CHAPTER 12 CONSISTENCY OF INDIRECT LEARNING METHODS
Intuition of proof ideas

- The decision rule is not explicit.
- However, we know that best classifiers minimizes some loss function or regularized loss functions.
- Thus,

\[ P(L(g_n) - L(g^*) > \epsilon) \leq P(L(g_n) - L_n(g_n) - L(g^*) + L_n(g^*) > \epsilon) \]

\[ \leq 2P(\sup_{g \in \mathcal{F}} |L_n(g) - L(g)| > \epsilon/2). \]

- We need control stochastic errors of such loss functions over the model space,

\[ \sup_{g \in \mathcal{F}} |L_n(g) - L(g)|. \]

- This uses concentration inequalities from empirical processes and relies on the model size of \( \mathcal{F} \).
Example 1: Maximum likelihood principle

- Let $\mathcal{F}$ be the class of decision functions with values in $[0, 1]$.
- The estimated decision function, $\eta_n(x) \in \mathcal{F}$, maximizes

$$l_n(\eta) = n^{-1} \sum_{i=1}^{n} [Y_i \log \eta(X_i) + (1 - Y_i) \log(1 - \eta(X_i))]$$

so the decision rule is $g_n(x) = I(\eta_n(X) > 1/2)$.
- Consistency result: suppose the true $\eta \in \mathcal{F}$ and $N_{[]} (\epsilon, \mathcal{F}, L_2(P)) \leq (K/\epsilon)^V$. Then $L(g_n) \rightarrow L(g^*)$ with probability one.
Steps of the proof

• Define

\[ \mathcal{F}_\epsilon = \{ \tilde{\eta} : \tilde{\eta} \in \mathcal{F}, L(\tilde{g}) > L(g^*) + \epsilon \} \]

where \( \tilde{g} = I(\tilde{\eta} > 1/2) \).

• Note

\[ P(L(g_n) - L(g^*) > \epsilon) = P(\eta_n \in \mathcal{F}_\epsilon) \leq P(\sup_{\tilde{\eta} \in \mathcal{F}_\epsilon} (l_n(\tilde{\eta}) - l_n(\eta)) > 0). \]

• On the other hand,

\[
E[l_n(\eta)] - E[l_n(\tilde{\eta})] = KL(\eta, \tilde{\eta}) \\
\geq E[\eta(X)(\sqrt{\eta(X)} - \sqrt{\tilde{\eta}(X)})^2] \\
+ (1 - \eta(X))(\sqrt{1 - \eta(X)} - \sqrt{1 - \tilde{\eta}(X)})^2] \\
\geq E[(\eta(X) - \tilde{\eta}(X))^2]/4
\]

and \( \epsilon \leq L(\tilde{g}) - L(g^*) \leq 2E[(\eta(X) - \tilde{\eta}(X))^2]. \)
Completing the proof

- Hence,

\[ P(L(g_n) - L(g^*) > \epsilon) \leq P(\eta_n \in \mathcal{F}_\epsilon) \]

\[ \leq P(\sup_{\tilde{\eta} \in \mathcal{F}_\epsilon} \{(l_n(\tilde{\eta}) - l_n(\eta)) - E[l_n(\tilde{\eta}) - l_n(\eta)]\} > \epsilon/8). \]

- Need to apply the concentration inequality from empirical processes to obtain an upper bound \( n^a e^{-n\epsilon^2} \).
Example 2: Empirical risk minimization

- $g_n$ minimizes the empirical Bayes risk

$$L_n(g) = n^{-1} \sum_{i=1}^{n} I(Y_i \neq g(X_i))$$

for $g \in \mathcal{C}$.

- Strong consistency: if $\mathcal{C}$ has a finite VC dimension $v$, then

$$L(g_n) - \inf_{g \in \mathcal{C}} L(g) \to_{a.s.} 0.$$ 

- Remark: if $g^* \in \mathcal{C}$, then $L(g_n) \to L(g^*)$ with probability one.
Proof

• Note

\[ L(g_n) - \inf_{g \in C} L(g) \leq L(g_n) - L_n(g_n) + \sup_{g \in C} (L_n(g) - L(g)) \]

\[ \leq 2 \sup_{g \in C} |L_n(g) - L(g)|. \]

• Thus,

\[ P(L(g_n) - \inf_{g \in C} L(g) > \epsilon) \]

\[ \leq P(\sup_{g \in C} |L_n(g) - L(g)| > \epsilon/2) \]

\[ \leq 8(\epsilon n/v)^v e^{-n\epsilon^2/128}. \]

• The last step uses the concentration inequality for empirical processes.
Concentration inequalities for VC-class

- For a class $\mathcal{F}$ with finite VC-dimension $v$,

$$P\left(\sup_{f \in \mathcal{F}} |P_n f - P f| > \epsilon\right) \leq 8\left(\frac{en}{v}\right)^v e^{-n\epsilon^2/32}.$$
Example 3: Empirical risk minimization with complexity regularization

- Consider a sequence of models
  \[ C^{(1)} \subset C^{(2)} \subset \ldots \]

- Each \( C^{(k)} \) has a finite VC-dimension \( v_k \).

- Let \( g_n^{(k)} \) be the best decision rule in model \( C^{(k)} \) minimizing \( L_n(g) \) for \( g \in C^{(k)} \).

- The best selected rule \( \tilde{g} \) is one of \( g_n^{(1)}, g_n^{(2)}, \ldots \) which minimizes
  \[ L_n(g_n^{(k)}) + R(k, n), \quad R(k, n) = \sqrt{32v_k \log(en)/n}. \]

- Consistency result: suppose \( \lim_{k} \inf_{g \in C^{(k)}} L(g) = L(g^*) \) and
  \[ \sum_{k=1}^{\infty} e^{-v_k} < \infty, \]
  then \( L(\tilde{g}) \to L(g^*) \) with probability one.
\textbf{Proof}

- Note

\[ L(\tilde{g}) - L(g^*) = L(\tilde{g}) - \inf_{k \geq 1} \left\{ L_n(g_n^{(k)}) + R(k, n) \right\} \]
\[ + \left( \inf_{k \geq 1} \left\{ L_n(g_n^{(k)}) + R(k, n) \right\} - L(g^*) \right). \]
First term in RHS

• For the first term,

$$P(L(\tilde{g}) - \inf_{k \geq 1} \left\{ L_n(g_n^{(k)}) + R(k, n) \right\} > \epsilon)$$

$$= P(L(\tilde{g}) - L_n(\tilde{g}) - R(\tilde{k}, n) > \epsilon)$$

$$\leq \sum_{k \geq 1} P(L(g_n^{(k)}) - L_n(g_n^{(k)}) > \epsilon + R(k, n))$$

$$\leq \sum_{k \geq 1} P(\sup_{g \in C^{(k)}} |L(g) - L_n(g)| > \epsilon + R(k, n))$$

$$\leq \sum_{k \geq 1} 8n v_k e^{-n(\epsilon + R(n,k))^2/32} \leq 8e^{-n\epsilon^2/32} \sum_k e^{-v_k}.$$  

• The last step uses the concentration inequality for empirical processes with VC-classes.
• First, there exists some $m$ such that

$$\inf_{g \in C(m)} L(g) < L(g^*) + \epsilon/4$$

and $R(n, m) < \epsilon/4$.

• Therefore,

$$P(\inf_{k \geq 1} (L_n(g_n^{(k)}) + R(k, n)) - L(g^*) > \epsilon)$$

$$\leq P(L_n(g_n^{(m)}) - L(g_n^{(m)}) > \epsilon/2)$$

$$\leq 8n^{v_m} e^{-n\epsilon^2/128}.$$

• It follows because $v_m < \infty$. 

---

**Second term in RHS**

- First, there exits some $m$ such that

$$\inf_{g \in C(m)} L(g) < L(g^*) + \epsilon/4$$

and $R(n, m) < \epsilon/4$.

- Therefore,

$$P(\inf_{k \geq 1} (L_n(g_n^{(k)}) + R(k, n)) - L(g^*) > \epsilon)$$

$$\leq P(L_n(g_n^{(m)}) - L(g_n^{(m)}) > \epsilon/2)$$

$$\leq 8n^{v_m} e^{-n\epsilon^2/128}.$$

- It follows because $v_m < \infty$. 

---
Remarks

• If $L(g^*) < \lim_{k} \inf_{g \in C^{(k)}} L(g)$, then the above proof implies

$$L(\tilde{g}) \rightarrow_{a.s.} \lim_{k} \inf_{g \in C^{(k)}} L(g)$$

instead of $L(g^*)$. 
CHAPTER 13 CONVERGENCE RATES
Some results

- Goal is to find some equality like

\[ E[L(g_n)] - L(g^*) < en^{-\alpha}. \]

- Unfortunately, this is impossible: let \( a_n \) decrease to zero and \( a_1 < 1/16 \). For every sequence classifier \( \{g_n\} \), there exits some a distribution of \((X, Y)\) such that \( L(g^*) = 0 \) but \( E[L(g_n)] \geq a_n \).

- However, such rate exists if we impose restrictive assumption of \((X, Y)\) distributions (smoothness).

- Many results on obtaining accurate convergence rates rely on using concentration inequalities as seen in consistency proofs.

- Open question: inference on \( E[L(g_n)] \)?
Resubstitution estimate

- Apparent error: \( n^{-1} \sum_{i=1}^{n} I(\hat{g}(X_i) \neq Y_i) \)
- It is biased towards zero.
- Using validation data to obtain
  \[
  m^{-1} \sum_{j=1}^{m} I(\hat{g}(X_{n+j}) \neq Y_{n+j}).
  \]
- Require large validation data.
Leave-one-out cross-validation estimate

- Assume $g_n$ to be a symmetric classifier.
- An upper bound result:
  \[ E \left\{ (\hat{L}_{CV} - L(g_n))^2 \right\} \leq n^{-1} + 6P(g_n(X) \neq g_{n-1}(X)). \]
- For $k$-NN rule,
  \[ P(g_n(X) \neq g_{n-1}(X)) \leq k/n. \]
- For kernel rule with $K(x)$ having support in the unit sphere and $K(x) \geq \beta$ for $\|x\| \leq \rho$,
  \[ P(g_n(X) \neq g_{n-1}(X)) \leq \frac{C_p}{\beta \sqrt{n}} (1 + \rho^{-1})^{p/2}. \]
Other error estimates

• Smooth error count:

\[ n^{-1} \sum_{i=1}^{n} [Y_i(1 - r(\eta_n(X_i))) + (1 - Y_i)r(\eta_n(X_i)))] \]

where \( r \) is an increasing function satisfying

\[ r(1/2 - u) + r(1/2 + u) = 1. \]

• It tends to have a smaller variability than the resubstitution estimate \((r(u) = I(u \geq 1/2))).\)

• Posterior probability estimate:

\[ n^{-1} \sum_{i=1}^{n} [I(\eta_n(X_i) \leq 1/2)\eta_n(X_i) + I(\eta_n(X_i) > 1/2)(1 - \eta_n(X_i))] \]

or its cross-validation version by replacing \( \eta_n(X - i) \) by \( \eta_{n,-i}(X_i). \)
- $K$-fold cross-validation estimate
- Bootstrap estimate: use a random sample to construct the classifier than predict the errors in those not selected; repeat many times then take average.
CHAPTER 15 CONCENTRATION INEQUALITIES
Concentration Inequalities

- What are concentration inequalities?
- They are essentially inequalities for the tail probability of the deviation of a random variable from its mean (or median)

\[
P(\xi(X_1, ..., X_n) - E[\xi(X_1, ..., X_n)] \geq x) \leq \exp\{-x^2/2\nu\}.
\]

- Large deviation or moderate deviation inequalities
Concentration Inequalities

- The oldest one

$$P(f(Z) - E[f(Z)] \geq x) \leq \exp\left\{-\frac{x^2}{2L}\right\},$$

where $Z \sim N(0, I_d)$ and $f$ is Lipschitz continuous with Lipschitz constant $L$. Also, $E[f(Z)]$ can be replaced by the median.
Technique 1 in obtaining concentration inequalities

- Cramer-Chernoff method (Chernoff inequality)

\[ P(Z \geq x) \leq \exp\{-\psi^*_Z(x)\}, \]

where \( \psi^*_Z(x) = \sup_\lambda \{\lambda x - \log E[e^{\lambda Z}]\} \) (Cramer transformation)
Resulting inequalities

- **Hoeffding’s inequality**: $X_1, ..., X_n$ are independent and $X_i \in [a_i, b_i]$,

\[
P\left(\sum_i X_i - \sum_i E[X_i] \geq x\right) \leq \exp\left\{-\frac{2x^2}{\sum_i (b_i - a_i)^2}\right\}.
\]
• **Bennett’s (Bernstein’s) inequality:** $X_1, \ldots, X_n$ are independent and square integrable with $X_i \leq b$ for some constant $b > 0$,

$$P\left( \sum_i X_i - \sum_i E[X_i] \geq x \right) \leq \exp\left\{ -\frac{\nu}{b^2} h\left( \frac{\nu x}{b} \right) \right\}$$

$$\leq \exp\left\{ -\frac{x^2}{2(\nu + bx/3)} \right\},$$

where $\nu = \sum_i E[X_i^2]$ and $h(y) = (1 + y) \log(1 + y) - y$. 
Technique 1 in obtaining concentration inequalities

- Bernstein’s inequality for unbounded variables:
  \( X_1, \ldots, X_n \) are independent and \( \nu = \sum_i E[X_i^2] \). If for
  \( k \geq 3, E[(X_i)_+^k] \leq k!\nu c^{k-2}/2 \), then

\[
P\left( \sum_i X_i - \sum_i E[X_i] \geq \sqrt{2\nu x + cx} \right) \leq \exp\{ -x \}.
\]
Technique 2 in obtaining concentration inequalities

- Information inequalities (Han’s inequality): \( Y = (Y_1, ..., Y_n) \) takes finite values in a product measure space and write 
  \( Y^{(i)} = (Y_1, ..., Y_{i-1}, Y_{i+1}, ..., Y_n) \). Then

\[
h_s(Y) \leq \frac{1}{n-1} \sum_{i=1}^{n} h_s(Y^{(i)})
\]

where \( h_s(Y) = - \sum_y P(Y = y) \log P(Y = y) \) (Shannon entropy).
Resulting inequalities

- **Bounded difference inequality**: If

  \[ |\xi(x_1, ..., x_i, ..., x_n) - \xi(x_1, ..., y_i, ..., x_n)| \leq c_i, \text{ then} \]

  \[ P(\xi(X_1, ..., X_n) - E[\xi(X_1, ..., X_n)] \geq x) \leq \exp\left\{-\frac{2x^2}{\sum_i c_i}\right\}. \]
Other techniques obtaining concentration inequalities

- $\phi$-entropy method
- Isoperimetric method
- Contraction inequality and Rademacher sequence
- etc...
Maximal inequalities

- They are concerned about the suprema of stochastic processes.
- They essentially rely on the previous concentration inequalities.
- Need conditions on the complexity/size of the index (VC-class, bounded entropy with bracket or other).
Maximal inequality (1)

- **Bounded difference process**: If $a_{it} \leq X_i(t) \leq b_{it}$ and let
  \[ L^2 = \sum_i \sup_t (b_{it} - a_{it})^2, \]
  then
  \[
P\left( \sup_{t \in T} \sum_i X_i(t) - E\left[ \sup_{t \in T} \sum_i X_i(t) \right] \geq x \right) \leq \exp\left\{ -\frac{2x^2}{2L} \right\}.
  \]
Maximal inequality (2)

- **Talagrand’s inequality** Let $\mathcal{F}$ be some countable family of measurable functions such that $\|f\|_{\infty} \leq b < \infty$. Let

$$Z = \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{n} (f(X_i) - E[f(X_i)]) \right| \text{ and } \sigma^2 = \sup_{f \in \mathcal{F}} \left[ \sum_{i=1}^{n} \text{Var}(f(X_i)) \right].$$

Then for any $\epsilon, x > 0$,

$$P(Z \geq (1 + \epsilon)E[Z] + \sigma \sqrt{2\nu x} + \nu(\epsilon)bx) \leq e^{-x}$$

with $\nu = 4$ and $\nu(\epsilon) = 2.5 + 32\epsilon^{-1}$ and

$$P(Z \geq (1 - \epsilon)E[Z] - \sigma \sqrt{2\nu'x} - \nu'(\epsilon)bx) \leq e^{-x}$$

with $\nu' = 5.4$ and $\nu'(\epsilon) = 2.5 + 43.2\epsilon^{-1}$. 
**Maximal inequality (3)**

- Let $\psi$ be convex, nondecreasing, nonzero and $\psi(0) = 0$ and

  $$\limsup_n \psi(x)\psi(y)/\psi(cxy) < \infty.$$  

  Then

  $$\left\| \max_{1 \leq i \leq m} X_i \right\|_{\psi} \leq K\psi^{-1}(m) \max_i \|X_i\|_{\psi}.$$

- For the same $\psi$ and if

  $$\|X(s) - X(t)\|_{\psi} \leq Cd(s, t),$$

  then

  $$\left\| \sup_{d(s,t) \leq \delta} |X(s) - X(t)| \right\|_{\psi} \leq K \left[ \int_0^\eta \psi^{-1}(D(\epsilon, d)) d\epsilon 
  + \delta \psi^{-1}(D^2(\eta, d)) \right].$$
Maximal inequality (3)

- A special case: if \( X(t) \) is sub-Gaussian with
  
  \[
P(|X(s) - X(t)| > x) \leq 2e^{-x^2/2d^2(s,t)},
  \]

  then

  \[
  E[\sup_{d(s,t) \leq \delta} |X(s) - X(t)|] \leq K \int_{0}^{\delta} \sqrt{\log D(\epsilon, d)} d\epsilon.
  \]
Maximal inequality (4)

- **Tail bounds for empirical processes:** If $\mathcal{F}$ is bounded by 1 and it satisfies

$$\sup_{Q} N(\epsilon, \mathcal{C}, L_2(Q)) \leq (K/\epsilon)^V$$

or

$$N(\epsilon, \mathcal{F}, L_2(P)) \leq (K/\epsilon)^V,$$

then for any $1 \geq \sigma^2 \geq \sup_{f \in \mathcal{F}} \text{Var}(f(X))$ and $\delta > 0$,

$$P^*(\|G_n\|_{\mathcal{F}} > t) \leq C\left(\frac{1}{\sigma}\right)^{2V}(1 \vee \frac{t}{\sigma})^{3V} + \delta e^{-\frac{1}{2} \frac{t^2}{\sigma^2 + (3+t)/\sqrt{n}}}.$$
Summary

• Concentration inequalities (and maximal inequalities) play essential role in the theory for statistical learning and high-dimensional problem.

• So far, they are only used in proving “consistency” in these problems.

• They or their further development can be potentially useful for “inference” (convergence rates or even asymptotic distributions).

• In conclusion, unless you hate SLHD theory, you should know concentration inequalities.