{f \in \mathcal{F} : \mathbb{E}f = s\}$ with $\hat{\mathbb{E}}f$ smaller than $\hat{f}$ for any f in $\{f \in \mathcal{F} : r_1 \leq \mathbb{E}f \leq r_2\}$, for
  1. (upper bound:) $[r_1, r_2] = [r^*, 1]$;
     $[r_1, r_2] = [r, r + \Delta r]$ with $r_{\epsilon,+} \leq r \leq r^*$.
  2. (lower bound:) $[r_1, r_2] = [0, 1/n]$;
     $[r_1, r_2] = [r, r + \Delta r]$ with $1/n \leq r \leq r_{\epsilon,-}$.

The near-optimal bound versus the fixed point

The difference can be enormous:

Theorem: For $x > 0$ and $n > N_0(x)$ there is a probability measure P and a bounded, star-shaped, Bernstein class $\mathcal{F}$, such that

$$\xi_n(r) = \begin{cases} (n+1)r & \text{if } 0 < r \leq 1/n, \\ r & \text{if } 1/n < r \leq 1/4, \\ 0 & \text{if } r > 1/4. \end{cases}$$

Fixed point is $\sup\{r > 0 : \xi_n(r) - r/4 \geq 0\} = 1/4$.
Maximizer of $\xi_n(r) - r$ is $1/n$, so with probability at least $1 - e^{-x}$,

$$\frac{1}{n}\left(1 - c\sqrt{\log n / n}\right) \leq \mathbb{E}\hat{f} \leq \frac{1}{n}.$$

The near-optimal bound versus the fixed point

But the difference cannot be estimated in general:

Theorem: For any $n > N_0$ there is a probability measure P and a pair of bounded, star-shaped, Bernstein classes, $\mathcal{F}_1, \mathcal{F}_2$, such that
1. For every $f \in \mathcal{F}_1$, $\mathbb{E}f \leq c/n$.
2. For every $f \in \mathcal{F}_2$, $\mathbb{E}f \geq 1/4$.
3. For every $X_1, \ldots, X_n$, $\Pi_{X_1^n}\mathcal{F}_1 = \Pi_{X_1^n}\mathcal{F}_2$.

That is, any statistic based only on values of functions on the data cannot lead to a general upper bound that is better than the fixed point.

And there are versions of these results for fixed function classes (i.e., not varying with n).

Outline

$$\text{Uniform convergence: } \mathbb{E}\hat{f} \le \sup\{r > 0 : \xi_n(0,1) - r \ge 0\} + \cdots,$$

$$\text{Local, } L_2 \text{ balls: } \mathbb{E}\hat{f} \le \sup\{r > 0 : \psi_n(0,r) - r \ge 0\} + \cdots,$$

$$\text{Local, } L_1 \text{ shells: } \mathbb{E}\hat{f} \le \sup\{r > 0 : \xi_n(r) - r \ge 0\} + \cdots,$$

$$\text{Optimal: } \mathbb{E}\hat{f} = \arg\max_{r>0}(\xi_n(r) - r) \pm \cdots.$$

- Can estimate local complexities (fixed points) from data.
- In general, cannot obtain better estimates from data than the fixed point.