44.006 — University of Seoul, Spring 2026

Lecture 3: Basic of M-Estimator

March, 2026

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

1 M-estimator와 경험적 확률과정

관측값 $X_1, \ldots, X_n$은 어떤 미지의 분포 P로부터 i.i.d.로 생성되었다고 가정한다. 이에 대응되는 경험적 분포(empirical measure)를

$$P_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$$

로 정의한다. 이 강의노트 전반에서 $Pf = \mathbb{E}_P[f(X)]$, $P_n f = \frac{1}{n}\sum_{i=1}^{n} f(X_i)$와 같은 표기를 사용한다.

Example 1.1. (*Maximum Likelihood Estimator*)
모수공간 Θ 위에서 정의된 확률밀도(또는 확률질량함수) p_θ, $\theta \in \Theta$를 고려하자. *MLE*는 주어진 데이터 하에서 로그우도를 최대화하는 모수로 정의된다:

$$\hat{\theta} = \arg\max_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \log p_\theta(X_i).$$

여기서 $m_\theta(x) := \log p_\theta(x)$ 라고 두면, *MLE*는 다음과 같이 경험적 평균을 최대화하는 문제로 표현된다:

$$\hat{\theta} = \arg\max_{\theta\in\Theta} P_n m_\theta.$$

한편, 이상적인(*population*) 관점에서는 $\theta_0 := \arg\max_{\theta\in\Theta} P m_\theta$ 를 정의할 수 있으며, 이는 흔히 "진짜 모수(*true parameter*)"로 해석된다.

Example 1.2. (*Least Squares Estimator*)
다음으로 회귀 문제를 고려하자. 관측값 (X_i, Y_i)가 주어졌을 때, 조건부 평균이 어떤 함수족 $\{f_\theta : \theta \in \Theta\}$로 근사된다고 가정한다.
*LSE*는 제곱오차의 경험적 평균을 최소화하는 모수로 정의된다:

$$\hat{\theta} = \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \big(Y_i - f_\theta(X_i)\big)^2.$$

이를 최대화 문제로 바꾸기 위해 $m_\theta(x,y) := -\big(y - f_\theta(x)\big)^2$ 라고 정의하면,

$$\hat{\theta} = \arg\max_{\theta\in\Theta} P_n m_\theta$$

로 다시 쓸 수 있다. 마찬가지로 *population* 기준의 최적 모수는 $\theta_0 := \arg\max_{\theta\in\Theta} Pm_\theta$ 로 정의된다. MLE와 LSE는 겉보기에는 서로 다른 추정 방법처럼 보이지만, 공통적으로 다음과 같은 형태를 갖는다:

$$\hat{\theta} = \arg\max_{\theta\in\Theta} P_n m_\theta, \qquad \theta_0 = \arg\max_{\theta\in\Theta} Pm_\theta.$$

이와 같이 경험적 평균 P_n을 기반으로 정의된 최대화 문제로부터 얻어지는 추정량을 *M-estimator*라고 부른다. 핵심적인 질문은 다음과 같다:

- $P_n m_\theta$는 Pm_θ를 얼마나 잘 근사하는가?
- 이 근사가 $\hat{\theta}$의 수렴성과 수렴률에 어떤 영향을 미치는가?

이 질문들에 답하기 위해 경험적 확률과정 이론이 필요하게 된다.

2 Excess Risk Analysis

확률공간 $(\mathcal{X}, \mathcal{A}, P)$ 위에서 관측치 $X_1, \ldots, X_n \overset{i.i.d.}{\sim} P$가 주어졌다고 하자. 각 $\theta \in \Theta \subset \mathbb{R}^d$에 대해 $m_\theta : \mathcal{X} \to \mathbb{R}$를 정의하고, 경험적 목적함수와 모집단 목적함수를 각각

$$M_n(\theta) := P_n m_\theta = \frac{1}{n}\sum_{i=1}^{n} m_\theta(X_i), \qquad M(\theta) := Pm_\theta$$

로 둔다. $M(\theta)$는 유일한 최대점 $\theta_0 := \arg\max_{\theta\in\Theta} M(\theta)$ 을 가진다고 가정한다. 경험적 M-estimator는 $\hat{\theta} \in \arg\max_{\theta\in\Theta} M_n(\theta)$ 로 정의된다. 이 때

$$\mathcal{R}(\theta) := M(\theta_0) - M(\theta) = P(m_{\theta_0} - m_\theta)$$

를 θ에서의 *excess risk*라 부른다. 즉, excess risk는 추정값 θ가 최적점 θ_0에서 벗어날 때 모집단 목적함수가 감소하는 크기를 정량화한 값이다.

2.1 ULLN

$\hat{\theta}$이 $M_n(\theta)$의 최대점이므로 항상

$$0 \le M_n(\hat{\theta}) - M_n(\theta_0).$$

양변에 $M(\hat{\theta}) - M(\theta_0)$를 더했다 빼면

$$\begin{aligned} M(\hat{\theta}) - M(\theta_0) &\leq \big(M(\hat{\theta}) - M_n(\hat{\theta})\big) + \big(M_n(\theta_0) - M(\theta_0)\big) \\ &\leq 2 \sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)|. \end{aligned}$$

왼쪽 $M(\theta_0) - M(\hat{\theta})$는 (최대화 문제에서의) **excess risk**이다. 즉,

$$\text{excess risk} \ \leq \ 2 \times (\text{ULLN 오차}).$$

따라서 ULLN이 성립하면 excess risk가 0으로 가고, 식별성 (특히 $M(\theta_0) > M(\theta)$ 과 분리특성)이 있으면 $\hat{\theta}$이 θ_0로 수렴한다.

2.2 Core decomposition

임의의 $\theta \in \Theta$에 대해 다음의 항등식이 성립한다:

$$\begin{aligned} M_n(\theta) - M_n(\theta_0) &= P_n(m_\theta - m_{\theta_0}) \\ &= (P_n - P)(m_\theta - m_{\theta_0}) + P(m_\theta - m_{\theta_0}). \end{aligned}$$

이를 'core decomposition'이라 부른다. 이 분해식에서 $(P_n - P)(m_\theta - m_{\theta_0})$ 는 표본으로 인한 확률적 요동(*stochastic fluctuation*)을 나타내며, $P(m_\theta - m_{\theta_0})$ 는 모집단 목적함수의 결정론적 감소(*deterministic drift*)를 나타낸다.

먼저 $(P_n - P)(m_\theta - m_{\theta_0})$ 표본 평균과 그 기대값의 차이로 이루어진 항으로, 고정된 θ에 대해서는 중심극한정리(CLT)에 의해 고정된 θ에 대해 일반적으로 $O_p(n^{-1/2})$의 크기를 가질것이라 기대한다. 이러한 기대를 만족시키는 가정이 있으며 최대우도추정량의 정규성 증명에 사용되는 정규성 가정(Regularity Condition)이 예 해당하는 특수한 조건이다. 한편 이 항은 표본 추출로 인한 무작위성에 의해 발생하며, 표본이 바뀔 때마다 값이 변하는 순수한 확률적 성분이므로 확률적 요동(*stochastic fluctuation*)이라 부른다. 여기서 더 엄밀하게 따져봐야하는 것은 θ가 θ_0 로 수렴하는 랜덤한 값인 경우 여전히 중심극한정리 해서 얻어지는 동이랑 결과 의미 하는지 혹은 글을 위해 필요한 조건이 무엇인지에 대한 것이다.

다음으로 $P(m_\theta - m_{\theta_0})$ 을 살펴 보자. 여기에는 랜덤한 값이 없다. $\theta \neq \theta_0$인 경우에 우리의 가정으로 부터 항상 $P(m_\theta - m_{\theta_0}) < 0$ 임을 할 수 있다. 즉, 추정값 θ가 θ_0와 다른 경우 발생하는 $M_n(\theta) - M_n(\theta_0)$의 감소분에 해당한다. 이 감소의 크기는 θ에 대한 m_θ의 변화량에 의존하며 많은 경우 m_θ 가 θ_0 근방에서 이차근사가 가능한 형태를 사용한다.

$\hat{\theta}$가 $M_n(\theta)$의 최대점이므로 항상 $0 \leq M_n(\hat{\theta}) - M_n(\theta_0)$ 가 성립한다. 이를 core decomposition에

대입하면

$$0 \le (P_n - P)(m_{\hat{\theta}} - m_{\theta_0}) + P(m_{\hat{\theta}} - m_{\theta_0}).$$

이를 정리하면 다음의 핵심 부등식을 얻는다:

$$P(m_{\theta_0} - m_{\hat{\theta}}) \le \left|(P_n - P)(m_{\hat{\theta}} - m_{\theta_0})\right|.$$

즉, M-estimator $\hat{\theta}$의 excess risk 는

$$\mathcal{R}(\hat{\theta}) \le \left|(P_n - P)(m_{\hat{\theta}} - m_{\theta_0})\right|.$$

로 경험과정의 확률적 요동에 의해 제한된다.

3 Consistency

M-estimator의 consistency는 다음 두 조건으로 요약된다.

1. (식별성) θ_0는 $M(\theta)$의 유일한 최대점이다.
2. (ULLN) $M_n(\theta)$는 $M(\theta)$로 균일 수렴한다:

$$\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \xrightarrow{p} 0.$$

이 두 조건이 만족되면, 경험적 최대점 $\hat{\theta}$는 최대점 θ_0로 수렴한다. 이제 이 균일 수렴 조건이 언제 성립하는지를 단계적으로 살펴본다.
(Parameter set이 유한한 경우) 먼저 $\Theta = \{\theta_1, \dots, \theta_K\}$가 유한 집합이라고 하자. 각 θ_k에 대해,대수의 법칙에 의해 $M_n(\theta_k) \xrightarrow{p} M(\theta_k)$ 이 성립한다. 한편 임의의 $\varepsilon > 0$에 대해

$$\left\{ \max_{1 \le k \le K} |M_n(\theta_k) - M(\theta_k)| > \varepsilon \right\} = \bigcup_{k=1}^{K} \left\{ |M_n(\theta_k) - M(\theta_k)| > \varepsilon \right\}.$$

따라서 확률공리에 의해

$$\mathbb{P}\left(\max_{1 \le k \le K} |M_n(\theta_k) - M(\theta_k)| > \varepsilon \right) \le \sum_{k=1}^{K} \mathbb{P}(|M_n(\theta_k) - M(\theta_k)| > \varepsilon).$$

K가 n에 의존하지 않는 고정된 값이므로 좌변항을 임의로 작게 만드는 sample size N을 잡을 수 있다 (구체적으로 설명해보아라!). 따라서

$$\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \xrightarrow{p} 0.$$

따라서 Θ 가 유한한 경우 ULLN 성립한다.
(Parameter set이 무한한 경우) 이제 Θ가 무한 집합(예: 구간, 유클리드 공간)이라고 하자. 각 고정된 θ에 대해서는 여전히 $M_n(\theta) \xrightarrow{p} M(\theta)$ 가 성립한다. 그러나 이것만으로는 $\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)|$ 의 수렴을 보장할 수 없다. 왜냐하면 우리가 이전 증명에서 사용한 부등식에서 유한한 K로 확률의 상한을 잡을 수 없기 때문이다. 이때부터 문제가 단순한 확률 문제가 아니라 Θ들을 얼마나 조밀하게 잘 쪼갤 수 있고 그 안에서 M 또는 M_n의 변화가 작은지 혹은 큰지를 확인하는 문제로 바뀌게 된다. 기본 적인 아이디어는 함수 값에 변화가 작은 구간을 만들었을 때 전체 Θ가 몇 개의 열린 공(open ball)으로 덮을 수 있는 가, 혹은 더 이상 공을 추가할 수 없을때 까지 채워 넣을 수 있는가의 문제가 된다. 전자에서 공의 개수는 Θ 에 대한 covering number, 후자는 packing number 라 부른다. 경험적 확률 과정에서는

$$\sup_{\theta\in\Theta} |(P_n - P)m_\theta| \xrightarrow{p} 0$$

에 대한 성질을 다룬다.

4 Rate of Convergence

우리의 1차적인 목표는 샘플 수가 증가함에 따라 추정량 $\hat{\theta}$ 이 θ_0로 가까이 가는 상황에서 $M_n(\theta) - M_n(\theta_0)$ 의 수렴속도를 밝히는 것이다. 수렴속도를 알게 되면 $M_n(\theta) - M_n(\theta_0)$ 의 선형근사를 통해 $\hat{\theta} - \theta_0$의 수렴속도를 알 수 있게 되고, 이를 통해 우리는 $\hat{\theta} - \theta_0$항의 점근적 근사를 통한 추론이 가능하게 된다. (중심극한정리를 생각해 보아라!). 먼저 core decomposition 을 통해 $M_n(\theta) - M_n(\theta_0)$를 구성하는 항을 각각 살펴보고 $\hat{\theta}$ 이 θ_0로 가는 상황에서 각 항이 어떤 특성을 가지게 될지 예상해보자.
먼저 결정적 항에 대해서는 모집단 목적함수의 국소적 식별성과 곡률을 다음과 같이 가정한다:

$$P(m_{\theta_0} - m_\theta) \geq c\|\theta - \theta_0\|^\alpha, \qquad \|\theta - \theta_0\| \text{ 작을 때.} \tag{H1}$$

여기서 $\alpha > 0$는 risk의 Hölder 차수이다. 예를 들어, smooth한 M-estimator의 경우 $\alpha = 2$이며, Manski의 maximum score에서는 $\alpha = 1$이 된다. 이를 m_θ에 대한 Hölder-type 가정을 통해 얻는다.
다음으로 확률적 요동에 대해 국소적인 modulus of continuity를 가정해보자. θ_0 근방에서 확률적 항이

다음과 같이 제한된다고 가정해보자.

$$\sup_{\|\theta-\theta_0\|\le\delta} \big|(P_n - P)(m_\theta - m_{\theta_0})\big| \le C\,\frac{1}{\sqrt{n}}\,\delta^\beta, \qquad \text{w.h.p.} \tag{H2}$$

$\beta > 0$는 함수 클래스의 Hölder 연속성을 나타내는 지수이다. 이는 경험과정의 CLT 스케일 $n^{-1/2}$과 θ 변화에 따른 함수 변동 크기를 동시에 반영한 조건이다. 우리가 여기서 주목해야 될 것은 1) 결정적 항과 확률적 항 모두 $\hat\theta$ 이 θ_0로 가는 상황에서 0으로 수렴한다는 사실이고 2) $\hat\theta - \theta_0$ 가 0으로 가는 속도에 따라 두 항이 0으로 가는 속도를 다르게 가정했다는 사실이다. 이런 경우 $M_n(\theta) - M_n(\theta_0)$의 수렴속도는 어떤 항에 의존하게 될까?
(H1), (H2)를 핵심 부등식에 적용하면

$$c\|\hat\theta - \theta_0\|^\alpha \le C\,\frac{1}{\sqrt{n}}\,\|\hat\theta - \theta_0\|^\beta.$$

이로부터 $\|\hat\theta - \theta_0\|^{\alpha-\beta} \lesssim \frac{1}{\sqrt{n}}$ 를 얻으며, 따라서 $\|\hat\theta - \theta_0\| = O_p\big(n^{-1/(2(\alpha-\beta))}\big)$ 를 얻게 된다.

Remark 4.1. 우리는 어떤 앞으로 무엇을 배워 할 것인지는 명확하다.

- $\sup_{\theta\in\Theta} |(P_n - P)m_\theta| \xrightarrow{p} 0$ as $n \to \infty$ 는 어떻게 보여야 할 것인가? 언제 이 명제가 성립하는가?
- $\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \xrightarrow{p} 0$. 는 언제 성립하는가?

한편 우리의 중요한 관찰은 다음과 같다. 위의 수렴 문제에서 본질적인 대상은 θ 자체가 아니라 $\mathcal{F} := \{m_\theta : \theta \in \Theta\}$ 라는 함수들의 집합이다. 즉,

$$\sup_{\theta\in\Theta}\big|(P_n - P)m_\theta\big| = \sup_{f\in\mathcal{F}}\big|(P_n - P)f\big|$$

로 다시 쓸 수 있으며, 문제는 이제 함수공간 $\mathcal{F}$ 위에서의 경험적 확률과정의 거동을 이해하는 것으로 바뀐다. 이 관점에서는 Θ가 유클리드 공간인지, 고차원인지, 혹은 비정형적인 집합인지는 부차적인 문제가 된다. 중요한 것은 $\mathcal{F}$가 어떤 거리(metric) 하에서 얼마나 "복잡한" 함수족인지이다.
이러한 이유로 경험적 확률과정은 단순히 $\sqrt{n}(P_n - P)f$ 를 고정된 함수 f에 대해 살펴보는 이론이 아니라, $\big\{\sqrt{n}(P_n - P)f : f \in \mathcal{F}\big\}$ 와 같은 확률과정으로 이해된다. 이 강의에서는 함수족 $\mathcal{F}$의 크기와 구조를 측정하는 도구들, 예를 들어 covering number, entropy, VC-type 조건 등을 통해 언제 위와 같은 균등수렴이 성립하는지를 살펴볼 것이다. 이러한 준비는 이후 M-estimator의 일관성(consistency)과 수렴률을 분석하는 데 핵심적인 역할을 하게 된다.

5 MLE 점근적 성질 증명을 위한 가정의 검토

이번 소절에서는 기호를 간단히 쓰기 위해 참분포 P_{θ_0}를 P로 표기하였다. P_{θ_0}의 확률밀도 f_{θ_0}가 존재하는 경우에 $Pg = \int g(x) f_{\theta_0}(x) dx$ 임을 의미한다. MLE는

$$\hat{\theta}_n \in \arg\max_{\theta \in K} M_n(\theta)$$

로 정의한다. 여기서 $K \subset \Theta$는 compact 집합이며 θ_0는 K의 interior point이다.

5.1 Classical MLE의 Regularity Conditions

최대우도추정량(MLE)의 점근적 정규성(asymptotic normality)은 일반적으로 여러 개의 정규성 가정(regularity conditions) 하에서 증명된다. 대표적인 조건들은 다음과 같다. 여기서는 일반적인 MLE의 증명에서 사용하는 log density를 $\ell_\theta = \log p_\theta$로 표기하나 여기서는 m_θ로 표기하겠다.

(R1) $M(\theta) = Pm_\theta$는 θ_0에서 유일한 최대점을 갖는다.

(R2) $\theta \mapsto m_\theta(x)$는 거의 모든 x에 대해 연속이며, 임의의 $\theta \in K$에 대해 θ의 근방 $U(\theta) \subset K$가 존재하여 $m_{U(\theta)}(x) := \sup_{\vartheta \in U(\theta)} m_\vartheta(x)$ 가 적분가능하다. 즉, $P|m_{U(\theta)}| < \infty$[1].

(R3) $m_\theta(x)$는 θ에 대해 θ_0 근방에서 두번 연속미분 (continuously differentiable) 가능하고 모든 θ에 대해 $\nabla P m_\theta = P \nabla m_\theta$ 와 $\nabla^2 P m_{\theta_0} = P\left[\nabla^2 m_{\theta_0}\right]$ 가 성립한다.

(R4) $I(\theta_0) := P\left[\nabla m_{\theta_0} \nabla m_{\theta_0}^\top\right]$ 는 유한하고 positive definite 다.

(R5) 어떤 $\delta_0 > 0$와 함수 $H : \mathcal{X} \to [0, \infty)$가 존재하여, 거의 모든 x에 대해 $\theta \mapsto m_\theta(x)$는 $B(\theta_0, \delta_0)$에서 두 번 연속미분 가능하고 $\sup_{\|\theta - \theta_0\| \le \delta_0} \left\|\nabla^2 m_\theta(x)\right\| \le H(x)$ a.s. 이며 $PH^2 < \infty$ 가 성립한다.

이 조건들 하에서 다음이 성립한다:

$$\sqrt{n}(\hat{\theta} - \theta_0) \Rightarrow N\left(0, I(\theta_0)^{-1}\right).$$

그러나 위 조건들은 서로 독립적인 가정이라기보다는, 결국 다음 두 가지 구조를 보장하기 위한 충분조건들로 이해할 수 있다.

[1] ULLN 이 필요하지 않은 경우에는 $Pm_{U(\theta)} < \infty$ 로 충분함.

5.2 Consistence 증명을 위한 Regularity Condition의 분석

Theorem 5.1. (*R1*)–(*R2*) 하에서 $\hat{\theta}_n \xrightarrow{p} \theta_0$.

Step 1: Wald 의 MLE 증명으로 부터 ULLN을 유도한다.
$\theta \in K$를 고정하자. 반지름이 0으로 감소하는 열린 공들의 열을 택하자. $U_j(\theta) := B(\theta, r_j) \cap K, \quad r_j \downarrow 0$. 그러면 집합 포함관계 $U_{j+1}(\theta) \subset U_j(\theta)$로부터 $m_{U_j(\theta)}(x) := \sup_{\vartheta \in U_j(\theta)} m_\vartheta(x)$는 j에 대해 단조감소한다:

$$m_{U_{j+1}(\theta)}(x) \leq m_{U_j(\theta)}(x).$$

또한 (R2)의 연속성으로 거의 모든 x에서 $m_{U_j(\theta)}(x) \downarrow m_\theta(x) \qquad (j \to \infty)$.[2] 어떤 j_0에 대해 $U_{j_0}(\theta) \subset U(\theta)$이고, 따라서 $m_{U_{j_0}(\theta)}(x) \leq m_{U(\theta)}(x)$, 이므로 (R2)에 의해 $Pm_{U_{j_0}(\theta)} \leq Pm_{U(\theta)} < \infty$를 얻는다. 단조수렴정리(monotone convergence theorem)에 따라

$$Pm_{U_j(\theta)} \downarrow Pm_\theta = M(\theta). \tag{2}$$

임의의 $\varepsilon > 0$에 대해 $A_\varepsilon := \{\theta \in K : \|\theta - \theta_0\| \geq \varepsilon\}$ 라 두자. (R1)에 의해 $\theta \in A_\varepsilon$이면 $M(\theta) < M(\theta_0)$이다. (2)의 수렴 $Pm_{U_j(\theta)} \downarrow M(\theta)$를 이용하면, 각 $\theta \in A_\varepsilon$에 대해 어떤 $j(\theta)$가 존재하여 $Pm_{U_{j(\theta)}(\theta)} < M(\theta_0)$. $U_\theta := U_{j(\theta)}(\theta)$로 두면, $\{U_\theta : \theta \in A_\varepsilon\}$는 A_ε의 open cover이다. 한편 A_ε가 compact이므로 유한 부분덮개가 존재한다: $A_\varepsilon \subset \bigcup_{k=1}^N U_k,\ U_k := U_{\theta_k}$. 각 k에 대해 m_{U_k}는 적분가능하므로(가정 (R2)), 대수의 법칙에 의해 $P_n m_{U_k} \xrightarrow{p} Pm_{U_k} \qquad (k = 1, \ldots, N)$. 유한 개이므로 $\max_{1 \leq k \leq N} \left| P_n m_{U_k} - Pm_{U_k} \right| \xrightarrow{p} 0$. 이 성립한다. 한편

$$\left\{ \sup_{\theta \in A_\varepsilon} |P_n m_\theta - Pm_\theta| > \varepsilon \right\} \subset \left\{ \max_{1 \leq k \leq N} |P_n m_{U_k} - Pm_{U_k}| > \varepsilon \right\}$$

이므로

$$\sup_{\theta \in A_\varepsilon} |P_n m_\theta - Pm_\theta| \xrightarrow{p} 0. \quad \text{(ULLN on } A_\varepsilon\text{)}$$

LLN으로 $M_n(\theta_0) = P_n m_{\theta_0} \to M(\theta_0)$임을 안다. 한편 확률적으로 충분히 큰 n에서 $M_n(\theta_0) > M(\theta_0) - \frac{1}{2}\eta_\varepsilon$. ULLN on A_ϵ은 의해 확률적으로 충분히 큰 n에서 $\sup_{\theta \in A_\varepsilon} |P_n m_\theta - Pm_\theta| \leq \varepsilon$를 의미하므로 $M_n(\theta_0)$의 확률적 수렴성을 통해 $\sup_{\theta \in B_\varepsilon} M_n(\theta) < M_n(\theta_0)$,을 알 수 있다. 왜냐하면 $M_n(\theta_0)$는 $M(\theta)$의 최대값으로 수렴해가고 θ_0를 포함하는 작은 ball 의 바깥(A_ϵ)에서는 $M_n(\theta)$ 이 $M(\theta)$로 균등수렴하므로 $\sup_{\theta \in B_\varepsilon} M_n(\theta) < M_n(\theta_0)$ w.h.p 의 결론을 얻는다. 이는 $\hat{\theta}$이 $M_n(\theta)$ 의 최대값이라는

[2]만약 $m_{U_j(\theta)}(x) \downarrow m_\theta(x)$ 가 성립하지 않는다고 하면 모순을 이끌어 낼 수 있다. $\theta_j \in U_j(\theta)$ 인 열 $\{\theta_j\}$은 $U_j(\theta)$의 정의에 의해 항상 θ로 수렴하는 수열이다. 하지만 $\lim_j m_{U_j(\theta)}(x) > m_\theta(x)$ 는 $m_{\theta_j}(x)$가 $m_\theta(x)$로 수렴하지 않는 수열 $\{\theta_j\}$를 잡을 수 있음을 의미하며 이는 (R2) 모순이다.

MLE의 정의에 의해 $\hat{\theta}_n \notin B_\varepsilon$ w.h.p를 의미한다. 따라서

$$P(\|\hat{\theta} - \theta_0\| \geq \varepsilon) \to 0.$$

5.3 Asymptotic Normality 증명을 위한 Regularity Condition의 분석

이제 위 조건들로부터 (H1)–(H2)가 따라옴을 보인다.

(H1): 결정적 곡률(Deterministic curvature)

Proposition 5.2. (*RC3*)–(*RC4*)가 성립하면 어떤 상수 $c > 0$, $\delta_0 > 0$가 존재하여 모든 $\|\theta - \theta_0\| \leq \delta_0$에 대해

$$P(m_{\theta_0} - m_\theta) = M(\theta_0) - M(\theta) \geq c\|\theta - \theta_0\|^2.$$

즉 (*H1*)이 $\alpha = 2$로 성립한다.

Proof (*sketch*). (RC3)는 θ_0 근방에서 $\nabla^2 M(\theta)$ 의 연속성[3]을 보장해준다. 그리고 θ_0 근방에서 M의 2차 Taylor 전개를 통해

$$M(\theta) = M(\theta_0) + (\theta - \theta_0)^\top \nabla M(\theta_0) + \frac{1}{2}(\theta - \theta_0)^\top \nabla^2 M(\tilde{\theta})(\theta - \theta_0)$$

($\tilde{\theta}$는 θ_0와 θ 사이의 점) 를 얻는다. 여기서 $-\nabla^2 M(\theta_0)$ 는 (R3)에 의해 $I(\theta_0)$와 같다. $\nabla^2 M(\theta)$ 의 연속성과 (R4)에 의해서 $\theta \to 0$인 $\tilde{\theta} = h\theta + (1-h)\theta_0$ 에 대해서 $\nabla^2 M(\tilde{\theta})$는 $\nabla^2 M(\theta_0)$에 operator norm 으로 가까우므로 충분히 작은 근방에서는

$$-\nabla^2 M(\tilde{\theta}) \succeq \frac{1}{2} I(\theta_0)$$

가 되도록 δ_0를 잡을 수 있다. 이에 대한 엄밀한 증명은 참고사항을 보아라. 결국

$$M(\theta_0) - M(\theta) = -\frac{1}{2}(\theta - \theta_0)^\top \nabla^2 M(\tilde{\theta})(\theta - \theta_0) \geq \frac{1}{4}(\theta - \theta_0)^\top I(\theta_0)(\theta - \theta_0) \geq c\|\theta - \theta_0\|^2.$$

MLE의 정규성 조건을 통해 (H1)을 보일 수 있다. □

(H2): Equicontinuity (H2)는

$$\sup_{\|\theta - \theta_0\| \leq \delta} \left|(P_n - P)(m_\theta - m_{\theta_0})\right| \leq \frac{C}{\sqrt{n}} \delta^\beta \text{ w.h.p.}$$

[3] $\theta \to \theta_0$가 $\|\nabla M(\theta) - \nabla M(\theta_0)\| \to 0$ 을 의미한다.

형태의 국소 modulus of continuity이다. MLE의 매끄러운(parametric smooth) 경우에는 $\beta = 1$이 된다.

Proposition 5.3 (R3–R5 $\Rightarrow$ (H2) with $\beta = 1$). (*R3*)–(*R5*)가 성립하면, 어떤 충분히 작은 $\delta \le \delta_0$에 대해

$$\sup_{\|\theta-\theta_0\|\le\delta} \left|(P_n - P)(m_\theta - m_{\theta_0})\right| = O_p\left(\frac{\delta}{\sqrt{n}}\right) + O_p(\delta^2).$$

Proof (*structured sketch*). 각 θ에 대해 평균값 정리를 쓰면

$$m_\theta - m_{\theta_0} = (\theta - \theta_0)^\top \nabla m_{\theta_0} + R_\theta, \tag{1}$$

여기서 remainder는

$$R_\theta(x) = \frac{1}{2}(\theta - \theta_0)^\top \left(\int_0^1 \nabla^2 m_{\theta_0 + t(\theta-\theta_0)}(x)\, dt\right)(\theta - \theta_0).$$

따라서 (R5)로부터 $|R_\theta(x)| \le \frac{1}{2}\|\theta - \theta_0\|^2 H(x)$ 와 같이 나머지 항의 상한을 잡을 수 있다.
먼저 (1)의 점근적 차수를 계산한다. (1)에 $(P_n - P)$에 적용하면

$$(P_n - P)(m_\theta - m_{\theta_0}) = (\theta - \theta_0)^\top (P_n - P)\nabla m_{\theta_0} + (P_n - P)R_\theta.$$

를 얻는다. 먼저 위 등식 오른쪽 첫번째 항에 대해서는 코시-슈바르츠 부등식에 의해

$$|(\theta - \theta_0)^\top (P_n - P)\nabla m_{\theta_0}| \le \|(\theta - \theta_0)\| \|(P_n - P)\nabla m_{\theta_0}\|$$

이므로

$$\sup_{\|\theta-\theta_0\|\le\delta} \left|(\theta - \theta_0)^\top (P_n - P)\nabla m_{\theta_0}\right| \le \delta \, \|(P_n - P)\nabla m_{\theta_0}\|$$

를 얻는다. (R5)에 의해 각 성분별 CLT로 $\|(P_n - P)\nabla m_{\theta_0}\| = O_p(1/\sqrt{n})$ 이므로

$$\delta \, \|(P_n - P)\nabla m_{\theta_0}\| = O_p\left(\frac{\delta}{\sqrt{n}}\right) \tag{2}$$

앞서 $|R_\theta(x)| \le \frac{1}{2}\|\theta - \theta_0\|^2 H(x)$ 이므로

$$\sup_{\|\theta-\theta_0\|\le\delta} |(P_n - P)R_\theta| \le \frac{1}{2}\delta^2 |(P_n - P)H|.$$

$PH^2 < \infty$이므로 $(P_n - P)H = O_p(n^{-1/2})$를 얻고 따라서

$$\sup_{\|\theta-\theta_0\|\le\delta} |(P_n - P)R_\theta| = O_p(\delta^2/\sqrt{n}). \tag{3}$$

따라서 (2) 와 (3)에 의해

$$\sup_{\|\theta-\theta_0\|\le\delta} \left|(P_n - P)(m_\theta - m_{\theta_0})\right| \le O_p\left(\frac{\delta}{\sqrt{n}}\right) + O_p\left(\frac{\delta^2}{\sqrt{n}}\right)$$

결론이 따른다. 작은 δ에서는 δ^2가 더 높은 차수이므로 □

5.4 정리: classical MLE regularity는 (H1)–(H2)의 충분조건

위 두 결과를 합치면, classical MLE의 다수 regularity conditions는 결국

- **M-estimator 의 ULLN**
- **모집단 목적함수의 국소적 곡률** (H1, $\alpha = 2$),
- **경험과정 요동의 국소 Lipschitz 제어** (H2, $\beta = 1$)

를 보장하기 위한 충분조건으로 해석된다.
MLE의 정규성 가정은 (H1)-(H2)를 의미하며 우도기반 추정에서 M-estimator 의 점금적 성질을 규명하기 위한 특별한 조건임을 확인하였다. M-estimator 의 일반적인 경우로 확장을 하기 위해서는 새로운 조규성 가정을 각 경우 마다 만드는 것이 아니라 (H1)과 (H2)를 보장할 수 있는 일반적인 가정을 탐구할 필요가 있다. 한편 (H1)–(H2)는 MLE의 매끄러운 경우를 포함하되, 매끄러움이 깨지는 경우에도 적용 가능하다. 예를 들어,

- 비매끄러운 목적함수에서는 (H1)의 지수 α가 2보다 작아질 수 있다(예: $\alpha = 1$).
- 함수족의 복잡도가 커지면 (H2)의 지수 β가 1보다 작아질 수 있으며, 이때 수렴률은

$$\|\hat{\theta} - \theta_0\| = O_p\big(n^{-1/(2(\alpha-\beta))}\big)$$

로 느려진다.

따라서 (H1)–(H2)는 classical MLE를 “특수한 매끄러운 경우”로 포함하면서도, 경험과정(covering number/entropy 등)을 통해 비모수적/비매끄러운 상황으로 확장될 수 있다.

6 보충 설명

Lemma 6.1 (Local uniform positive definiteness). *1.* $\theta \mapsto \nabla^2 M(\theta)$는 θ_0에서 *operator norm* $\|\cdot\|_{\text{op}}$에 대해 연속이다.

2. $I(\theta_0) := -\nabla^2 M(\theta_0)$는 *positive definite*이다.

그러면 어떤 $\delta_0 > 0$가 존재하여, $\|\theta - \theta_0\| \le \delta_0$를 만족하는 모든 θ에 대해 $-\nabla^2 M(\theta) \succeq \frac{1}{2} I(\theta_0)$ 가 성립한다.

Proof. $A(\theta) := -\nabla^2 M(\theta)$로 두면 $A(\theta_0) = I(\theta_0) \succ 0$ 이다. 따라서 $A(\theta_0)^{1/2}$ 및 $A(\theta_0)^{-1/2}$가 존재한다. 정규화된 행렬을

$$B(\theta) := A(\theta_0)^{-1/2} A(\theta) A(\theta_0)^{-1/2}$$

로 정의하면 $B(\theta_0) = I_d$이다. 이제 $C := A(\theta_0)^{-1/2}$로 두면

$$B(\theta) - I_d = C\big(A(\theta) - A(\theta_0)\big)C$$

이므로 operator norm의 submultiplicativity[4]에 의해

$$\|B(\theta) - I_d\|_{\text{op}} \le \|C\|_{\text{op}}^2 \, \|A(\theta) - A(\theta_0)\|_{\text{op}}.$$

가정 1로 $\|A(\theta) - A(\theta_0)\|_{\text{op}} \to 0$ 이므로 $\|B(\theta) - I_d\|_{\text{op}} \to 0$ 이다. 따라서 어떤 $\delta_0 > 0$가 존재하여 $\|\theta - \theta_0\| \le \delta_0$이면

$$\|B(\theta) - I_d\|_{\text{op}} \le \frac{1}{2}.$$

임의의 단위벡터 u에 대해

$$u^\top B(\theta) u = u^\top I_d u + u^\top (B(\theta) - I_d) u \ge 1 - \|B(\theta) - I_d\|_{\text{op}} \ge \frac{1}{2}$$

이므로 $B(\theta) \succeq \frac{1}{2} I_d$이다. 이제 임의의 v를 잡고 $w := A(\theta_0)^{1/2} v$ 라고 두면,

$$v^\top A(\theta) v = v^\top A(\theta_0)^{1/2} B(\theta) A(\theta_0)^{1/2} v = w^\top B(\theta) w.$$

[4]operator norm의 submultiplicativity

$$\|XYZ\|_{\text{op}} \le \|X\|_{\text{op}} \|Y\|_{\text{op}} \|Z\|_{\text{op}}$$

그런데 $B(\theta) \succeq \frac{1}{2} I_d$ for $\|\theta - \theta_0\| \leq \delta_0$이므로

$$w^\top B(\theta) w \geq \frac{1}{2} w^\top w = \frac{1}{2}\|w\|^2 = \frac{1}{2} v^\top A(\theta_0) v.$$

따라서 임의의 v에 대해

$$w^\top B(\theta) w - \frac{1}{2} v^\top A(\theta_0) v = v^\top \Big(A(\theta) - \tfrac{1}{2} A(\theta_0) \Big) v \geq 0$$

가 성립하고, 이는 곧

$$A(\theta) \succeq \frac{1}{2} A(\theta_0) = \frac{1}{2} I(\theta_0)$$

임을 의미한다.
□

Remark (Frobenius 노름의 연속성) 앞선 Lemma에서는 $\nabla^2 M(\theta)$가 operator norm $\|\cdot\|_{\text{op}}$에 대해 θ_0에서 연속이라고 가정하였다. 그러나 실제로는 Frobenius 노름 $\|\cdot\|_F$에 대한 연속성만으로도 충분하다. 유한차원에서 두 행렬 노름은 서로 동치이므로, 특히 다음 부등식이 항상 성립한다:

$$\|A\|_{\text{op}} \leq \|A\|_F.$$

이를 간단히 확인해보면,

$$\|A\|_{\text{op}} = \sup_{\|x\|=1} \|Ax\| = \sqrt{\lambda_{\max}(A^T A)},$$

한편

$$\|A\|_F^2 = \text{tr}(A^T A) = \sum_{i=1}^{d} \lambda_i(A^T A) \geq \lambda_{\max}(A^T A),$$

이므로

$$\|A\|_{\text{op}} \leq \|A\|_F.$$

따라서 만약

$$\|A(\theta) - A(\theta_0)\|_F \to 0 \quad (\theta \to \theta_0)$$

이면,

$$\|A(\theta) - A(\theta_0)\|_{\text{op}} \to 0$$

도 자동으로 성립한다. 결국 $\nabla^2 M(\theta)$가 Frobenius 노름에 대해 θ_0에서 연속이면, operator norm에

대해서도 연속이 되고, 앞서 증명한 국소적 양의 정부호성 결과

$$-\nabla^2 M(\theta) \succeq \frac{1}{2} I(\theta_0)$$

는 동일하게 성립한다.