44.006 — University of Seoul, Spring 2026

Lecture 5: Empirical Process: Applications to Asymptotic Analysis

June, 1, 2026

Lecturer: *Jong-June Jeon* *Scribe*: *Jong-June Jeon*

1 Persistent Estimator

관측치 $Z^i = (Y^i, X_1^i, \ldots, X_m^i) \sim$ i.i.d. F_n $(i = 1, \ldots, n)$ 로 주어진다. 예측 모형은 $\beta \in B^n \subset \mathbb{R}^m$로 매개변수화된 함수족 $\{g_\beta\}$으로 선형 모형의 경우 $g_\beta(X_1, \ldots, X_m) = \sum_{j=1}^m \beta_j X_j$로 주어진다. 손실함수는 $l(Y, g_\beta(X)) \equiv l(\beta, Z)$로 표시한다. 제곱 손실함수는 $l(\beta, Z) = (Y - \sum \beta_j X_j)^2$가 된다.
변수의 개수가 표본 수보다 많은 고차원 환경($m \gg n$, 예: $m = n^\alpha, \alpha > 1$)을 가정한다. 이를 분석하기 위해 표본 크기 n이 증가함에 따라 파라미터 차원 m도 증가하는 삼각형 배열(Triangular array) 구조를 사용한다.

- $L_{F_n}(\beta) = \mathbb{E}_{F_n} l(\beta, Z)$: population risk function
- $L_{\hat{F}_n}(\beta) = \frac{1}{n}\sum_{i=1}^n l(\beta, Z^i)$:
- $\beta^*_{F_n} = \arg\min_{\beta \in B^n} L_{F_n}(\beta)$: optimal predictor under the true risk

Definition 1.1 (Persistent Estimator). 어떤 estimator $\hat{\beta}_n \in B_n$의 열이 모든 $\epsilon > 0$에 대해 다음을 만족하면 $\{\hat{\beta}_n\}$이 B^n에 대해 persistent하다고 정의함.

$$\sup_{F_n \in \mathcal{F}_n} \mathbb{P}_{F_n}(L_{F_n}(\hat{\beta}_n) - L_{F_n}(\beta^*_{F_n}) > \epsilon) \to 0 \quad (\text{as } n \to \infty)$$

Assumption 모든 ε에 대하여, 충분히 큰 n일 때 만약 $L_{F_n}(\beta) > L_{F_n}(\beta^*_{F_n}) + 2\varepsilon$ 이라면, $T_\varepsilon \equiv \min(l(\beta, Z), M(\varepsilon))$이 다음을 만족하도록 하는 $M(\varepsilon)$이 존재한다:

$$E_{F_n} T_\varepsilon > L_{F_n}(\beta^*_{F_n}) + \varepsilon$$

- $L_{F_n}(\beta) > L_{F_n}(\beta^*_{F_n}) + 2\varepsilon$ 이라는 것은 '해당 β 값이 true risk function의 optimal 값보다 조금 큰 경우'를 의미
- $E_{F_n} T_\varepsilon > L_{F_n}(\beta^*_{F_n}) + \varepsilon$ 이라는 것은 그러한 β에 대해서 $L_{F_n}(\beta^*_{F_n})$값이 임의로 커질 일이 없다는 뜻.

- 다음을 만족하는 유한개의 파라미터 집합 $\beta^1, \ldots, \beta^N$의 log cardinality를 ϵ-entropy 라 부른다.

$$\forall \beta \in B,\ \exists \beta^j \in \{\beta^1, \ldots, \beta^N\} \quad \text{s.t.} \quad \sup_{F \in \mathcal{F}} |L_F(\beta^j) - L_F(\beta)| < \varepsilon$$

- $A(k_n)$은 희소성 조건 (Sparsity Condition)을 만족하는 집합. 파라미터 벡터 β의 성분 중 0이 아닌(비제약) 성분의 개수가 최대 k_n개 이하임.

$$\|\beta\|_0 \leq k_n$$

Theorem 1.2. 조건 *1*을 만족하는 삼각형 배열(*triangular array*)이 주어졌을 때, 다음을 가정하자:

- 모든 분포 열 F_n ($n = 1, 2, \ldots$)에 대하여, $\beta^*_{F_n}$은 원점(*origin*)을 중심으로 하고 유클리드 부피가 R_n인 $k_n = o(n/\log(n))$ 차원의 큐브(*cube*)에 속한다. 여기서 $\log(R_n) = o(n)$이다.
- 리스크 함수 $L_{F_n}(\beta)$는 다음의 *Lipschitz condition*을 만족한다: 임의의 $\varepsilon > 0$에 대하여, $\delta > 0$와 $\gamma > 0$가 존재하여, 만약 $\beta, \beta' \in A(k_n)$이고 $\|\beta - \beta'\|_2 < \delta n^{-\gamma}$이면, $F_n \in \mathcal{F}^n$에 대해 균등하게 (*uniformly*) 다음이 성립한다.

$$|L_{F_n}(\beta) - L_{F_n}(\beta')| < \varepsilon$$

이때, 모든 $\varepsilon > 0$에 대하여 *persistent estimator* 가 존재한다.

(증명) 이론적 최적 파라미터 $\beta^*_{F_n}$를 유한 격자점 공간 위로 투영(Projection)시킨 근사점을 β^o 라고 하자. $\hat{\beta}_n$은 유한 격자점 집합 C_n($\frac{\varepsilon}{2}$-grid) 위에서 경험적 리스크(Empirical Risk)를 최소화하는 해라고 하자.

- $L_{F_n}(\hat{\beta}_n) - L_{F_n}(\beta^*_{F_n}) = \Big(L_{F_n}(\hat{\beta}_n) - L_{\hat{F}_n}(\hat{\beta}_n)\Big) + \Big(L_{\hat{F}_n}(\hat{\beta}_n) - L_{F_n}(\beta^*_{F_n})\Big)$: 격자 위에서 수렴 (개수를 나중에 세기)
- $L_{\hat{F}_n}(\hat{\beta}_n) - L_{F_n}(\beta^*_{F_n}) \leq L_{\hat{F}_n}(\beta^\circ) - L_{F_n}(\beta^*_{F_n})$: empirical minimizer
- $L_{\hat{F}_n}(\beta^\circ) - L_{F_n}(\beta^*_{F_n}) = \Big(L_{\hat{F}_n}(\beta^\circ) - L_{F_n}(\beta^\circ)\Big) + \Big(L_{F_n}(\beta^\circ) - L_{F_n}(\beta^*_{F_n})\Big)$: 격자 위에서 수렴 (개수를 나중에 세기) + libshitz conti.

$L_{F_n}(\hat{\beta}_n) - L_{F_n}(\beta^*_{F_n}) \leq \frac{3}{2}\varepsilon$를 보일 수 있다. 격자위에서 수렴을 exponental convergence 를 사용할 수 있다면 결국 uniformly convergence 를 보일 때 필요한 것이 격자의 개수에 해당한다.
크기가 k_n인 좌표(변수)의 서브셋 가짓수는 m^{k_n}개보다 적다. 해당 k_n-차원 큐브를 꼭짓점의 길이가 $\delta/\sqrt{k_n}n^\gamma$인 서로소인 작은 큐브들로 분할한다. 부피가 R_n인 큐브 안의 각 점은 작은 큐브 중 하나의 중심으로부터 유클리드 거리 기준 $\delta n^{-\gamma}$ 이내에 존재하게 되며, 이는 특히 최적점인 $\beta^*_{F_n}$에 대해서도 성립한다.

정의된 ε-grid의 원소 수(Cardinality)는 다음과 같음:

$$R_n/[\delta/\sqrt{k_n}n^{\gamma}]^{k_n} = \exp\left(\log(R_n) + [\log(\tfrac{1}{\delta}) + \log(\sqrt{k_n}) + \gamma\log(n)]k_n\right) \equiv A_n$$

이러한 서브셋의 개수는 $B_n \equiv m^{k_n} = \exp(\alpha\log(n)k_n)$보다 적음. 최대 k_n개의 비제약(nonzero) 좌표를 가지며 가정 (i)의 큐브에 속하는 모든 예측기 집합에 대해 ε-grid를 구축하는 데 필요한 총 점의 개수 N은 다음과 같이 유계됨:

$$N = A_n \times B_n$$

N의 증가속도가 다항항수의 속도이기 때문에 수렴.

2 High-Dimensional Generalized Linear Models and the Lasso

2.1 Hoeffding's lemma

$X \in [a,b]$이고 $E(X) = 0$ 인 확률변수를 생각한다 (단, $a < 0 < b$). 임의의 s에 대해

$$E[e^{sX}] \leq \exp\left(\frac{s^2(b-a)^2}{8}\right)$$

(증명) 함수 $f(x) = e^{sx}$는 임의의 s에 대해 볼록 함수이다. $X \in [a,b]$이므로, X는 다음과 같이 두 끝점 a, b의 볼록 결합(Convex combination)으로 표현할 수 있다. X를 a와 b의 $(1-t) : t$의 내분점 ($t \in [0,1]$) $X = tb + (1-t)a$ 로 놓자. 여기서 이 등식을 t 에 대해 표현하면 $t = \frac{X-a}{b-a}$가 된다.
$f(x)$가 볼록 함수이므로 $f(X) = f(tb + (1-t)a) \leq tf(b) + (1-t)f(a)$ 가 성립한다. 이를 다시 쓰면

$$\underbrace{e^{sX}}_{f(X)} \leq \underbrace{te^{sb}}_{tf(b)} + \underbrace{(1-t)e^{sa}}_{(1-t)f(a)} = \frac{X-a}{b-a}e^{sb} + \frac{b-X}{b-a}e^{sa}$$

양변에 X에 대해 기댓값을 취하면 $E[X] = 0$를 이용하여 다음과 같이 정리된다.

$$E[e^{sX}] \leq \frac{-a}{b-a}e^{sb} + \frac{b}{b-a}e^{sa}$$

식의 단순화를 위해 $u = s(b-a)$ 와 $\theta = \frac{-a}{b-a}$ (이때 $a \leq 0 \leq b$ 이므로 $0 \leq \theta \leq 1$)를 놓는다. 이를 부등식 좌변에 대입하여 u에 관한 함수 $\phi(u)$로 좌변을 다시 쓴다.

$$\phi(u) = \theta e^{(1-\theta)u} + (1-\theta)e^{-\theta u}$$

$\psi(u) = \ln \phi(u)$에 대한 상한을 구해보면 $u = 0$ 근방에서 테일러 전개를 통해서

$$\psi(u) = \frac{u^2}{8}$$

를 얻을 수 있다. 즉

$$\phi(u) \leq \exp\left(\frac{u^2}{8}\right)$$

원래의 변수 $u = s(b-a)$를 다시 대입하면 최종 상계가 유도된다.

$$E[e^{sX}] \leq \exp\left(\frac{s^2(b-a)^2}{8}\right)$$

Theorem 2.1. 독립인 확률변수 $X_1, X_2, \ldots, X_n$이 각각 $X_i \in [a_i, b_i]$를 만족하고 $\mathbf{E}(X_i) = 0$ 이다.이 때 임의의 양수 $t > 0$에 대하여 다음 부등식이 성립한다.

$$P\left(\bar{X} \geq t\right) \leq \exp\left(\frac{-2n^2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

(증명) 양수 $s > 0$를 놓으면

$$P\left(\bar{X} \geq t\right) = P\left(\sum_{i=1}^{n} X_i \geq nt\right) = P\left(e^{s\sum_{i=1}^{n} X_i} \geq e^{snt}\right)$$

이 성립한다. 이 식에 마르코프 부등식을 적용하면 다음이 성립한다.

$$P\left(e^{s\sum_{i=1}^{n} X_i} \geq e^{snt}\right) \leq e^{-snt} E\left[e^{s\sum_{i=1}^{n} X_i}\right]$$

$X_1, \ldots, X_n$이 독립이므로

$$E\left[e^{s\sum_{i=1}^{n} X_i}\right] = E\left[\prod_{i=1}^{n} e^{sX_i}\right] = \prod_{i=1}^{n} E\left[e^{sX_i}\right]$$

여기에 각 $X_i \in [a_i, b_i]$에 대한 *Hoeffding's lemma* $E[e^{sX_i}] \leq \exp\left(\frac{s^2(b_i-a_i)^2}{8}\right)$를 적용한다.

$$\prod_{i=1}^{n} E\left[e^{sX_i}\right] \leq \prod_{i=1}^{n} \exp\left(\frac{s^2(b_i - a_i)^2}{8}\right) = \exp\left(\frac{s^2\sum_{i=1}^{n}(b_i - a_i)^2}{8}\right)$$

임의의 $s > 0$에 대해 성립하는 상한 선을 얻는다.

$$P\left(\bar{X} \geq t\right) \leq \exp\left(-snt + \frac{s^2 \sum_{i=1}^{n}(b_i - a_i)^2}{8}\right)$$

좌변의 하한을 구한다. 이차방정식을 이용하면 되므로

$$P\left(\bar{X} \geq t\right) \leq \exp\left(\frac{-2n^2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

를 얻는다.

Lemma 2.2. 독립인 확률변수 $X_1, X_2, \ldots, X_n$이 각각 $X_i \in [a_i, b_i]$를 만족하고 $E[X_i] = \mu_i$이다. 표본평균을 $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, 기댓값의 평균을 $\bar{\mu} = \frac{1}{n}\sum_{i=1}^{n} \mu_i$라 할 때, 임의의 양수 $t > 0$에 대하여 다음 부등식이 성립한다.

$$P\left(\bar{X} - \bar{\mu} \geq t\right) \leq \exp\left(\frac{-2n^2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

(증명) 새로운 확률변수 $Y_i = X_i - \mu_i$를 정의한다. 기댓값의 선형성에 의해 $E[Y_i] = 0$이며, 구간의 길이는 $(b_i - \mu_i) - (a_i - \mu_i) = b_i - a_i$로 보존된다.$Y_i$의 표본평균은 $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \bar{X} - \bar{\mu}$이므로, 구하고자 하는 확률 식은 다음과 같이 치환된다.

$$P\left(\bar{X} - \bar{\mu} \geq t\right) = P\left(\bar{Y} \geq t\right)$$

임의의 양수 $s > 0$에 대하여 다음이 성립한다.

$$P\left(\bar{Y} \geq t\right) = P\left(\sum_{i=1}^{n} Y_i \geq nt\right) = P\left(e^{s\sum_{i=1}^{n} Y_i} \geq e^{snt}\right)$$

마르코프 부등식을 적용하면 다음 상계를 얻는다.

$$P\left(e^{s\sum_{i=1}^{n} Y_i} \geq e^{snt}\right) \leq e^{-snt} E\left[e^{s\sum_{i=1}^{n} Y_i}\right]$$

$Y_1, \ldots, Y_n$의 독립성에 의해 기댓값이 분리된다.

$$E\left[e^{s\sum_{i=1}^{n} Y_i}\right] = \prod_{i=1}^{n} E\left[e^{sY_i}\right]$$

$E[Y_i] = 0$이므로 각 항에 *Hoeffding's lemma*를 적용한다.

$$\prod_{i=1}^{n} E\left[e^{sY_i}\right] \leq \prod_{i=1}^{n} \exp\left(\frac{s^2(b_i - a_i)^2}{8}\right) = \exp\left(\frac{s^2 \sum_{i=1}^{n}(b_i - a_i)^2}{8}\right)$$

이를 결합하면 임의의 $s > 0$에 대해 다음 부등식이 유도된다.

$$P\left(\bar{Y} \geq t\right) \leq \exp\left(-snt + \frac{s^2 \sum_{i=1}^{n}(b_i - a_i)^2}{8}\right)$$

우변의 지수 이차식을 최소화하는 최적의 해 $s^* = \frac{4nt}{\sum_{i=1}^{n}(b_i - a_i)^2}$를 대입하면 다음 상한을 얻는다.

$$P\left(\bar{Y} \geq t\right) \leq \exp\left(\frac{-2n^2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

$\bar{Y} = \bar{X} - \bar{\mu}$를 원식에 대입하면 증명이 완료된다.

$$P\left(\bar{X} - \bar{\mu} \geq t\right) \leq \exp\left(\frac{-2n^2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

2.2 Concentration inequality

통계적 학습 이론에서는 *empirical quantity*와 *population quantity*의 차이를 제어해야 한다. 하나의 함수 f에 대해서는 다음 항을 고려한다.

$$\frac{1}{n}\sum_{i=1}^{n} f(Z_i) - Ef(Z).$$

그러나 추정량이 데이터에 의해 선택되는 경우에는 하나의 고정된 함수가 아니라 함수족 전체에 대한 *uniform bound*가 필요하다.

$$\sup_{f \in \mathcal{F}} |(P_n - P)f|.$$

따라서 단일 함수에 대한 *concentration*과 경험과정 *supremum*에 대한 *concentration*은 구분되어야 한다.

*Hoeffding's lemma*는 *bounded random variable*의 *moment generating function*을 제어하는 결과이다. 이를 독립합에 적용하면 고정된 유계 함수 (f)에 대해

$$P\left(|(P_n - P)f| > t\right) \leq 2\exp(-Cnt^2)$$

형태의 *tail bound*를 얻는다. 따라서 *Hoeffding's lemma*는 기본적으로 $(P_n - P)f$ 와 같이 고정된 하나의 함수에 대한 *empirical average*를 제어하는 도구이다.
다음과 같은 경험과정 $\sup_{f\in\mathcal{F}} |(P_n - P)f|$. 을 제어할때 *Hoeffding's inequality* 는 $\log N(\epsilon, \mathcal{F}, d)$ 가 복잡도 항으로 사용되어

$$P\left(\sup_{f\in\mathcal{F}} |(P_n - P)f| > t\right) \lesssim N(\epsilon, \mathcal{F}, d) \exp(-Cnt^2)$$

와 같은 형태를 가지게 된다.
반면 *Bousquet inequality*는 경험과정 *supremum*

$$Z = \sup_{f\in\mathcal{F}} |(P_n - P)f|$$

자체를 하나의 확률변수로 보고, Z가 그 평균 EZ 주변에 집중한다는 형태의 부등식을 제공한다. 따라서 이 접근에서는 먼저 EZ를 *symmetrization*, *contraction*, *Rademacher complexity* 등을 이용해 상계를 구하고, 그 이후 $Z - EZ$의 *deviation*을 *concentration inequality*로 제어한다.
경험과정 *supremum* 자체를 하나의 확률변수로 놓고 *deviation*을 통제하는 방법을 생각해보자.

$$\sup_{\gamma\in\Gamma} \left| \frac{1}{n}\sum_{i=1}^{n} \gamma(Z_i) - E\gamma(Z_i) \right|.$$

각 $\gamma \in \Gamma$에 대해 $|\gamma|_\infty \leq \eta_n$ 이고

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{var}(\gamma(Z_i)) \leq \tau_n^2$$

라고 하자. 그러면 *Bousquet inequality*는 다음 형태의 *bound*를 제공한다.

$$Z \leq EZ + z\sqrt{2(\tau_n^2 + 2\eta_n EZ)} + \frac{2z^2\eta_n}{3}$$

with high probability. 즉, *Bousquet inequality*는 경험과정 *supremum* Z가 평균 EZ 주변에서 얼마나 벗어나는지를 제어한다. 특별히 이러한 방식을 EZ가 *0*으로 수렴하는 속도를 제어할 수 있을 때 유용한다.
*Bousquet inequality*의 핵심은 다음과 같은 형태의 확률상계가 주어진다는 점이다.

$$Z \lesssim EZ + \sqrt{vt} + bt$$

with high probability. 여기서 v는 분산에 해당하는 양이고, b는 함수족의 *envelope* 또는 *uniform bound*에 해당하는 상수이다. 위 식에 대응할때는 $z^2 = t$ 라고 보면 된다.

2.3 Submodel Class 에 Concentration inequality의 적용

이때 $E(Z)$는 경험과정 *supremum*의 평균적 크기를 나타낸다. 즉, $Z = \sup_{f\in\mathcal{F}} |(P_n - P)f|$ 에 대해

$$E\left[\sup_{f\in\mathcal{F}} |(P_n - P)f|\right]$$

는 함수족 $\mathcal{F}$가 표본 위에서 평균적으로 얼마나 크게 *fluctuation*할 수 있는지를 측정한다. 고정된 하나의 함수 f에 대해서는 $E(P_n - P)f = 0$ 이지만, *supremum*을 취하면 일반적으로 $E\left[\sup_{f\in\mathcal{F}} |(P_n - P)f|\right] > 0$ 이다. 이는 표본에 따라 가장 크게 흔들리는 함수가 선택되기 때문이다. 따라서 경험과정 *supremum* (Z)의 *concentration*은 *0* 주변이 아니라 ($E(Z)$) 주변에서 이루어진다. 또한 ($E(Z)$)는 표본크기 (n)에 의존한다. 일반적으로 함수족이 충분히 작거나 복잡도가 통제되면 ($E(Z)$)는 (n)이 증가함에 따라 감소한다. 이 논문에서는 특정한 조건 (제한된 함수공간)하에서 *symmetrization*과 *contraction*을 통해 $E(Z(M)) \lesssim M\sqrt{\frac{\log m}{n}}$ 과 같은 상계를 얻을 수 있다. 구체적으로 살펴보자.
국소 함수족을 $\{f_\theta : \theta \in \Theta,\ I(\theta - \theta^*) \le M\}$ 라고 하자. 관심 대상은 *fluctuation* 인 다음 *empirical process* 의 상계다.

$$Z(M) = \sup_{f_\theta\in\mathcal{F}_M} |(P_n - P)(\gamma_{f_\theta} - \gamma_{f_{\theta^*}})|\,.$$

다음 *symmetrization lemma*와 간단한 부등식 계산을 통해

$$EZ(M) \le 4ME\left[\max_{1\le k\le m}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\frac{\psi_k(X_i)}{\sigma_k}\right|\right].$$

여기에 기대값의 *maximal inequality* 를 적용한다. (*Lemma A.1*)

$$E\left[\max_{1\le k\le m}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\frac{\psi_k(X_i)}{\sigma_k}\right|\right] \le \sqrt{\frac{2\log(2m)}{n}} + \frac{K_m\log(2m)}{n}.$$

$\bar{a}_n$ 으로 상계를 표시하여

$$\mathbf{E}Z(M) \le a_n M$$

을 얻는다. 여기서 $\bar{a}_n$ 은 n이 증가하면서 *0*으로 수렴해야하는 조건을 만족해야할 것이다.

2.4 논문에서 활용

*Bousquet inequality*에 조건들을 대입하고 EZ의 상한을 대입하면

$$Z(M) \leq \bar{a}_n M + z\sqrt{2(M^2 + 2MK_m\bar{a}_n M)} + \frac{2z^2 MK_m}{3}$$

가 높은 확률로 성립한다. 기호들을 도입하여 루트 안을 정리여 아래와 같은 꼴을 만든다.

$$Z(M) \leq \bar{a}_n M \left[1 + t\sqrt{2(1 + 2\bar{a}_n K_m)} + \frac{2t^2\bar{a}_n K_m}{3}\right].$$

따라서

$$\bar{\lambda}_{n,0}(t) := \bar{a}_n \left[1 + t\sqrt{2(1 + 2\bar{a}_n K_m)} + \frac{2t^2\bar{a}_n K_m}{3}\right]$$

로 정의하면,

$$Z(M) \leq \bar{\lambda}_{n,0}(t)M$$

을 얻는다. *Bousquet inequality*에서 $z = \bar{a}_n t$를 대입하여

$$Z(M) \leq \bar{\lambda}_{n,0}(t)M$$

을 얻는다. 이때 확률은 $1 - \exp(-n\bar{a}_n^2 t^2)$ 이다.