Empirical Minimization and Risk Bounds

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

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

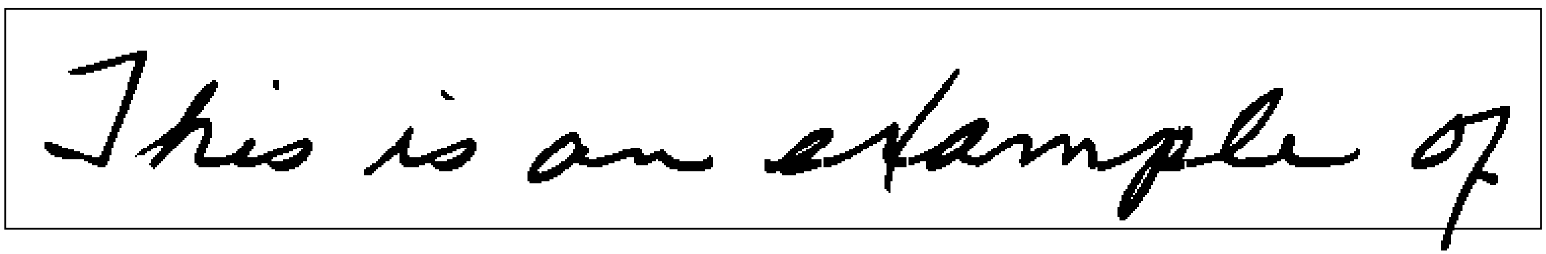
slides at http://www.stat.berkeley.edu/∼bartlett/talks

Empirical risk minimization in prediction problems

Optical Character Recognition:

$X =$ grey-scale image of a sequence of characters

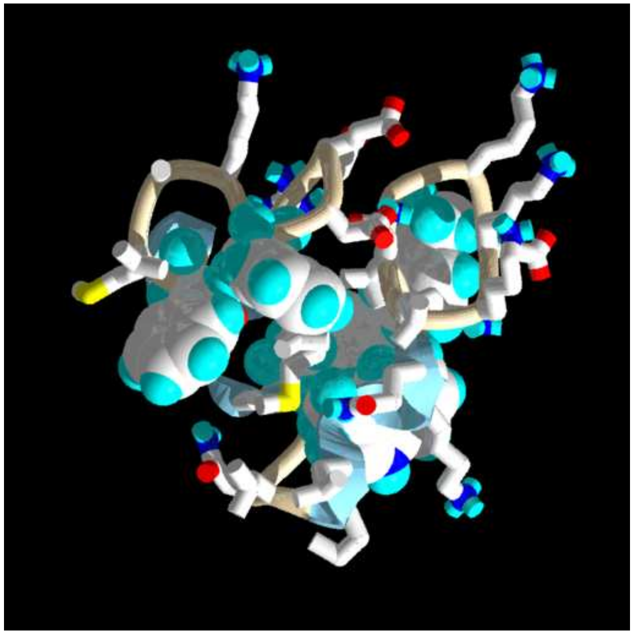
$Y =$ sequence of characters



This	is	an	example	of

Predicting Protein Function

$X =$ amino acid sequence, protein family alignments, protein-protein interactions, gene expression profiles, ...
$Y =$ protein function (metabolism, energy, cell cycle, transcription, ...)



Prediction problems

- 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

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

  or a regularized version.

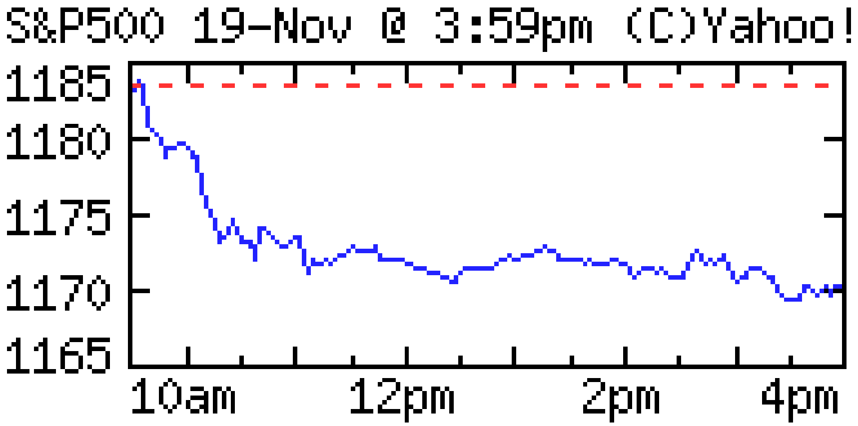
Empirical risk minimization in prediction problems

Financial prediction:

$X =$ features of a financial instrument.

$Y =$ future price.

$\ell(y, \hat{y}) = (y - \hat{y})^2$.



Empirical risk minimization in prediction problems

Anomaly detection in network security:

$X =$ features of network traffic
Aim is to flag anomalous Xs:

$$X \in \left\{ x : \frac{dP}{d\mu}(x) < \epsilon \right\},$$

where P is the probability measure of X and μ is a reference measure.

Equivalent to a classification problem. (Steinwart, Hush, Scovel, 2004)

Empirical Risk Minimization in Large Margin Classifiers

For pattern classification, minimization of empirical risk is often intractable:

- Replace 0-1 loss, ℓ, with a convex surrogate, $\phi : \mathbb{R} \to \mathbb{R}^+$.
- Two-class: $Y \in \{\pm 1\}$.
- Define the **ϕ-risk** of $f : \mathcal{X} \to \mathbb{R}$ as $\mathbb{E}\phi(Yf(X))$.
- Choose f to minimize empirical ϕ-risk,

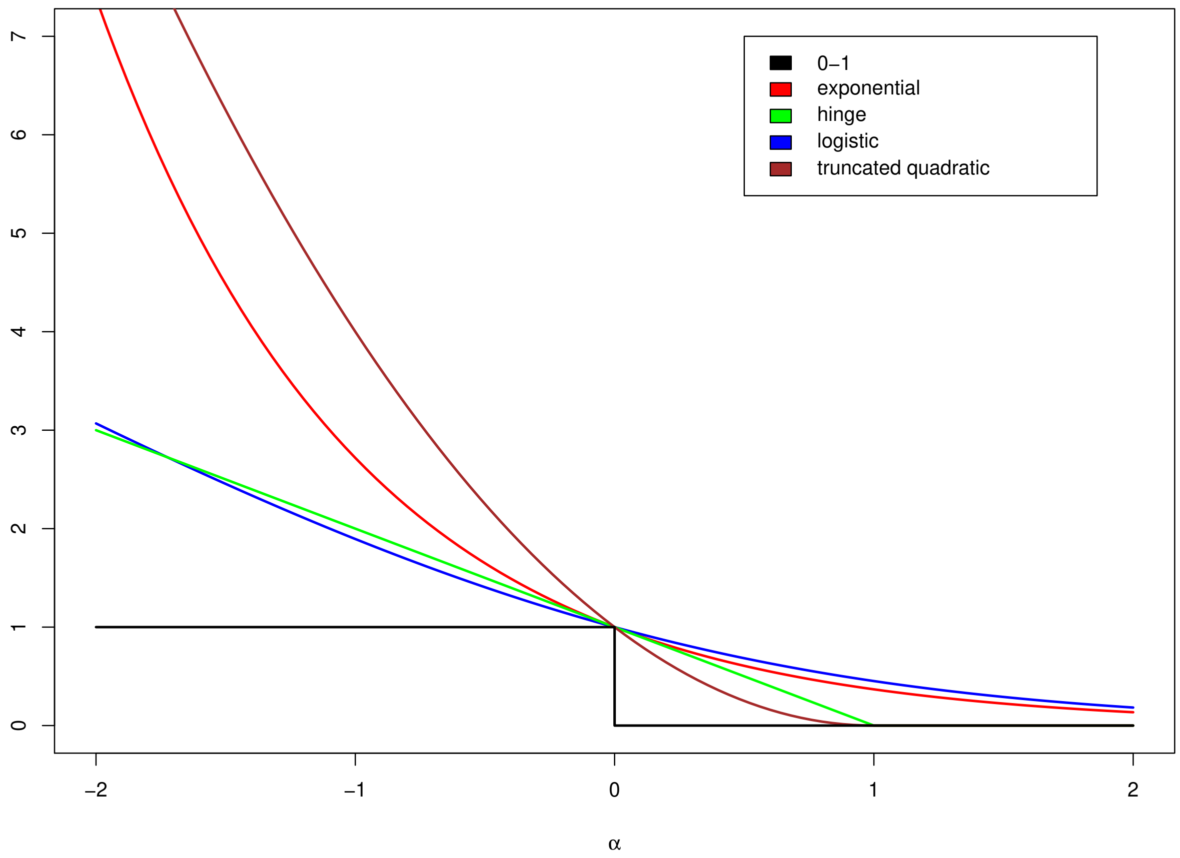
$$\hat{\mathbb{E}}\phi(Yf(X)).$$

- Predict using $\hat{y} = \operatorname{sign}\left(\hat{f}(x)\right)$.

Large Margin Algorithms

- Adaboost:
  - $\mathcal{F} = \mathrm{span}(\mathcal{G})$ for a VC-class $\mathcal{G}$,
  - $\phi(\alpha) = \exp(-\alpha)$,
- Support vector machines with 1-norm soft margin.
  - $\mathcal{F}$ = reproducing kernel Hilbert space, $\mathcal{H}$.
  - $\phi(\alpha) = \max(0, 1 - \alpha)$.
- Neural networks,
- etc

Large Margin Algorithms



Empirical Risk Minimization: Questions

- Consistency?
- Convergence of excess risk to 0:
  What is $\mathbb{E}\ell(Y, \hat{f}(X)) - \inf_{f \in \mathcal{F}} \mathbb{E}\ell(Y, f(X))$?
- Hence, design of algorithms:
  model selection/regularization methods.

Outline

- Empirical minimization, notation.
- Classical risk bounds: uniform convergence.
- Improvements: local complexities.
- Estimating local complexities.
- The risk of the empirical minimizer.

Empirical minimization

- 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 $\hat{\mathbb{E}}g$ over $\mathcal{G}$ is equivalent to empirical risk minimization over $\mathcal{F}$.
- $\mathbb{E}g$ is excess risk.

Empirical Minimization

From now on, we'll consider:

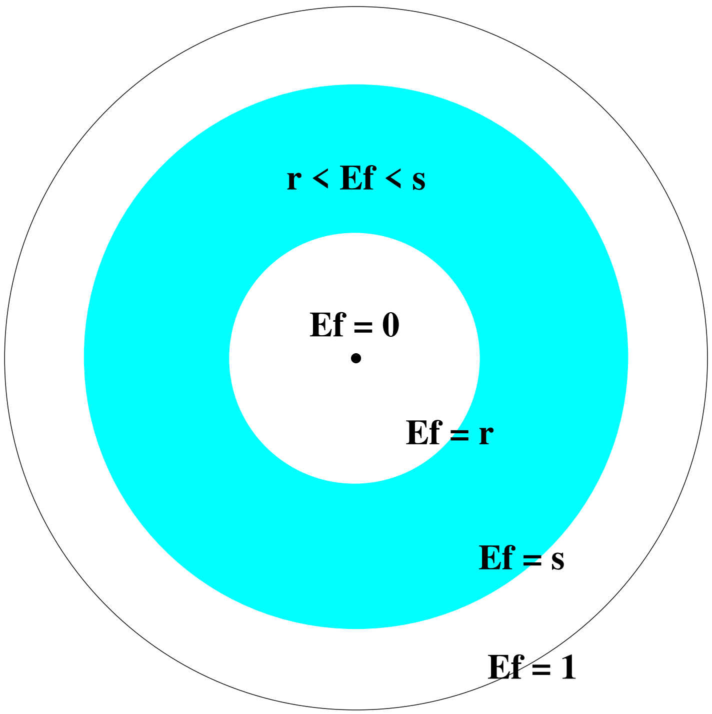
- i.i.d. $X, X_1, \ldots, X_n$ from $\mathcal{X}$,
- a class $\mathcal{F}$ of $[-1, 1]$-valued functions on $\mathcal{X}$ (with $\mathbb{E}f \geq 0$),
- $\hat{f} \in \arg\min_{f \in \mathcal{F}} \hat{\mathbb{E}} f$.

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

Empirical Processes: Notation

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

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



- sup is maximal deviation of expectation from sample average.
- Measures complexity of set.
- Decreases with n.
- Consider *shells* defined by $\mathbb{E}f$.
- Assume $\mathbb{E}f \le 1$.

Risk Bounds

Classical risk bounds are based on uniform convergence:

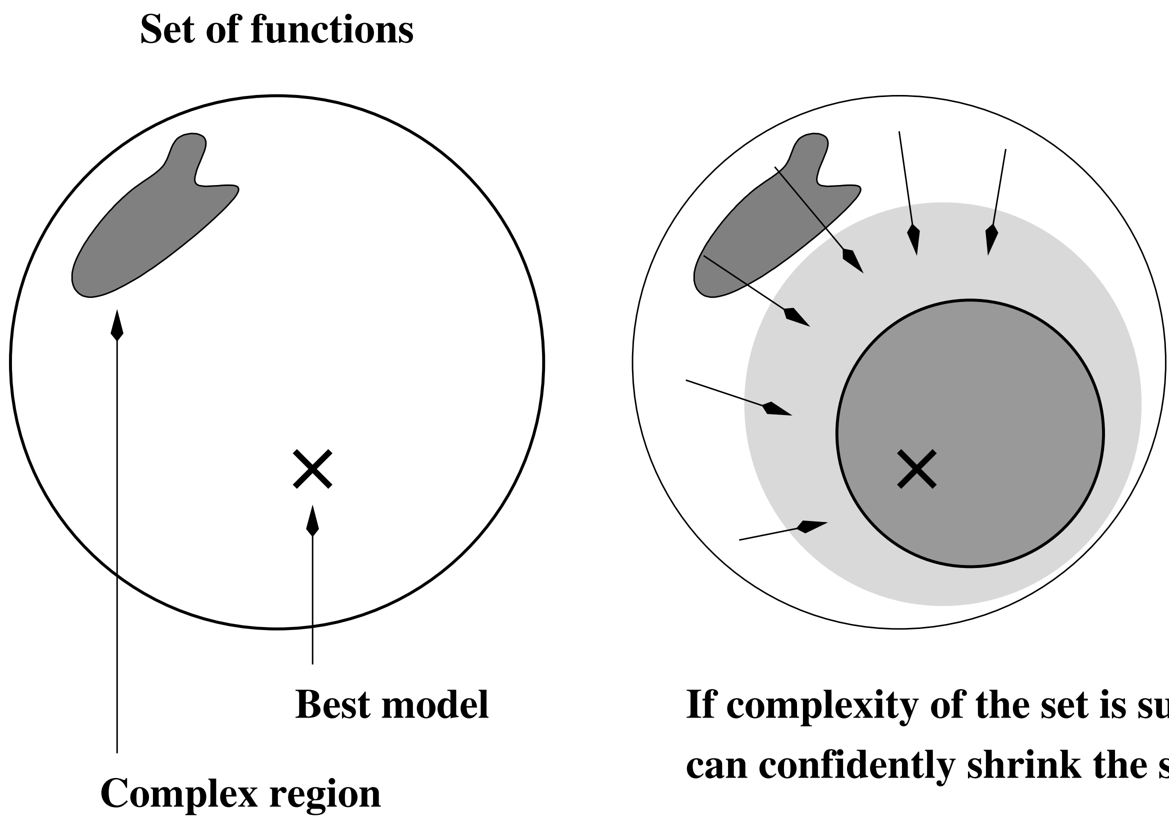
$$\mathbb{E}\hat{f} \leq \xi_n(0,1) + \dots$$

$$\xi_n(0,1) \leq c\sqrt{\frac{\mathrm{VCdim}(\mathcal{F})}{n}}. \qquad \text{(Vapnik and Chervonenkis)}$$

$$\xi_n(0,1) \leq c\int_0^\infty \sqrt{\frac{\log \mathcal{N}(\mathcal{F}, \epsilon, L_2(P))}{n}}\, d\epsilon. \qquad \text{(Dudley)}$$

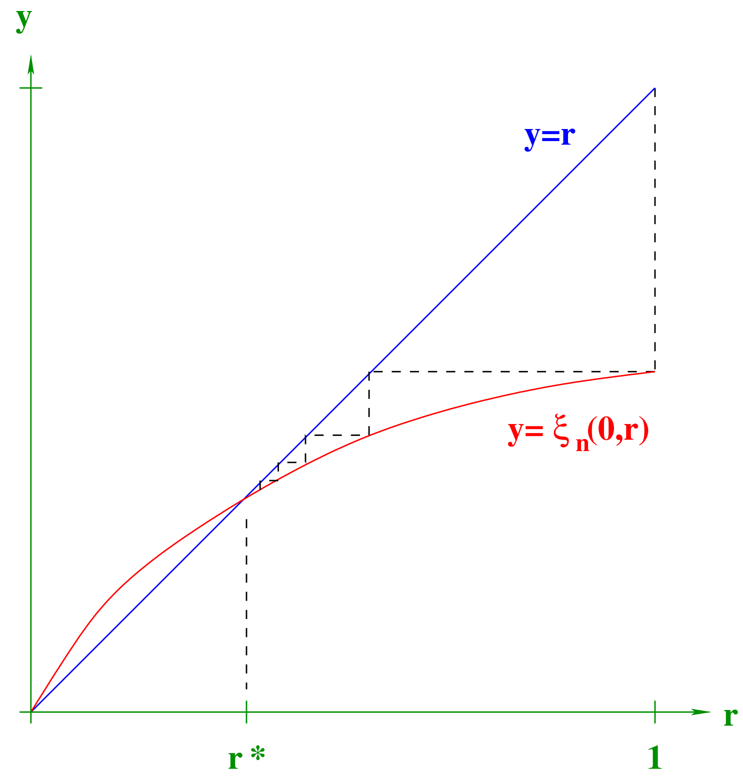
Aim: Tight bounds on $\mathbb{E}\hat{f}$, that do not rely on knowledge of P.

Risk bounds based on local complexity

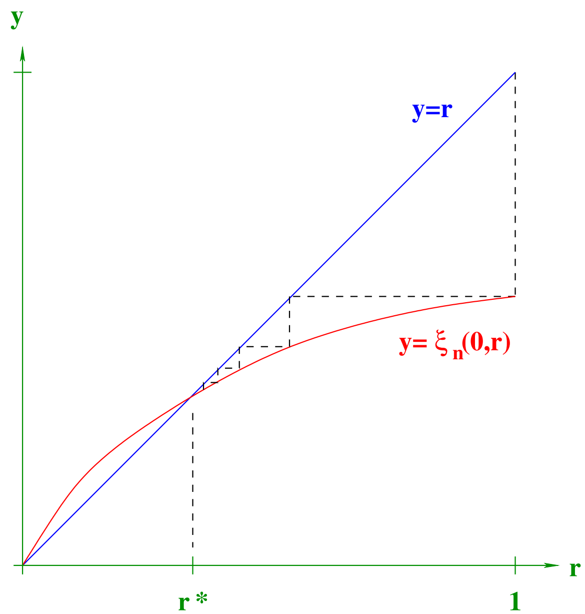


If complexity of the set is sufficiently small, can confidently shrink the set.

Risk bounds based on local complexity



Risk bounds based on local complexity



- Global uniform convergence is stronger than necessary.
- (van de Geer, 2000)
- (Koltchinskii and Panchenko, 2000), (Lugosi and Wegkamp, 2004), (B, Bousquet and Mendelson, 2004), (Koltchinskii, 2004).

Outline

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

c.f. unif. convergence: $\mathbb{E}\hat{f} \leq \sup\{r > 0 : \xi_n(0,1) - r \geq 0\} + \cdots$,

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

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

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

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

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

Examples of Bernstein classes:

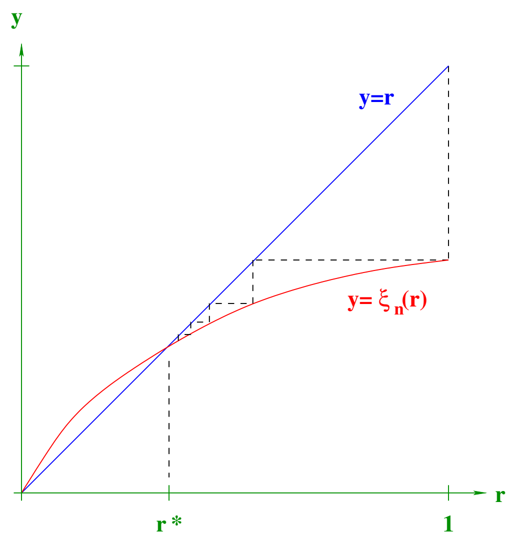
- non-negative functions.
- excess loss class from a convex function class and a strictly convex loss.
- excess loss class for low-noise classification.

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

Risk bounds

Theorem: If $\mathcal{F}$ is bounded, star-shaped, and Bernstein, then with probability at least $1 - e^{-x}$, the empirical minimizer satisfies

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



Recall:

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

Proof idea: Isomorphic coordinate projections

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 $f \in \mathcal{F}$,

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

Recall: $\hat{\mathbb{E}}f = \frac{1}{n}\sum_{i=1}^{n} f(X_i)$.

Isomorphic coordinate projections: Proof idea

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

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

$$\Longrightarrow \Pi_{X_1^n} \text{ is an } \epsilon\text{-isomorphism for } \{f \in \mathcal{F} : \mathbb{E}f = r\}$$

$$\Longrightarrow \xi_n(r) < (1+\alpha)\epsilon r.$$

Proof: Talagrand's functional Bernstein inequality

(the Bernstein property controls the variance term).

Isomorphic coordinate projections: Proof idea

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

Proof: $\mathcal{F}$ star-shaped $\Longrightarrow$
any f with $\mathbb{E}f \geq r$ has a scaled version, f', with $\mathbb{E}f' = r$.

Isomorphic coordinate projections: Proof idea

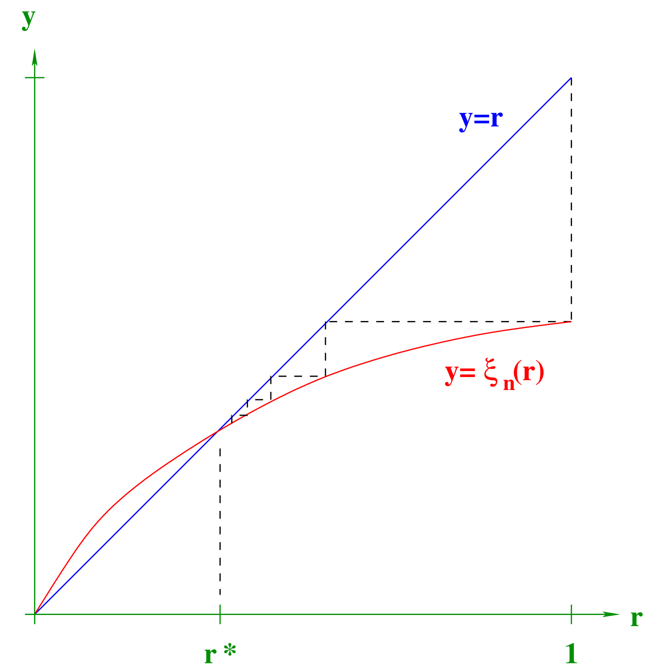
Combining gives:

Theorem: If

$$r \geq 4\xi_n(r) \vee \frac{cx}{n},$$

then with probability at least $1 - e^{-x}$, every $f \in \mathcal{F}$ satisfies

$$\mathbb{E}f \leq 2\hat{\mathbb{E}}f \vee r.$$



But $0 \in \mathcal{F} \implies \hat{\mathbb{E}}\hat{f} \leq 0 \implies \mathbb{E}\hat{f} \leq r$.

Isomorphic coordinate projections

Theorem: If $\mathcal{F}$ is bounded, star-shaped, and Bernstein, then with probability at least $1 - e^{-x}$, the empirical minimizer satisfies

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

Isomorphic coordinate projections

Theorem: If $\mathcal{F}$ is bounded, star-shaped, and Bernstein, 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 $X_1, \ldots, X_n$, define

$$R_n f = \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(X_i) \qquad (\sigma_i \text{ independent, uniform } \{\pm 1\})$$

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

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

(‘empirical Rademacher averages’)

Estimating the fixed point from data

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

Symmetrization:

$$\xi_n(r) = \mathbb{E} \sup \left\{ \mathbb{E}f - \hat{\mathbb{E}}f : \mathbb{E}f = r \right\} \leq 2\mathbb{E}R_n \{ f \in \mathcal{F} : \mathbb{E}f = r \}.$$

Concentration: with prob. $1 - e^{-x}$,

$$2\mathbb{E}R_n \{ f \in \mathcal{F} : \mathbb{E}f = r \} \leq 4\mathbb{E}_\sigma R_n \{ f \in \mathcal{F} : \mathbb{E}f = r \} + \frac{x}{n}$$

Isomorphic coordinate projections: if $r \geq r^* \vee \frac{cx}{n}$, w. prob. $1 - e^{-x}$,

$$\{ f \in \mathcal{F} : \mathbb{E}f = r \} \subseteq \left\{ f \in \mathcal{F} : \frac{r}{2} \leq \hat{\mathbb{E}}f \leq \frac{3r}{2} \right\}$$

Estimating the fixed point from data

Consider the fixed point $\hat{r}^* = \sup\left\{r > 0 : \hat{\xi}_n(r) \geq \frac{r}{4}\right\}$, with

$$\hat{\xi}_n(r) = 4\mathbb{E}_\sigma R_n \left\{f \in \mathcal{F} : \frac{r}{2} \leq \hat{\mathbb{E}} f \leq \frac{3r}{2}\right\} + \frac{r}{c}.$$

Using binary search (with $O(\log n)$ steps), we can compute an accurate estimate $\hat{r}$ of $\hat{r}^*$:

Theorem: If $\mathcal{F}$ is bounded, star-shaped, and Bernstein, then with probability at least $1 - cne^{-x}$, the empirical minimizer satisfies

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

Outline

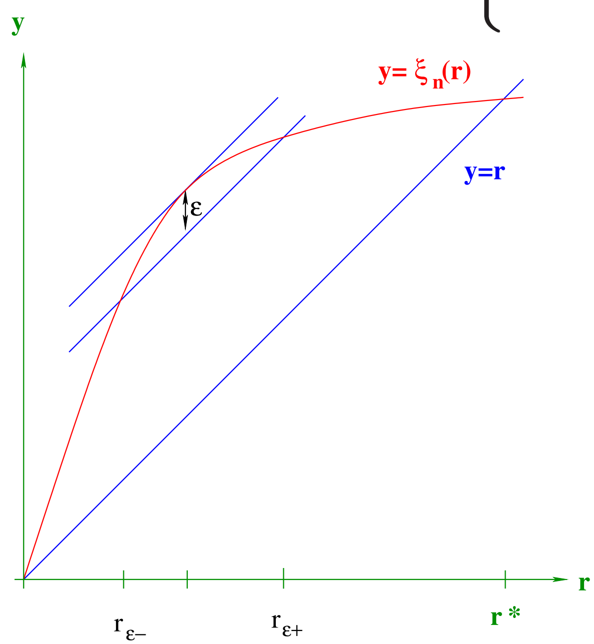
1.	Local, shells:	$\mathbb{E}\hat{f} \leq \sup\{r > 0 : \xi_n(r) - r \geq 0\} + \cdots$.
c.f.	unif. convergence:	$\mathbb{E}\hat{f} \leq \sup\{r > 0 : \xi_n(0,1) - r \geq 0\} + \cdots$,
and	local balls:	$\mathbb{E}\hat{f} \leq \sup\{r > 0 : \xi_n(0,r) - r \geq 0\} + \cdots$.
2.		Estimating the fixed point, $\xi_n(r) = r$ from data, using Rademacher averages.
3.	Optimal:	$\mathbb{E}\hat{f} = \arg\max_{r>0}(\xi_n(r) - r) \pm \cdots$.
4.	Examples:	The improvement can be enormous. But in general, the better bound cannot be estimated from data.

A near-optimal bound

Define:

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

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



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 $1/n \leq r \leq r_{\epsilon,-}$.

Outline

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

c.f. unif. convergence: $\mathbb{E}\hat{f} \leq \sup\{r > 0 : \xi_n(0,1) - r \geq 0\} + \cdots$,

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

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

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

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

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

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

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

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

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 and probability distributions (i.e., not varying with n).

Summary

- Empirical risk minimization methods for classification, regression, anomaly detection.
- Risk bounds:

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

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

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

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

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