Empirical Entropy, Minimax Regret and Minimax Risk

Alexandre Tsybakov,
Laboratoire de Statistique, CREST-ENSAE
joint work with **Karthik Sridharan** and **Alexander Rakhlin**

Moscow, June 26, 2013

Goal

Understand the relationship between minimax rates for the problems of learning and estimation.

Problem I: Estimation

Model:

$$Y_i = f(X_i) + \xi_i$$

where ξ_i satisfy $\mathbb{E}(\xi_i | X_i) = 0$ and $f \in \mathcal{F}$.

- $\mathcal{X}$ is any set and $\mathcal{Y} \subseteq \mathbb{R}$
- $\mathcal{F}$ is a class of functions from $\mathcal{X}$ to $\mathcal{Y}$
- $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ i.i.d. from P_{XY} on $\mathcal{X} \times \mathcal{Y}$

Fix marginal P_X and conditional distribution of $\xi = Y - f(X)$ given X

Minimax risk:

$$W_n(\mathcal{F}) = \inf_{\hat{f}} \sup_{f \in \mathcal{F}} \mathbb{E}_f \|\hat{f} - f\|^2$$

where the norm is $L_2(P_X)$

Problem II: Statistical Learning

Model: any distribution P_{XY}

Expected loss of f:

$$L(f) = \mathbb{E}_{XY}[(f(X) - Y)^2]$$

Minimax Regret:

$$V_n(\mathcal{F}) = \inf_{\hat{f}} \sup_{P_{XY} \in \mathcal{P}} \left\{ \mathbb{E}L(\hat{f}) - \inf_{f \in \mathcal{F}} L(f) \right\}$$

Statistical Learning and Oracle Inequalities

Minimization of $L(f)$ over all measurable functions yields

$$\eta(x) = \mathbb{E}_{XY}[Y|X = x]$$

and

$$\mathbb{E}L(\hat{f}) - \inf_{f \in \mathcal{F}} L(f) = \mathbb{E}\|\hat{f} - \eta\|^2 - \inf_{f \in \mathcal{F}} \|f - \eta\|^2$$

Hence, minimax regret can be written as

$$V_n(\mathcal{F}) = \inf_{\hat{f}} \sup_{P_{XY} \in \mathcal{P}} \left\{ \mathbb{E}\|\hat{f} - \eta\|^2 - \inf_{f \in \mathcal{F}} \|f - \eta\|^2 \right\}$$

Upper bounds are known as *exact oracle inequalities*.

Well-Specified vs Misspecified Models

Once again,

$$V_n(\mathcal{F}) = \inf_{\hat{f}} \sup_{P_{XY} \in \mathcal{P}} \left\{ \mathbb{E}\|\hat{f} - \eta\|^2 - \inf_{f \in \mathcal{F}} \|f - \eta\|^2 \right\}$$

and

$$W_n(\mathcal{F}) = \inf_{\hat{f}} \sup_{f \in \mathcal{F}} \mathbb{E}_f \|\hat{f} - f\|^2$$

Clearly,

$$W_n(\mathcal{F}) \leq V_n(\mathcal{F})$$

Question

Is there a gap in the rates?

What is known: well-specified model

The behavior of minimax risk $W_n(\mathcal{F})$ has been analyzed to a great extent in the past 30 years. Entropic conditions on $\mathcal{F}$ go back to (Ibragimov and Khasminskii 1980), (Birgé, 1983), (Le Cam 1973, 1986), (Van de Geer 1990), (Birgé and Massart 1994), (Yang and Barron 1999), ...

What is known: misspecified model

The precise behavior of $V_n(\mathcal{F})$ is known for aggregation problem (T., 03):

- Finite class $\mathcal{F} = \{f_1, \dots, f_M\}$: $\frac{\log(M)}{n}$
- Convex hull of a finite class conv $(\{f_1, \dots, f_M\})$:

$$\frac{M}{n} \wedge \sqrt{\frac{\log(1 + M/\sqrt{n})}{n}}$$

- Linear combinations: $\frac{M}{n}$
- Sparse convex combinations - up to logs - (Lounici 08):

$$\frac{s \log(eM/s)}{n} \wedge \sqrt{\frac{\log M}{n}}$$

(Rigollet and T. 2011) [fixed design]

What is known: upper bounds

- For VC-classes $\mathcal{F}$ that are convex (Lee et al. 1998):

$$V_n(\mathcal{F}) = O\left(\frac{\log n}{n}\right)$$

or with $L^* = 0$ (Vapnik and Chervonenkis 1981):

$$V_n(\mathcal{F}) = O\left(\frac{1}{n}\right)$$

Empirical distance on the sample S:

$$d_S(f,g) = \left(\frac{1}{n}\sum_{(x,y)\in S}(f(x)-g(x))^2\right)^{1/2}$$

For VC-classes the empirical covering numbers satisfy:

$$\mathcal{N}(\mathcal{F},\epsilon,d_S) = O(\epsilon^{-V})$$

.

What is known: upper bounds

- For convex classes $\mathcal{F}$ with entropy condition $\log \mathcal{N}(\mathcal{F}, \epsilon, d_S) = O(\epsilon^{-p})$:

$$V_n(\mathcal{F}) = O\left(n^{-\frac{2}{2+p}}\right) \quad , \quad p \in (0,2) \quad \text{(Mendelson, Koltchinskii...)}$$

  Example: $p = d/\beta$ (dimension/smoothness) $\longrightarrow n^{-\frac{2\beta}{2\beta+d}}$

- Without the convexity assumption, Koltchinskii 2011 obtained non-sharp oracle inequalities for ERM. Case $p > 2$?
- There seems to be a gap between sharp and non-sharp inequalities for ERM.

What methods are used for the misspecified case?

- empirical risk minimization
- penalized ERM
- mixtures (e.g. exponential weights)
- ...

Interestingly

- Selectors are known to be suboptimal even for finite class.
- ERM fails for non-convex cases.

Assumption

Bounded noise and $\mathcal{F}$ is a class of bounded functions.

- For simplicity, we will work with $\mathcal{Y} = [0, 1]$.
- Random design
- Will have $\log(1/\delta)$ dependence on confidence δ

Proposed method called the aggregation-of-leaders procedure:

- split sample into three equal parts
- use first to construct empirical cover of $\mathcal{F}$
- *use second to find empirical minimizers over the resulting partitions*
- use third to aggregate empirical minimizers

Split sample $D_{3n} = S \cup S' \cup S''$. Define empirical distance on the X-part of S:

$$d_S(f, g) = \left(\frac{1}{n} \sum_{(x,y) \in S} (f(x) - g(x))^2 \right)^{1/2}$$

Fix $\epsilon > 0$ and let $N = \mathcal{N}_2(\mathcal{F}, \epsilon, d_S)$ Fix some cover centers $c_1, \ldots, c_N$ and define partition

$$\hat{\mathcal{F}}_i^S(\epsilon) = \hat{\mathcal{F}}_i^S = \left\{ f \in \mathcal{F} \ : \ i \in \arg \min_{j \in \{1, \ldots, N\}} d_S(f, c_j) \right\}$$

Find ERM's

$$\hat{f}_i^{S,S'} \in \arg \min_{f \in \hat{\mathcal{F}}_i^S} \ \frac{1}{n} \sum_{(x,y) \in S'} (f(x) - y)^2$$

for each $i \in \{1, \ldots, N\}$ and using S'' define an aggregate

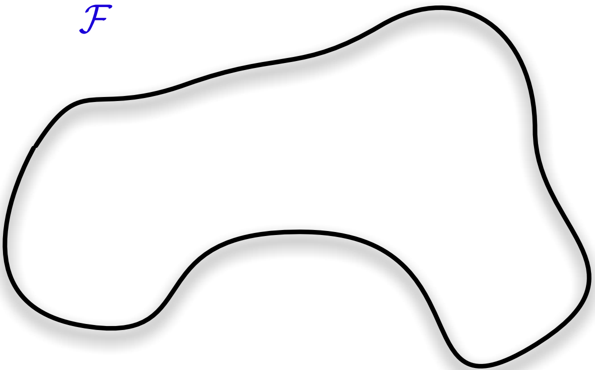
$$\hat{f} = \sum_{i=1}^{N} p_i \hat{f}_i^{S,S'}$$

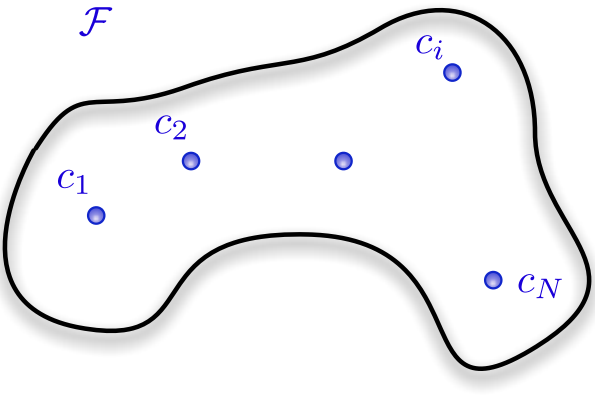
High-probability aggregation result

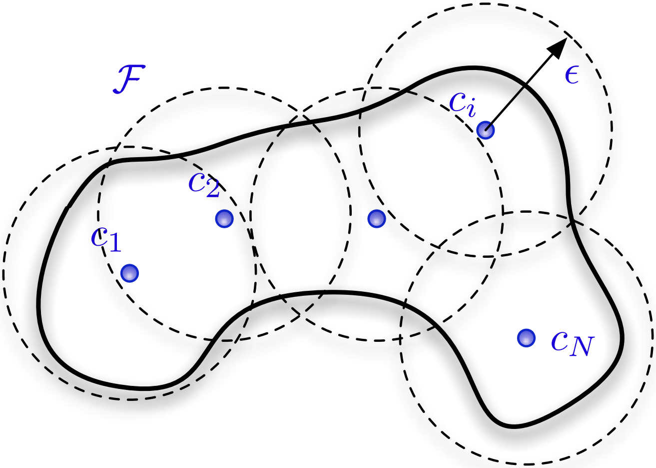
From (Audibert 2007, Lecué and Mendelson 2008, Lecué and Rigollet 2012):

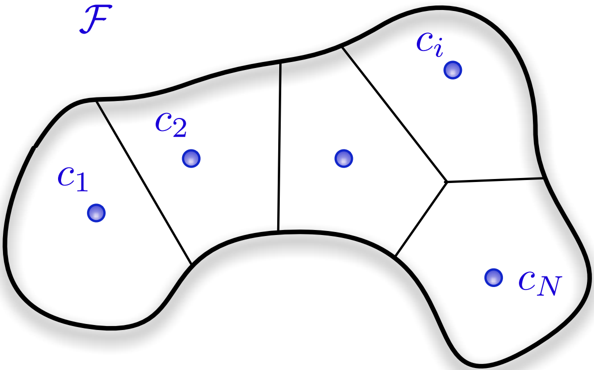
$$L(\hat{f}) \leq \inf_{i=1,\dots,N} L(\hat{f}_i^{S,S'}) + C\frac{\log(N/\delta)}{n}$$

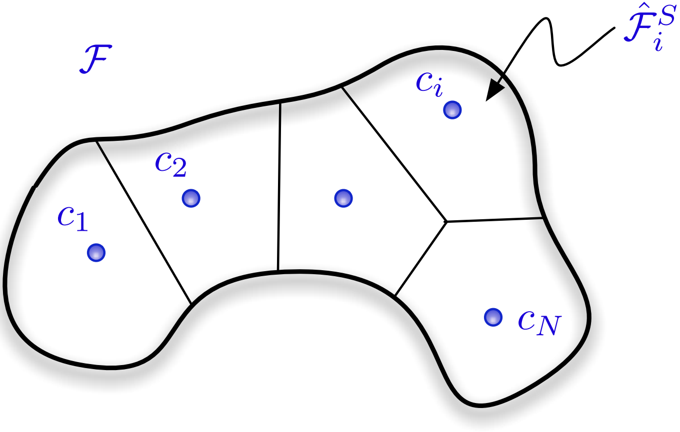
with probability at least $1-\delta$ over the sample S''.

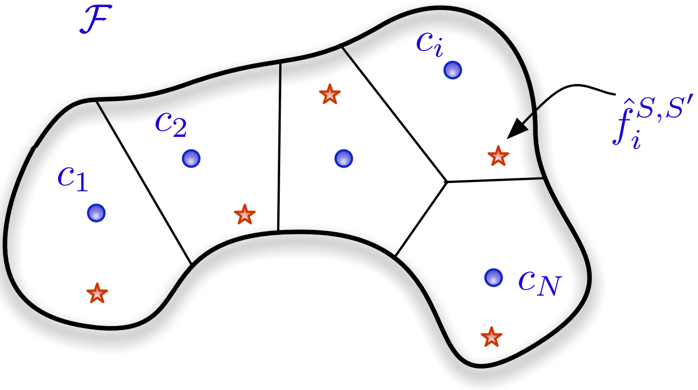

$\mathcal{F}$

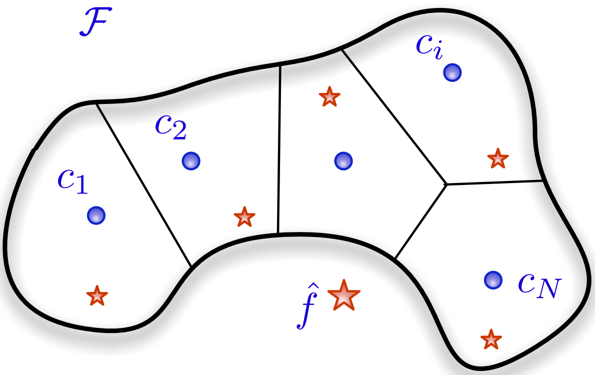

$\mathcal{F}$
c_i
c_2
c_1
c_N

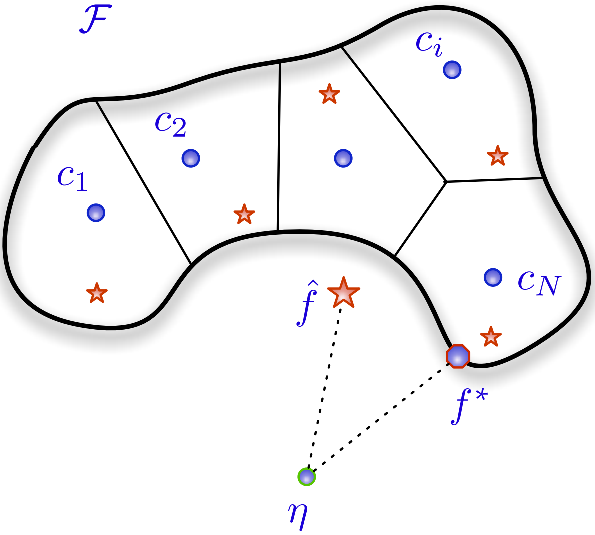

$\mathcal{F}$
c_i
ϵ
c_2
c_1
c_N


$\mathcal{F}$
c_i
c_2
c_1
c_N


$\hat{\mathcal{F}}_i^S$
$\mathcal{F}$
c_i
c_2
c_1
c_N


$\mathcal{F}$
c_i
c_2
c_1
c_N
$\hat{f}_i^{S,S'}$


$\mathcal{F}$
c_i
c_2
c_1
$\hat{f}$
c_N


$\mathcal{F}$
c_i
c_2
c_1
c_N
$\hat{f}$
f^*
η

Main Result

In the case of polynomial growth $O(\epsilon^{-p})$ of empirical entropy $\log \mathcal{N}_2(\mathcal{F}, \epsilon, d_S)$,

- For misspecified models (sharp oracle inequalities), the rates obtained by the proposed method with $\epsilon = n^{-\frac{1}{2+p}}$ are
  - $V_n(\mathcal{F}) = O\left(n^{-\frac{2}{2+p}}\right)$ for $p \in (0, 2]$
  - $V_n(\mathcal{F}) = O\left(n^{-1/p}\right)$ for $p \in (2, \infty)$
- For well-specified models, the same method attains the rate $W_n(\mathcal{F}) = O\left(n^{-\frac{2}{2+p}}\right)$ for all $p > 0$.

For polynomial growth $\left(\frac{1}{\epsilon}\right)^v$ of covering numbers $\mathcal{N}_2(\mathcal{F}, \epsilon, d_S)$, method attains

- $V_n(\mathcal{F}) = O\left(\frac{v \log(n/v)}{n}\right)$ rates for VC subgraph classes
- $O\left(\frac{s \log(eM/s)}{n} \wedge \sqrt{\frac{\log(1+M/\sqrt{n})}{n}}\right)$ for s-sparse convex aggregation.

Main Result

In the case of polynomial growth $O(\epsilon^{-p})$ of empirical entropy $\log \mathcal{N}_2(\mathcal{F}, \epsilon, d_S)$,

- For misspecified models (sharp oracle inequalities), the rates obtained by the proposed method with $\epsilon = n^{-\frac{1}{2+p}}$ are
  - $V_n(\mathcal{F}) = O\left(n^{-\frac{2}{2+p}}\right)$ for $p \in (0, 2]$
  - $V_n(\mathcal{F}) = O\left(n^{-1/p}\right)$ for $p \in (2, \infty)$
- For well-specified models, the same method attains the rate $W_n(\mathcal{F}) = O\left(n^{-\frac{2}{2+p}}\right)$ for all $p > 0$.

For polynomial growth $\left(\frac{1}{\epsilon}\right)^v$ of covering numbers $\mathcal{N}_2(\mathcal{F}, \epsilon, d_S)$, method attains

- $V_n(\mathcal{F}) = O\left(\frac{v \log(n/v)}{n}\right)$ rates for VC subgraph classes
- $O\left(\frac{s \log(eM/s)}{n} \wedge \sqrt{\frac{\log(1+M/\sqrt{n})}{n}}\right)$ for s-sparse convex aggregation.

Main Result

In the case of polynomial growth $O(\epsilon^{-p})$ of empirical entropy $\log \mathcal{N}_2(\mathcal{F}, \epsilon, d_S)$,

- For misspecified models (sharp oracle inequalities), the rates obtained by the proposed method with $\epsilon = n^{-\frac{1}{2+p}}$ are
  - $V_n(\mathcal{F}) = O\left(n^{-\frac{2}{2+p}}\right)$ for $p \in (0, 2]$
  - $V_n(\mathcal{F}) = O\left(n^{-1/p}\right)$ for $p \in (2, \infty)$
- For well-specified models, the same method attains the rate $W_n(\mathcal{F}) = O\left(n^{-\frac{2}{2+p}}\right)$ for all $p > 0$.

For polynomial growth $\left(\frac{1}{\epsilon}\right)^v$ of covering numbers $\mathcal{N}_2(\mathcal{F}, \epsilon, d_S)$, method attains

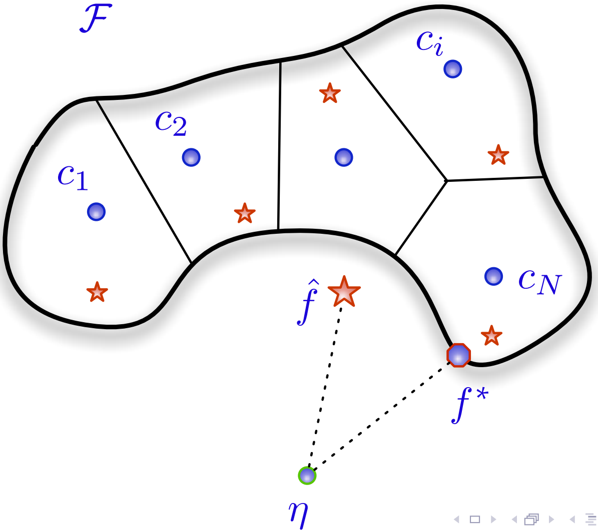
- $V_n(\mathcal{F}) = O\left(\frac{v \log(n/v)}{n}\right)$ rates for VC subgraph classes
- $O\left(\frac{s \log(eM/s)}{n} \wedge \sqrt{\frac{\log(1+M/\sqrt{n})}{n}}\right)$ for s-sparse convex aggregation.

Lower Bounds

- Lower bounds of $n^{-\frac{2}{2+p}}$ in the well-specified case are obtained in e.g. (Yang & Barron 1999)
- There exists a VC-subgraph class $\mathcal{F}$ with VC dimension v such that

$$W_n(\mathcal{F}) \geq C \frac{v \log(n/v)}{n}$$

Remark: On importance of ERM in partitions



Aggregation step:

$$L(\hat{f}) \leq \inf_j L(\hat{f}_j^{S,S'}) + C\frac{\log(N/\delta)}{n}$$

with probability at least $1-\delta$ over the sample S''.

$$L(\hat{f}) - \inf_j L(\hat{f}_j^{S,S'}) = \|\hat{f} - \eta\|^2 - \inf_j \|\hat{f}_j^{S,S'} - \eta\|^2$$

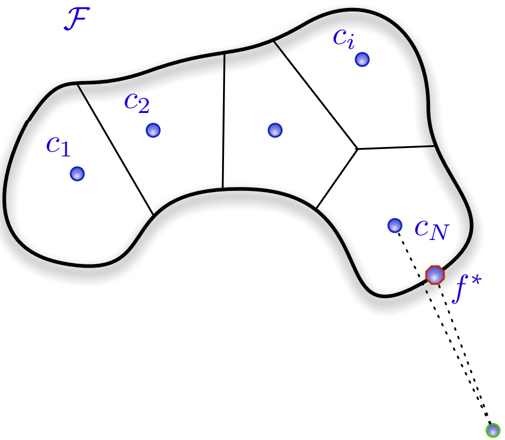
Skeleton aggregation: aggregate elements c_j of ϵ-net.

$$L(\hat{f}) \leq \inf_j L(c_j) + C\frac{\log(N/\delta)}{n}$$

with probability at least $1-\delta$ over the sample S''.

$$L(\hat{f}) - \inf_j L(c_j) = \|\hat{f} - \eta\|^2 - \inf_j \|c_j - \eta\|^2$$

Skeleton aggregation



However $(a+\epsilon)^2 - a^2 = O(\epsilon)$ and balancing $\epsilon = \frac{\log \mathcal{N}(\mathcal{F}, \epsilon, d_S)}{n}$ yields a wrong rate.

Remark: On importance of ERM in partitions

This does not happen for well-specified models! (e.g. (Birgé 1983), (Yang and Barron 1999)): $a = 0$ and $(a+\epsilon)^2 - a^2 = \epsilon^2$.

The minimax rate arises from

$$\epsilon^2 = \frac{\log \mathcal{N}(\mathcal{F}, \epsilon, d_S)}{n}.$$

Emerging Picture: misspecified case

Regime	our $\hat{f}$	Skeleton agg.	ERM
(Finite) $\lvert\mathcal{F}\rvert = M$	$\frac{\log M}{n}$	$\frac{\log M}{n}$	$\sqrt{\frac{\log M}{n}}$
("parametric") $VC(\mathcal{F}) = d$	$\frac{d \log(n/d)}{n}$	$\sqrt{\frac{d \log(n/d)}{n}}$	$\sqrt{\frac{d}{n}}$
$\log \mathcal{N}_2(\mathcal{F}, \epsilon) = \epsilon^{-p}$, $p \in (0, 2]$	$n^{-\frac{2}{2+p}}$	$\gg n^{-\frac{1}{p+1}} \vee n^{-\frac{1}{2}}$	$n^{-\frac{1}{2}}$
$p \in (2, \infty)$	$n^{-\frac{1}{p}}$	$n^{-\frac{1}{p+1}}$	$n^{-\frac{1}{p}}$

Discussion

- For finite class $\mathcal{F}$ aggregation-of-leaders and skeleton aggregation achieve the optimal excess risk rate $\frac{\log M}{n}$. Global ERM has a suboptimal rate.
- For very massive $\mathcal{F}$, when the empirical entropy is ϵ^{-p} with $p \geq 2$ both ERM and aggregation-of-leaders have the rate $n^{-1/p}$. Skeleton aggregation is suboptimal.
- For all other cases: aggregation-of-leaders is optimal, both ERM and skeleton aggregation are suboptimal.
- Unless $\mathcal{F}$ is finite, skeleton aggregation does not improve upon ERM in the misspecified case.
- Well-specified case. Aggregation-of-leaders and skeleton aggregation achieve the optimal rate for the minimax risk. The global ERM is, in general, suboptimal.

Including the approximation error when $p > 2$

Theorem

Let $\mathcal{Y} = [0,1]$, $\mathcal{F} \subseteq \{f : 0 \le f \le 1\}$, *and* $\log \mathcal{N}_2(\mathcal{F}, \rho) \le A\rho^{-p}$, $\forall\, \rho > 0$, *with* $p > 2$. *Consider an aggregation-of-leaders estimator* $\hat{f}$ *with the covering radius* $\epsilon = n^{-\frac{1}{2+p}}$. *For any joint distribution* P_{XY} *:*

$$\mathbb{E}\|\hat{f} - \eta\|^2 - \inf_{f \in \mathcal{F}} \|f - \eta\|^2 \le C_p \bar{\psi}_{n,p}(\Delta)$$

where $\Delta^2 = \inf_{f \in \mathcal{F}} \|f - \eta\|^2$, $C_p > 0$ *is a constant depending only on* p *and* A, *and*

$$\bar{\psi}_{n,p}(\Delta) = \begin{cases} n^{-\frac{2}{2+p}} & \textit{if } \Delta^2 \le n^{-2/(2+p)}, \\ \Delta^2 & \textit{if } n^{-2/(2+p)} \le \Delta^2 \le n^{-1/p}, \\ n^{-1/p} & \textit{if } \Delta^2 \ge n^{-1/p}. \end{cases}$$

Linking Statistical Learning and Estimation

Introduce the class of Δ-misspecified models

$$\mathcal{P}_\Delta(\mathcal{F}) = \left\{ P_{XY} \in \mathcal{P} : \inf_{f \in \mathcal{F}} \|f - \eta\| \le \Delta \right\}, \quad \Delta \ge 0,$$

and define the Δ-*misspecified regret* as

$$V_n^\Delta(\mathcal{F}) = \inf_{\hat{f}} \sup_{P_{XY} \in \mathcal{P}_\Delta(\mathcal{F})} \left\{ \mathbb{E}\|\hat{f} - \eta\|^2 - \inf_{f \in \mathcal{F}} \|f - \eta\|^2 \right\}.$$

- $V_n^\Delta(\mathcal{F})$ measures the minimax regret for statistical estimation problem with approximation error $\le \Delta$.
- By definition, $V_n^\Delta(\mathcal{F}) = W_n(\mathcal{F})$ when $\Delta = 0$ and $V_n^\Delta(\mathcal{F}) = V_n(\mathcal{F})$ when $\Delta = 1$ (the diameter of $\mathcal{F}$).
- Case $p > 2$. By the above theorem, Δ-misspecified regret admits the bound $V_n^\Delta(\mathcal{F}) \le C_p \bar{\psi}_{n,p}(\Delta)$.
- Smooth transition between learning and estimation rates.

Contributions

- Rates of estimation and learning match for $p \leq 2$
- Single algorithm obtains optimal rates in both well-specified and misspecified cases, for all regimes
- Adaptation: prior knowledge of well-specified vs misspecified model is not needed
- Optimal rates of aggregation
- ERM over partitions step is crucial: skeleton aggregation fails in misspecified case.

Theorem

Let $\mathcal{Y} = [0,1]$ *and* $0 \le f \le 1$ *for all* $f \in \mathcal{F}$. *Let* $r^* = r^*(\mathcal{G})$ *denote a localization radius of* $\mathcal{G} = \{(f-g)^2 : f, g \in \mathcal{F}\}$. *Consider an aggregation-of-leaders estimator* $\hat{f}$. *Then* $\exists\ C > 0$ *such that for any* $\delta > 0$, *with probability at least* $1 - \delta$,

$$L(\hat{f}) \le \inf_{f \in \mathcal{F}} L(f) + C\left(\frac{\log(\mathcal{N}_2(\mathcal{F}, \epsilon, d_S)/\delta)}{n} + \Xi(n, \epsilon, S')\right),$$

$$\Xi(n, \epsilon, S') = \gamma\sqrt{r^*} + \frac{1}{\sqrt{n}}\int_0^{C\gamma} \sqrt{\log \mathcal{N}_2(\mathcal{F}, \rho, d_{S'})}\, d\rho$$

with $\gamma = \sqrt{\epsilon^2 + r^* + \beta}$ *and* $\beta = (\log(1/\delta) + \log\log n)/n$.