Review: Localization analysis

IDEA lab

Department of Statistics, Seoul National University

May 20, 2026

Introduction

- Our goal is to characterize the upper bounds of the excess risk of ERM.
- Classical approaches rely on the global complexity of the function class, which often yields slow rates.
- By localizing the function class, we can obtain more sharp excess risk bounds.

Outline

Setup

- $(\mathcal{I}, \mathfrak{A}, P)$: probability space
- $Z_{1:n} := (Z_1, \ldots, Z_n)$: observations
- P : true distribution, P_n: empirical distribution
- $\mathcal{F}$: class of measurable functions from $\mathcal{I}$ to a measurable space $\mathcal{S}$
- $\ell : \mathcal{I} \times \mathcal{S} \to [0, \infty)$: loss function
- $\ell \circ f : \mathcal{I} \to [0, \infty)$ be a function s.t. $\ell \circ f(z) = \ell(z, f(z))$ for any $z \in \mathcal{I}$.

Setup

- $P[\ell \circ f]$: Risk of f

  $P_n[\ell \circ f]$: Empirical risk of f

- $f^\star \in \arg\min_{f \in \mathcal{F}} P[\ell \circ f]$: Bayes Predictor

  $\hat{f}_n \in \arg\min_{f \in \mathcal{F}} P_n[\ell \circ f]$: ERM (Empricial Risk Minimizer)

- Excess Risk of f

$$\mathcal{E}(f) := P[\ell \circ f] - P[\ell \circ f^\star]$$

- Assume that there exists an absolute constant $B > 0$ such that $\|\ell \circ f - \ell \circ f^\star\|_\infty \leq B$ for all $f \in \mathcal{F}$.

Outline

Global Analysis $\Rightarrow$ slow rate

- From the Basic Inequality($P_n(\ell \circ \hat{f}_n) \leq P_n(\ell \circ f)$), when $f^\star \in \mathcal{F}$,

$$\begin{aligned}\mathcal{E}(\hat{f}_n) &= P[\ell \circ \hat{f}_n] - P[\ell \circ f^\star] \\ &= (P - P_n)[\ell \circ \hat{f}_n - \ell \circ f^\star] + P_n[\ell \circ \hat{f}_n - \ell \circ f^\star] \\ &\leq (P - P_n)[\ell \circ \hat{f}_n - \ell \circ f^\star]\end{aligned}$$

- Employing Global bound:

$$\mathcal{E}(\hat{f}_n) \;\leq\; \sup_{f \in \mathcal{F}} \left|(P_n - P)\big(\ell \circ f - \ell \circ f^\star\big)\right|.$$

Talagrand

Lemma 2.1

Let $\mathcal{G}$ be a class of functions on $\mathcal{I}$ such that $\sup_{g\in\mathcal{G}} \|g\|_\infty \le B$ for some $B > 0$. Then

$$\mathbb{P}\left[\sup_{g\in\mathcal{G}}(P-P_n)[g] \;\ge\; 2\mathbb{E}[\sup_{g\in\mathcal{G}}(P-P_n)[g]] + \sqrt{\frac{2t}{n}\sup_{g\in\mathcal{G}}\mathrm{Var}_P(g)} + \frac{4Bt}{3n}\right] \le e^{-t} \qquad (1)$$

- Replace $\mathcal{G}$ with $\mathcal{L}(\mathcal{F}_{-f^\star}) := \{\ell\circ f - \ell\circ f^\star \;\; : f\in\mathcal{F}\}$
- Applying Lemma2.1, we can control Global Upper Bound:

$$\sup_{f\in\mathcal{F}} |(P_n - P)(\ell\circ f - \ell\circ f^\star)| \;\approx\; O_p(n^{-1/2})$$

$\Rightarrow$ **slow rate**

Outline

Main idea: focusing on a "good" local region

- A standard approach controls

$$\sup_{f \in \mathcal{F}} |(P_n - P)(\ell \circ f - \ell \circ f^\star)|$$

  to derive an upper bound on the excess risk.
- However, taking the supremum over the whole class $\mathcal{F}$ is not tight enough, so we introduce a **localized class**:

$$\mathcal{F}(\delta) := \{f \in \mathcal{F} : \mathcal{E}(f) \leq \delta\}.$$

- The goal is to control the empirical process over this smaller class:

$$\sup_{f \in \mathcal{F}(\delta)} |(P_n - P)(\ell \circ f - \ell \circ f^\star)|$$

  as precisely as possible in terms of δ.
- Intuition: If $\mathcal{E}(f)$ is small, then f is close to $f^\star$, and the complexity near $f^\star$ is much smaller than the global complexity.[1]

[1] The right neighborhood size is determined by a fixed point, which will be explained later.

Intuition: find δ as small as possible

- Assumption 1 states that
$$\sup_{f \in \mathcal{F}} \|\ell \circ f - \ell \circ f^\star\|_\infty \leq B.$$
- In the bound from (1), we apply it with $g = \ell \circ f - \ell \circ f^\star$ and $\mathcal{G} = \mathcal{F}(\delta)$.
$\rightarrow$ The resulting bound depends on n, δ, and t. We denote it by
$U_n(\delta, t) := 2\mathbb{E}\Big[\sup_{f \in \mathcal{F}(\delta)} (P - P_n)[f]\Big] + \sqrt{\frac{2t}{n} \sup_{f \in \mathcal{F}(\delta)} \mathrm{Var}_P(f)} + \frac{4Bt}{3n}$.

Intuition: find δ as small as possible

- Assumption 1 states that
$$\sup_{f \in \mathcal{F}} \|\ell \circ f - \ell \circ f^\star\|_\infty \le B.$$
- In the bound from (1), we apply it with $g = \ell \circ f - \ell \circ f^\star$ and $\mathcal{G} = \mathcal{F}(\delta)$.
$\rightarrow$ The resulting bound depends on n, δ, and t. We denote it by
$U_n(\delta, t) := 2\mathbb{E}\Big[\sup_{f \in \mathcal{F}(\delta)} (P - P_n)[f]\Big] + \sqrt{\frac{2t}{n} \sup_{f \in \mathcal{F}(\delta)} \mathrm{Var}_P(f)} + \frac{4Bt}{3n}$.
- First, set $\delta_1 := B$, so that $\mathcal{F}(\delta_1) = \mathcal{F}$. Then define
$$\mathfrak{A}_1 := \left\{ Z_{1:n} \in \mathcal{I}^n : \sup_{f \in \mathcal{F}(\delta_1)} (P - P_n)[\ell \circ f - \ell \circ f^\star] \le \delta_2 := U_n(\delta_1, t_1) \right\}.$$
Then $\mathbb{P}(\mathfrak{A}_1) \ge 1 - e^{-t_1}$. On this event,
$$\begin{aligned} \mathcal{E}(\hat{f}_n) &\le \sup_{f \in \mathcal{F}(\delta_1)} (P - P_n)[\ell \circ f - \ell \circ f^\star] \qquad \text{(by the basic inequality)} \\ &\le \delta_2 \qquad \text{(by Talagrand's inequality)}. \end{aligned}$$

Intuition: find δ as small as possible

- Assumption 1 states that
$$\sup_{f \in \mathcal{F}} \|\ell \circ f - \ell \circ f^\star\|_\infty \le B.$$
- In the bound from (1), we apply it with $g = \ell \circ f - \ell \circ f^\star$ and $\mathcal{G} = \mathcal{F}(\delta)$.
$\rightarrow$ The resulting bound depends on n, δ, and t. We denote it by
$U_n(\delta, t) := 2\mathbb{E}\Big[\sup_{f \in \mathcal{F}(\delta)} (P - P_n)[f]\Big] + \sqrt{\frac{2t}{n} \sup_{f \in \mathcal{F}(\delta)} \mathrm{Var}_P(f)} + \frac{4Bt}{3n}$.
- First, set $\delta_1 := B$, so that $\mathcal{F}(\delta_1) = \mathcal{F}$. Then define
$$\mathfrak{A}_1 := \left\{ Z_{1:n} \in \mathcal{I}^n : \sup_{f \in \mathcal{F}(\delta_1)} (P - P_n)[\ell \circ f - \ell \circ f^\star] \le \delta_2 := U_n(\delta_1, t_1) \right\}.$$
Then $\mathbb{P}(\mathfrak{A}_1) \ge 1 - e^{-t_1}$. On this event,
$$\begin{aligned} \mathcal{E}(\hat{f}_n) &\le \sup_{f \in \mathcal{F}(\delta_1)} (P - P_n)[\ell \circ f - \ell \circ f^\star] \qquad \text{(by the basic inequality)} \\ &\le \delta_2 \qquad \text{(by Talagrand's inequality)}. \end{aligned}$$
- Since $\mathcal{E}(\hat{f}_n) \le \delta_2$, we have $\hat{f}_n \in \mathcal{F}(\delta_2)$. So we can repeat the same argument and define δ_3 by
$$\mathfrak{A}_2 := \left\{ Z_{1:n} \in \mathcal{I}^n : \sup_{f \in \mathcal{F}(\delta_2)} (P - P_n)[\ell \circ f - \ell \circ f^\star] \le \delta_3 := U_n(\delta_2, t_2) \right\}.$$
Again, $\mathbb{P}(\mathfrak{A}_2) \ge 1 - e^{-t_2}$.

Intuition

- Then,

$$\begin{aligned}\mathbb{P}(\mathcal{E}(\hat{f}_n) > \delta_3) &\leq \mathbb{P}(\{\mathcal{E}(\hat{f}_n) > \delta_3\} \cap \mathfrak{A}_1) + \mathbb{P}(\mathfrak{A}_1^C) \\ &\leq e^{-t_1} + e^{-t_2}.\end{aligned}$$

- Repeating this argument, the sequence δ_j decreases and eventually converges. In particular,

$$\mathcal{E}(\hat{f}_n) \leq \delta_N \quad \text{with probability at least} \quad 1 - \sum_{j=1}^{N} e^{-t_j}.$$

- But is δ_j decreasing?

Intuition

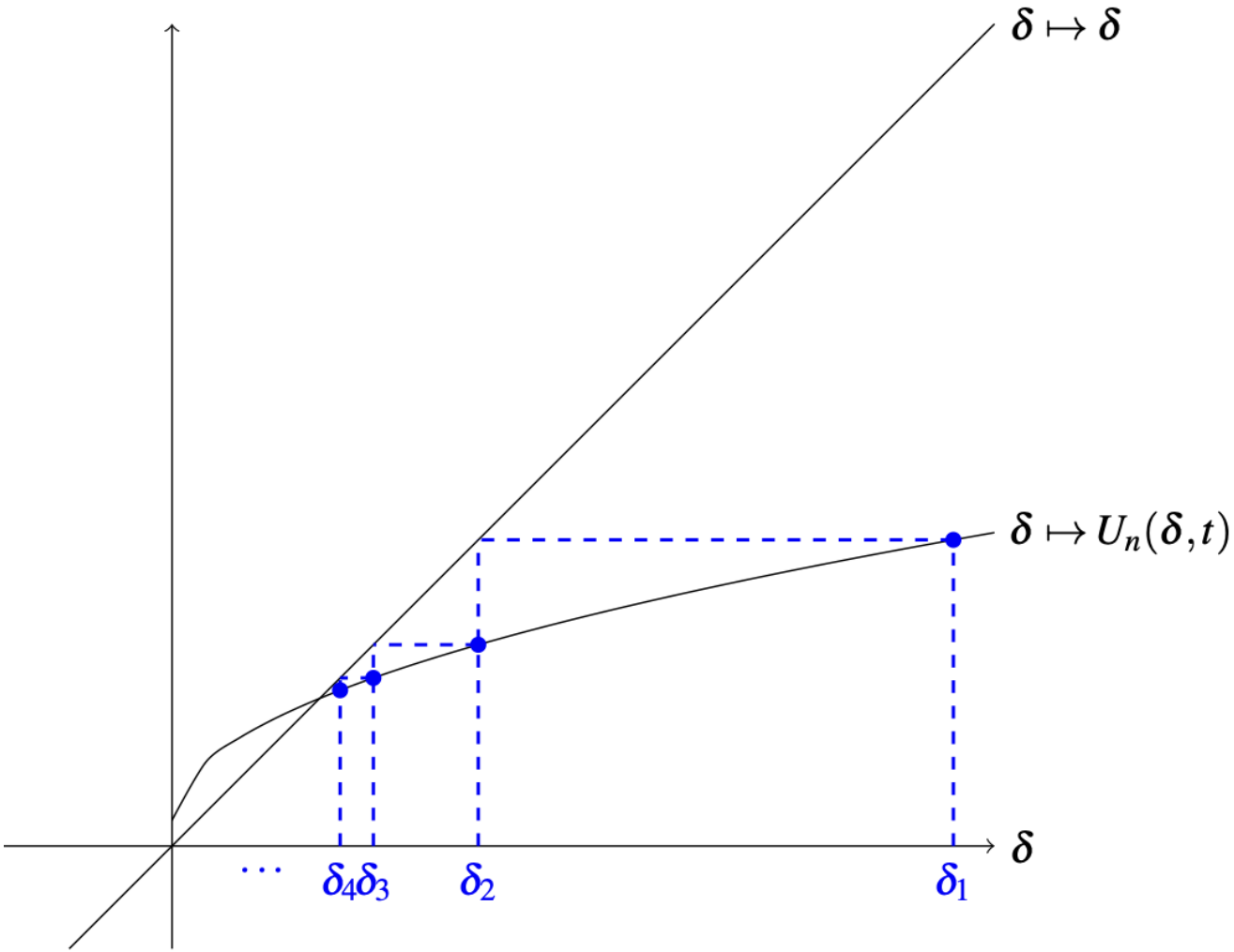

Figure: Illustration of iterative arguments

Fixed-point method: the rate is determined by the fixed point of $\Psi_n(\delta)$

- Instead of iterative arguments, we introduce new method called fixed point method.
- Fixed point method proceeds as follows:
  - We make any decreasing sequence δ_j, and positive number t_j.
  - And find $\Psi_n(\delta)$ such that $\Psi_n(\delta_j) \geq U_n(\delta_j, t_j)$ for all j.
  - Then we can find minimum δ_j called fixed point such that $\delta_n^\dagger := \sup\{\delta : \ \delta \leq \Psi_n(\delta)\}$.
- Thus, excess risk can be controlled by the fixed point $\delta_n^\dagger$.
- When does the fixed point occurs at a smaller δ?

Outline

Finding a good $\Psi_n(\delta)$

- Therefore, we need to find a function $\Psi_n(\delta)$ that upper bounds $U_n(\delta, t)$ and has a shape that makes its fixed point more smaller.
- Desirable shape is a slope less than 1 and a small residual term.
- The sub-root dependence make ensure both.
- For example, $\Psi_n(\delta) = c_1\sqrt{\frac{\delta}{n}} + \frac{1}{n}$

$$\delta \leq \Psi_n(\delta)$$

$$\delta \leq \frac{c_1}{2}\delta + \frac{c_1}{2n} + \frac{1}{n} \quad \text{(By AM-GM, } \sqrt{\frac{\delta}{n}} \leq \frac{1}{2}\delta + \frac{1}{2n}\text{)}$$

$$\delta_n^\dagger \leq \frac{3c_1}{2 - c_1}\frac{1}{n} = O(n^{-1}).$$

Conclusion

- We will get upper bounds of each terms in

$$U_n(\delta, t) = 2\mathbb{E}\Big[\sup_{f\in\mathcal{F}(\delta)} (P - P_n)[\ell \circ f - \ell \circ f^\star]\Big] + \sqrt{\frac{2t}{n} \sup_{f\in\mathcal{F}(\delta)} \mathrm{Var}_P(\ell \circ f - \ell \circ f^\star)} + \frac{4Bt}{3n}$$

- We will get result:

$$\mathbb{E}\Big[\sup_{f\in\mathcal{F}(\delta)} (P - P_n)[\ell \circ f - \ell \circ f^\star]\Big] \le \frac{1}{16}\delta + C_\rho(\phi_{1,n})^{2/(2-\rho)} + \phi_{0,n} \quad \rho \in (0, 1]$$

$$\sqrt{\frac{2t}{n} \sup_{f\in\mathcal{F}(\delta)} \mathrm{Var}_P(\ell \circ f - \ell \circ f^\star)} \le \frac{1}{16}\delta + C_\kappa \left(\frac{Rt}{n}\right)^{1/(2-\kappa)} \quad \kappa \in (0, 1]$$

Main Theorem

$$\mathcal{E}(\hat{f}) \lesssim \max\left\{\bar{\phi}_n(\mathcal{F}), (Rt/n)^{1/(2-\kappa)}, Bt/n\right\},$$

where

$$\bar{\phi}_n(\mathcal{F}) := (\phi_{1,n}(\mathcal{F}))^{2/(2-\rho)} \vee \phi_{0,n}(\mathcal{F}).$$

Thus, we can get tight upper bounds than the global $n^{-1/2}$ rate.

Conditions for fast rates 1: Bernstein

- First, we introduce some useful inequality and definition.

Definition

Let $\kappa \in (0, 1]$ and $R > 0$. We say that $\mathcal{G}$ is a (κ, R)-*Bernstein class* with respect to a probability measure $\mathbf{P}$, if

$$\mathrm{Var}_{\mathbf{P}}[g] \leq R(\mathbf{P}[g])^{\kappa}$$

for every $g \in \mathcal{G}$. We call κ the *Bernstein exponent*.

Young's inequality

Let $p > 1$ and $q > 1$ be conjugate indices such that $1/p + 1/q = 1$. Then

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}$$

for any $a \geq 0$ and $b \geq 0$.

Conditions for fast rates 1: Bernstein

- We now derive an upper bound for the variance term

$$\sqrt{\frac{2t}{n} \sup_{f \in \mathcal{F}(\delta)} \mathrm{Var}_P(\ell \circ f - \ell \circ f_\star)}.$$

Assumption 2

The class $\ell(\mathcal{F}_{-f_\star}) := \{\ell \circ f - \ell \circ f_\star : f \in \mathcal{F}\}$ is a (κ, R)-*Bernstein class* with respect to a probability measure $\mathbb{P}$.
That is, $\mathrm{Var}_{\mathbf{P}}(\ell \circ f - \ell \circ f_\star) \le R\,[\mathbf{P}[\ell \circ f - \ell \circ f^\star]]^\kappa = R(\mathcal{E}(f))^\kappa$ for any $f \in \mathcal{F}$.

- Using Assumption 2 and the Young's inequality, the variance term of $U_n(\delta, t)$ can be bounded as

$$\sqrt{\frac{2t}{n} \sup_{f \in \mathcal{F}(\delta)} \mathrm{Var}_P(\ell \circ f - \ell \circ f_\star)} \le \sqrt{\frac{2Rt\delta^\kappa}{n}} \le \frac{1}{16}\delta + C_\kappa \left(\frac{Rt}{n}\right)^{1/(2-\kappa)},$$

- Therefore, the usual $n^{-1/2}$ rate can be improved to $n^{-1/(2-\kappa)}$, which is significantly faster when $\kappa \in (0, 1]$.

Control expectation of empirical process

- We now derive an upper bound for the expectation term in $U_n(\delta, t)$. We impose the following assumption.

Assumption3: Sub-root complexity

There exists a constant $\rho \in (0, 1]$ and sequences of positive numbers $(\phi_{0,n})_{n\in\mathbb{N}}$ and $(\phi_{1,n})_{n\in\mathbb{N}}$ such that

$$\varphi_n(\delta, \mathcal{F}) := \mathbb{E}\left[\sup_{f\in\mathcal{F}(\delta)} (\mathbf{P} - \mathbf{P}_n)(\ell \circ f - \ell \circ f_\star)\right] \leq \phi_{1,n}\delta^{\rho/2} + \phi_{0,n} \tag{2}$$

for any $\delta > 0$.

- Usually, the sequence $(\phi_{1,n})_{n\in\mathbb{N}}$ is not faster than $n^{-1/2}$.

Control expectation of empirical process

- Applying Young's inequality to (2), we obtain:

$$\phi_{1,n}\delta^{\rho/2} + \phi_{0,n} \leq \frac{1}{16}\delta + C_\rho(\phi_{1,n})^{2/(2-\rho)} + \phi_{0,n}$$

- We define the quantity $\bar{\phi}_n$:

$$\bar{\phi}_n(\mathcal{F}) := (\phi_{1,n})^{2/(2-\rho)}(\mathcal{F}) \vee \phi_{0,n}(\mathcal{F}) \tag{3}$$

- Finally, the excess risk is bounded as follows:

$$\max\left\{\bar{\phi}_n(\mathcal{F}), (Rt/n)^{1/(2-\kappa)}, Bt/n\right\}.$$

- Since $1 < 2/(2-\rho) \leq 2$ for $\rho \in (0, 1]$, the rate $\bar{\phi}_n(\mathcal{F})$ can be faster than $n^{-1/2}$ satisfying

$$\phi_{0,n} = o(n^{-1/2}) \qquad \phi_{1,n} = O(n^{-a})$$

with $a > \frac{2-a}{\rho}$

End