# Bayesian Asymptotics

IDEA lab

Department of Statistics, Seoul National University

March 31, 2026

# Outline

# About MLE convergence

**Setup**

- Observations: $X_1, \ldots, X_n \overset{i.i.d.}{\sim} P_0$
- Function class: $\Theta$ (finite or infinite dimension)
- Log likelihood: $\ell_\theta(x) = \log p_\theta(x)$
- Target:

$$\theta^* \in \underset{\theta \in \Theta}{\operatorname{argmax}} P_0 \ell_\theta \tag{1}$$

  where $Pg = \int g dP$. ($P_0 = P_{\theta^*}$ if well-specified)

- MLE:

$$\hat{\theta}_n \in \underset{\theta \in \Theta}{\operatorname{argmax}} P_n \ell_\theta \tag{2}$$

  where $P_n g := \frac{1}{n} \sum_{i=1}^n g(X_i)$.

# About MLE convergence

- What we want:

$$d(\hat{\theta}_n, \theta^*) = O_p(r_n) \tag{3}$$

  for some distance $d$ (e.g. Hellinger, L2, ...).
  - Parametric: usually $r_n = n^{-1/2}$
  - Nonparametric: $r_n$ becomes slower depend on model complexity
- What we consider - excess risk:

$$P_0(\ell_{\theta^*} - \ell_{\hat{\theta}_n}) \tag{4}$$

- Then we consider some relationship between $d$ and the excess risk.
- Also, we have $P_0(\ell_{\theta^*} - \ell_{\hat{\theta}_n}) \leq (P_n - P_0)(\ell_{\hat{\theta}_n} - \ell_{\theta^*})$, since $P_n \ell_{\hat{\theta}_n} \geq P_n \ell_{\theta^*}$.

# About MLE convergence

- Example on Hellinger distance:

$$d_H^2(p,q) = \frac{1}{2}\int \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 d\mu(x) \tag{5}$$

$$d_H^2(p_{\hat{\theta}_n}, p_{\theta^*}) \lesssim P_0(\ell_{\theta^*} - \ell_{\hat{\theta}_n}) \leq (P_n - P_0)(\ell_{\hat{\theta}_n} - \ell_{\theta^*}) \tag{6}$$

# About MLE convergence

**Point-wise convergence**

- We can use CLT to yield point-wise convergence.

$$\sqrt{n}(P_n - P_0)(\ell_\theta - \ell_{\theta^*}) \Rightarrow \mathcal{N}(0, \mathrm{Var}_{P_0}(\ell_\theta(X) - \ell_{\theta^*}(X))) \tag{7}$$

  for each $\theta$.

- However, this cannot be implemented to MLE, since $\hat{\theta}_n$ is random depend on the observations.

# About MLE convergence

**How to yield MLE convergence?**

- Basic strategy:

$$\left|(P_n - P_0)(\ell_{\hat{\theta}_n} - \ell_{\theta^*})\right| \leq \sup_{\theta \in \Theta} |(P_n - P_0)(\ell_\theta - \ell_{\theta^*})| \tag{8}$$

  We need to control the supremum term.

- To control the supremum term, we need to control following tail probabilities:

$$\Pr\left(\sup_{\theta \in \Theta} |(P_n - P_0)(\ell_\theta - \ell_{\theta^*})| \geq t\right) \tag{9}$$

  We call the probabilistic bounds on the tail probabilities as **large deviation bounds**.

# About MLE convergence

**Additional techniques**

- When $\Theta$ is too big, one can consider sieve $\Theta_n \subset \Theta$.
- Sieve MLE:

$$\hat{\theta}_n \in \underset{\theta \in \Theta_n}{\operatorname{argmax}} P_n \ell_\theta \tag{10}$$

- Let $\theta_n^\dagger \in \underset{\theta \in \Theta_n}{\operatorname{argmax}} P_0 \ell_\theta$, then we can derive the convergence rate as follows:

$$d(\hat{\theta}_n, \theta^*) \lesssim d(\theta_n^\dagger, \theta^*) + r_n \tag{11}$$

where the first term (of right-hand side) implies approximation error and the second term implies estimation error.

# Outline

# Set-up

- $X_1, \ldots, X_n \overset{i.i.d.}{\sim} f(x|\theta^*)$ for some $\theta^*$ in $\Theta$.
- Let $\hat{\theta}$ be the maximum likelihood estimator of $\theta$.
- We are going to prove that $\sqrt{n}(\hat{\theta} - \theta^*)$ converges in distribution to the normal distribution with mean 0 and covariance matrix $\mathbb{I}^{-1}(\theta^*)$, where $\mathbb{I}(\theta) = -\mathbb{E}(\partial^2 f(X|\theta)/\partial\theta\partial\theta)$.

# Prove

- Let $\ell_n(\theta) = \sum \log f(X_i|\theta)$ and let $s_n(\theta) = \partial \ell_n(\theta)/\partial\theta$.
- Note that $s_n(\hat{\theta}) = 0$ by the definition of the MLE.
- In addition, we have $\mathbb{E}(\partial \log f(X|\theta)/\partial\theta)\big|_{\theta=\theta^*} = 0$, which implies that $\mathbb{E}(s_n(\theta^*)) = 0$.
- Thus, the central limit theorem implies that

$$\frac{s_n(\theta^*)}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, \mathbb{E}(s(\theta^*)s^\top(\theta^*))),$$

where $s(\theta) = \partial \log f(X|\theta)/\partial\theta$.

# Prove

- Now, Taylor expansion yields

$$0 = s_n(\hat{\theta}) \approx s_n(\theta^*) + [\partial s_n(\theta)/\partial\theta)](\hat{\theta} - \theta^*).$$

- Thus we have

$$\hat{\theta} - \theta^* \approx [\partial s_n(\theta)/\partial\theta)]^{-1} s_n(\theta^*),$$

and so

$$\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{d} \mathcal{N}(0, [\partial s_n(\theta)/\partial\theta)/n]^{-1}\mathbb{E}(s(\theta^*)s^\top(\theta^*)))[\partial s_n(\theta)/\partial\theta)/n]^{-1})$$

- Finally, by changing the integration and derivative operators, we can show that

$$[\partial s_n(\theta)/\partial\theta)/n]^{-1}\mathbb{E}(s(\theta^*)s^\top(\theta^*)))[\partial s_n(\theta)/\partial\theta)/n]^{-1} \to \mathbb{I}(\theta^*)^{-1}.$$

# Outline

# Problem

- We call that a given model $\{f(x|\theta), \theta \in \Theta\}$ is nonparametric if the dimension of $\Theta$ is infinity.
- A popular way of estimating $\theta$ is a sieve MLE.
- We call $\Theta_n, n = 1, \ldots$ a sieve if $\Theta_n$ is increasing, the dimension of $\Theta_n$ is finite and $\Theta_n \approx \Theta$ well.
- Let $\hat{\theta}$ be the sieve MLE (i.e. the maximizer of the log-likelihood on $\Theta_n$).
- We want to know how fast $\hat{\theta}$ converges to $\theta^*$ in terms of $n$.
- A problem is that it is hardly possible to say something about the convergence of $\hat{\theta}$ to $\theta^*$ since the dimension of $\theta^*$ is infinite.

# Excess Risk

- Instead, we try to say something about the convergence rate of $\mathcal{E}(\hat{\theta})$, where

$$\mathcal{E}(\theta) = \mathbb{E}\ell(X, \theta) - \mathbb{E}\ell(X, \theta^*)$$

  for a given loss $\ell(X, \theta)$.
- Here, $\hat{\theta}$ is the minimizer of $\sum_{i=1}^{n} \ell(X_i, \theta)$ on $\theta \in \Theta_n$. and $\theta^* = \operatorname{argmin}_{\theta \in \Theta} \mathbb{E}\ell(X, \theta)$.
- When $\ell(X, \theta)$ is the negative log-likelihood, $\hat{\theta}$ becomes a sieve MLE.

# Techniques

- Let $\mathbb{E}_n \ell(X, \theta) = \sum_{i=1}^n \ell(X_i, \theta)/n$.
- By the law of large numbers, we have

$$\mathbb{E}_n \ell(X, \theta) \approx \mathbb{E}\ell(X, \theta).$$

- Under regularity conditions, we can show that this convergence holds uniformly in $\theta$.
- If $\mathbb{E}\ell(X, \theta)$ is convex and has the minimizer at $\theta^*$, we expect that $\hat{\theta}$ is close to $\theta^*$ and thus $\mathcal{E}(\hat{\theta})$ converges to 0.
- Read references for details.

# Outline

# Paper

Maximum Likelihood Estimation of Misspecified Models. White. 1982.

# Problem

- $X_1, \ldots, X_n \overset{i.i.d.}{\sim} G(x)$
- Since $G$ is unknown, we choose $F(x|\theta)$ which may or may not contain true structure $G$.
- We define quasi-log-likelihood of the sample as $L_n(X, \theta) = n^{-1} \sum_i^n \log f(X_i, \theta)$ and $\hat{\theta}_n = \text{argmax}_{\theta \in \Theta} L_n(X, \theta)$.
- We also define

$$\begin{aligned} \theta^* &= \text{argmin}_{\theta \in \Theta} KL(g(x)||f(x;\theta)) \\ &= \text{argmin}_{\theta \in \Theta} (\mathbb{E}_g[\log(g(x))] - \mathbb{E}_g[\log(f(x;\theta)]) \\ &= \text{argmax}_{\theta \in \Theta} \mathbb{E}_g[\log f(x;\theta)] \end{aligned}$$

# Assumptions

## Assumption1, A1

The independent random 1 x M vectors $X_t, t = 1, ..., n$, have common joint distribution function $G$ on $\Omega$, a measurable Euclidean space, with measurable Radon-Nikodym density $g = dG/d\nu$.

## Assumption2, A2

The family of distribution $F(x, \theta)$ has Randon-Nikodym densities $f(x, \theta) = dF(x, \theta)/d\nu$ which are measurable in $x$ for every $\theta$ in $\Theta$, a compact subset of a p-dimensional Euclidean space, and continuous in $\theta$ for every $x$ in $\Omega$.

## Assumption3, A3

- (a) $\mathbb{E}_g(\log(g(X)))$ exists and $|\log(f(x, \theta))| \leq m(x)$ for all $\theta$ in $\Theta$, where m is integrable with respect to G.
- (b) $KL(g(x)||f(x; \theta))$ has a unique minimum at $\theta^*$ in $\Theta$.

# Consistency

## THEOREM 1

Given A1-A3, $\hat{\theta}_n \to \theta^*$ as $n \to \infty$ for almost every sequence $(x_t)$; i.e., $\hat{\theta}_n \xrightarrow{a.s.} \theta^*$.

- By A3 (a), $\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} \log f(x_i; \theta) - \mathbb{E}_g[\log f(X; \theta)] \right| \xrightarrow{a.s.} 0 \quad$ as $n \to \infty$.
- By A3 (b) and A2, $\Theta$ is compact and $\mathbb{E}_g[\log(f(X; \theta)]$ has unique maximum for $\theta$.
- Using Argmax Theorem, $\operatorname{argmax}_{\theta \in \Theta} L_n(X, \theta) \to \operatorname{argmax}_{\theta \in \Theta} \mathbb{E}_g[\log f(X; \theta)]$.
- Therefore, $\hat{\theta}_n \to \theta^*$.

# Assumptions

## Assumption4, A4

$\frac{\partial \log(f(x,\theta))}{\partial \theta_i}, i = 1, ..., p$,are measurable functions of $x$ for each $\theta$ in $\Omega$ and continuously differentiable functions of $\theta$ for each $x$ in $\Omega$.

## Assumption5, A5

$|\frac{\partial^2 \log(f(x,\theta))}{\partial \theta_i \partial \theta_j}|$ and $|\frac{\partial \log(f(x,\theta))}{\partial \theta_i} \cdot \frac{\partial \log(f(x,\theta))}{\partial \theta_j}|, i, j = 1, ..., p$ are dominated by functions integrable with respect to $G$ for all $x$ in $\Omega$ and $\theta$ in $\Theta$.

## Assumption6, A6

- (a) $\theta^*$ is interior to $\Theta$.
- (b) $\mathbb{E}_g[\frac{\partial \log(f(x,\theta))}{\partial \theta_i} \cdot \frac{\partial \log(f(x,\theta))}{\partial \theta_j}]\big|_{\theta=\theta^*}$ is nonsingular.
- (c) $\theta^*$ is a regular point of $\mathbb{E}_g[\frac{\partial^2 \log(f(x,\theta))}{\partial \theta_i \partial \theta_j}]$

# Asymptotic Normality

- For convenience,

$$A_n(\theta) = \left\{ n^{-1} \sum_{t=1}^{n} \partial^2 \log f(x_t, \theta)/\partial\theta_i \partial\theta_j \right\}$$

$$B_n(\theta) = \left\{ n^{-1} \sum_{t=1}^{n} \partial \log f(x_t, \theta)/\partial\theta_i \cdot \partial \log f(x_t, \theta)/\partial\theta_j \right\}$$

$$A(\theta) = \mathbb{E}_g[(\partial^2 \log f(x_t, \theta)/\partial\theta_i \partial\theta_j)]$$

$$B(\theta) = \mathbb{E}_g[(\partial \log f(x_t, \theta)/\partial\theta_i \cdot \partial \log f(x_t, \theta)/\partial\theta_j)]$$

- By A4, A5,

$$\sup_{\theta \in \Theta} |A_n(\theta) - A(\theta)| \xrightarrow{a.s.} 0 \quad \text{as } n \to \infty \tag{12}$$

$$A_n(\theta^*) \xrightarrow{p} A(\theta^*)$$

# Asymptotic Normality

- Let $s_n(\theta) = \partial L_n(\theta)/\partial\theta$ and $s(\theta) = \partial \log f(X|\theta)/\partial\theta$.
- In addition, we have $\mathbb{E}_g[\partial \log f(X|\theta)/\partial\theta]\big|_{\theta=\theta^*} = 0$.
- Thus, the central limit theorem implies that

$$\sqrt{n}s_n(\theta^*) \xrightarrow{d} \mathcal{N}(0, \mathbb{E}_g(s(\theta^*)s^T(\theta^*))$$

where $\mathbb{E}_g[s(\theta^*)s^T(\theta^*)] = B(\theta^*)$

# Asymptotic Normality

## Theorem 3

Given Assumptions A1-A6, $\sqrt{n}(\hat{\theta} - \theta^*) \sim N(0, A(\theta^*)^{-1}B(\theta^*)A(\theta^*)^{-1})$.

- Now, Taylor expansion yields

$$0 = s_n(\hat{\theta}) \approx s_n(\theta^*) + [\partial s_n(\theta)/\partial\theta)](\hat{\theta} - \theta^*).$$

- Thus we have

$$\hat{\theta} - \theta^* \approx -[\partial s_n(\theta)/\partial\theta)]\big|^{-1}_{\theta=\theta*} s_n(\theta^*),$$

and so

$$\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{d} \mathcal{N}(0, (-A(\theta^*))^{-1}B(\theta^*)(-A(\theta^*))^{-1})$$

- However,

$$A(\theta^*)^{-1}B(\theta^*)A(\theta^*)^{-1} \not\to A(\theta^*)^{-1}.$$

# Outline

# Paper

- Convergence Rate of Sieve Estimates. Shen and Wong. 1994.

# Problem

- Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} f(x|\theta^*)$ for some $\theta^*$ in $\Theta$.
- Let $l : \mathcal{X} \times \Theta \to \mathcal{R}$ be a suitably chosen function.
- Let $\hat{\theta}_n$ be an estimate defined by maximizing the empirical criterion $L_n(\theta) = (1/n) \sum_{i=1}^{n} l(X_i, \theta)$ in the following sense:

$$L_n(\hat{\theta}_n) \geq \sup_{\theta \in \Theta} L_n(\theta) - \eta_n$$

  where $\eta_n \to 0$ as $n \to \infty$.
- If $l(x, \theta) = \log f(x|\theta)$, The MLE $\hat{\theta}_n$ is obtained by maximizing $L_n(\theta)$.
- The MLE may be inconsistent if the size of the underlying parameter space is too large.

# Problem

- Let $\Theta_n, n = 1, \ldots$ be a sieve. ($\Theta_n \subset \Theta$, $\Theta_n \nearrow \Theta$)
- Assume for any $\theta \in \Theta$, there exist $\pi_n\theta \in \Theta_n$ such that, for an appropriate pseudo-distance $\rho$, $\rho(\pi_n\theta, \theta) \to 0$ as $n \to \infty$.
- Let $\hat{\theta}_n$ be the sieve estimates, which is required to satisfy

$$L_n(\hat{\theta}_n) \geq \sup_{\theta \in \Theta_n} L_n(\theta) - \eta_n \tag{13}$$

- The authors want to know how fast $\hat{\theta}$ converges to $\theta^*$ in terms of $n$.
- For positive $\varepsilon_n$, the convergence rate is $O(\varepsilon_n)$ if $\rho(T_n, \theta_0)$ is $O_{\mathbb{P}}(\varepsilon_n)$ under $\mathbb{P}$. => For any $\delta > 0$, there exist $M > 0$ and $N \in \mathbb{N}$ such that

$$\forall n > N, \quad \mathbb{P}(\rho(T_n, \theta_0) \geq M\varepsilon_n) \leq \delta$$

# Main Results.

- Let $\rho(\cdot, \cdot)$ be a pseudo-distance on $\Theta$.
- The authors call the quantity $l(x, \theta_0) - l(x, \theta)$ the criterion difference at $\theta$.
- The authors focus on the sieve MLE, where $l(x, \theta) = \log f(x, \theta)$.
- All result apply to any estimate $\hat{\theta}_n$ maximizing $L_n(\theta)$ over $\Theta_n$, as long as $l(x, \theta)$ satisfies the regularity conditions.

# Main Results.

## Condition C1.

For some constants $A_1 > 0$ and $\alpha > 0$, and for all small $\varepsilon > 0$,

$$\inf_{\{\rho(\theta,\theta_0)\geq\varepsilon,\theta\in\Theta_n\}} \mathbb{E}\left[l\left(X,\theta_0\right) - l(X,\theta)\right] \geq 2A_1\varepsilon^{2\alpha}.$$

- In order to have convergence, the expected criterion difference should be zero at $\theta = \theta_0$ and positive otherwise.
- **Condition C1** simply specifies the rate of increase of the expected criterion difference as $\theta$ moves away from $\theta_0$. On the other hand, as $\theta \to \theta_0$, the criterion difference should approach zero.

# Main Results.

## Condition C2.

For some constants $A_2 > 0$ and $\beta > 0$, and for all small $\varepsilon > 0$,

$$\sup_{\{\rho(\theta,\theta_0)\leq\varepsilon,\theta\in\Theta_n\}} \mathrm{Var}\left(l\left(X,\theta_0\right) - l(X,\theta)\right) \leq A_2\varepsilon^{2\beta}.$$

- **Condition C2** basically controls the rate of decay of its variance as $\theta$ approaches $\theta_0$.

# Main Results.

## Condition C3.

Let

$$\mathcal{F}_n = \{l(\cdot, \theta) - l(\cdot, \pi_n \theta_0) : \theta \in \Theta_n\}.$$

For some constants $r_0 < \frac{1}{2}$ and $A_3 > 0$,

$$H(\varepsilon, \mathcal{F}_n) \leq A_3 n^{2r_0} \varepsilon^{-r} \quad \text{for all small } \varepsilon > 0,$$

where $H(\varepsilon, \mathcal{F}_n)$ is the $L_\infty$-metric entropy of the space $\mathcal{F}_n$, that is, $\exp(H(\varepsilon, \mathcal{F}_n))$ is the number of $\varepsilon$-balls in the $L_\infty$-metric needed to cover the space $\mathcal{F}_n$.

# Main Results.

- **Condition C3** controls the size of $\mathcal{F}_n$ induced by $\theta \in \Theta_n$, that is, it controls the effective size of the approximating space $\Theta_n$.
- In the case when unrestricted maximization can be used, one has $r_0 = 0$, and $r$ depends on the characteristics of the function class.
- $n^{2r_0}$ disappears because it is done at $\Theta$ without maximizing at $\Theta_n$.
- Thus, in the case of sieve estimation, the sieve approximation error $\rho(\pi_n \theta_0, \theta_0)$ decreases in $n$ and the entropy count increases in $n$ as $n^{r_0}$.

# Main Results.

## Theorem (Main Theorem)

*Suppose* **Conditions C1** *and* **C3** *hold and* $\hat{\theta}_n$ *satisfies (13) with* $\eta_n = o\left(n^{-\omega}\right)$, *where*

$$\omega = \begin{cases} \frac{2(1-2r_0)}{2} - \frac{\log\log n}{\alpha \log n}, & \text{if } r = 0^+, \\ \frac{2(1-2r_0)}{2+r}, & \text{if } 0 < r < 2, \\ \frac{1-2r_0}{2} - \frac{\log\log n}{\log n}, & \text{if } r = 2, \\ \frac{1-2r_0}{r}, & \text{if } r > 2. \end{cases}$$

$\varepsilon^{-0^+}$ *is understood to represent* $\log(1/\varepsilon)$.

# Main Results.

## Theorem (Main Theorem)

*In addition,* **Condition C2** *is also supposed to hold for the case of* $0^+ \leq r < 2$. *Then,*

$$\rho(\hat{\theta}_n, \theta_0) = O_{\mathbb{P}}\left(\max\left(n^{-\tau}, \rho\left(\pi_n\theta_0, \theta_0\right), K^{1/2\alpha}\left(\pi_n\theta_0, \theta_0\right)\right)\right),$$

*where* $K\left(\pi_n\theta_0, \theta_0\right) = \mathbb{E}\left(l\left(\theta_0, Y\right) - l\left(\pi_n\theta_0, Y\right)\right)$ *and*

$$\tau = \begin{cases} \frac{1-2r_0}{2\alpha} - \frac{\log\log n}{2\alpha\log n}, & \text{if } r = 0^+, \beta \geq \alpha, \\ \frac{1-2r_0}{4\alpha-2\beta}, & \text{if } r = 0^+, \beta < \alpha, \\ \frac{1-2r_0}{4\alpha-\min(\alpha,\beta)(2-r)}, & \text{if } 0 < r < 2, \\ \frac{1-2r_0}{4\alpha} - \frac{\log\log n}{2\alpha\log n}, & \text{if } r = 2, \\ \frac{1-2r_0}{2\alpha r}, & \text{if } r > 2. \end{cases}$$

# Main Results.

1. In the case of MLE, the expected criterion difference $\mathbb{E}_{\theta_0}\left(l\left(\theta_0, Y\right) - l(\theta, Y\right)$ reduces to the KL pseudo-distance,

$$K\left(\theta, \theta_0\right) = \mathbb{E}_{\theta_0} \log\left(p\left(\theta_0, Y\right)/p(\theta, Y)\right).$$

2. If we choose $\rho\left(\theta, \theta_0\right)$ to be the Hellinger distance, then **Condition C1** holds with $\alpha = 1$.
3. If $K(\pi_n\theta_0, \theta_0) = O(n^{-2\alpha\tau})$, it will not enter the rate calculation.
4. The extra $\log n$ factor in Theorem 1 for the case of $r = 0^+$ can be removed if an extra continuity assumption is made on the criterion difference.

# Peeling : Main technique of proof.

- The authors obtain bounds for the tail probability for a decreasing sequence of rates $\varepsilon_n^{(k)}$ where $k = 1, 2, \ldots$.
- **Strategy**: A recursive probability decomposition based on the division of the parameter space into $B_n^{(k)}$

$$\begin{aligned} \mathbb{P}\left(\rho(\hat{\theta}_n, \theta_0) \geq D\epsilon_n^{(k)}\right) &= \mathbb{P}\left(\{\rho(\hat{\theta}_n, \theta_0) \geq D\epsilon_n^{(k-1)}\} \cup B_n^{(k)}\right) \\ &\leq \mathbb{P}\left(\rho(\hat{\theta}_n, \theta_0) \geq D\epsilon_n^{(k-1)}\right) + \mathbb{P}\left(B_n^{(k)}\right) \end{aligned} \tag{14}$$

  where $B_n^{(k)} = \{D\varepsilon_n^{(k)} \leq \rho(\hat{\theta}_n, \theta_0) < D\varepsilon_n^{(k-1)}\}$ for $k = 1, \ldots$.
- $\mathbb{P}\left(\rho(\hat{\theta}_n, \theta_0) \geq D\epsilon_n^{(1)}\right)$ is bounded by **Lemma (2)**.
- Assume $D > 1$ and the authors only prove the case of $4\alpha \geq \frac{\beta(2-r)}{2}$.

# Used lemmas.

## Lemma (2)

*Suppose* **Conditions C1** *and* **C2** *hold. Assume also that* **Condition C3** *holds if* $0^+ \leq r < 2$.

*Let* $\varepsilon_n^{(1)} = n^{-\min(\alpha_1,(1-2r_0)/[\alpha(r+2)])}$, *where* $\alpha_1 = (1-2r_0)/(4\alpha)$. *Then there exists an* $M_1 > 0$ *such that, for any* $D > 0$, *we have*

$$\mathbb{P}\left(\rho(\hat{\theta}_n, \theta_0) \geq D\varepsilon_n^{(1)}\right) \leq 5\exp\left(-(1-\varepsilon)\max\left(D^{4\alpha}, D^{2\alpha}\right) M_1 n^{2r_0}\right).$$

# Used lemmas.

## Lemma (3)

*Suppose* **Conditions C1** *and* **C2** *hold. Assume also that* **Conditior C3** *holds if* $0^+ \leq r < 2$. *If at Step* $k-1$ *we have a rate*

$$\varepsilon_n^{(k-1)} > \max\left(n^{-(1-2r_0)/[\alpha(r+2)]}, \rho\left(\pi_n\theta_0, \theta_0\right), K^{1/2\alpha}\left(\pi_n\theta_0, \theta_0\right)\right),$$

*so that*

$$\mathbb{P}\left(\rho(\hat{\theta}_n, \theta_0) \geq D\varepsilon_n^{(k-1)}\right) \leq 5G_n.$$

- $G_n = \exp\left(-(1-\varepsilon)\max\left(D^{4\alpha}, D^{2\alpha}\right)M_1 n^{2r_0}\right) + (k-1)\exp\left(-Ln^{\delta_0}\right)$.
- $\delta_0 = \min\left(\frac{r+4r_0}{r+2}, \frac{\beta r(1-2r_0)}{4\alpha} + r_0\right)$.
- $L = (1-\varepsilon)\min\left(M_2 D^{2\alpha}, M_3 D^{4\alpha-\beta(2-r)/2}\right)$.

# Used lemmas.

## Lemma (3)

*Then at Step $k$, we can find an improved rate*

$$\mathbb{P}\left(\rho\left(\hat{\theta}_n, \theta_0\right) \geq D\varepsilon_n^{(k)}\right) \leq$$
$$5\left[\exp\left(-(1-\varepsilon)\max\left(D^{4\alpha}, D^{2\alpha}\right) M_1 n^{2r_0}\right) + k\exp\left(-L_n^{\delta_0}\right)\right].$$

- $\varepsilon_n^{(k)} = \max\left(n^{-\alpha_k}, n^{-(1-2r_0)/[\alpha(r+2)]}, \rho\left(\pi_n\theta_0, \theta_0\right), K^{1/2\alpha}\left(\pi_n\theta_0, \theta_0\right)\right)$.
- $\alpha_k = \frac{1-2r_0}{4\alpha} + \frac{\alpha_{k-1}\beta(2-r)}{4\alpha} \to \frac{1-2r_0}{4\alpha-\beta(2-r)}$ as $k \to \infty$ if $\beta < \alpha$.

# Key Tool: Lemma 1

## Lemma (1)

*If*

$$H\left(\varepsilon, \mathcal{F}_n\right) \leq A_0 n^{2r_0} \varepsilon^{-r},$$

*for* $\varepsilon \in (0, a]$, *where* $a$ *is a small positive number. Then,*

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}_n} \nu_n(f) > M\right) \leq 5 \exp\left(-(1-\epsilon)\psi_1(M, v, \mathcal{F}_n)\right)$$

- **Notation and condition in Lemma (1)**:
  - $\nu_n(f) = n^{-1/2} \sum_{i=1}^n (f(Y_i) - \mathbb{E}f(Y_i))$.
  - $M$: The threshold that the process must exceed, whose condition varies depending on the metric entropy index $r \propto$ **Condition C3**.
  - $v \geq \sup_{\mathcal{F}_n} Var(f(Y))$
  - $\psi_1(M, v, \mathcal{F}_n) = M^2 / \left[2v\left(1 + M/3n^{1/2}v\right)\right]$
  - $\mathcal{F}_n$: The class of functions, whose complexity is controlled by metric entropy.

# Applying lemma 1 to lemma 3

- Main part of whole procedure:

$$\mathbb{P}(B_n^{(k)}) \leq \mathbb{P}\left(\sup_{f \in \mathcal{F}_n} \nu_n(f) > M\right)$$

Table: Matching notation between Lemma 1 and Lemma 3

| **In Lemma 1** | **In Lemma 3** |
|---|---|
| $f(Y)$ | $l(\theta, Y) - l(\pi_n\theta_0, Y)$ |
| $\nu_n(f)$ | $n^{1/2}\left(L_n(\theta) - L_n(\pi_n\theta_0) - \mathbb{E}[L_n(\theta) - L_n(\pi_n\theta_0)]\right)$ |
| $M$ | $A_1 n^{1/2}(D\epsilon_n^{(k)})^{2\alpha} \propto$ **Condition C1** |
| $\sup_{f \in \mathcal{F}_n} \mathrm{Var}(f(Y))$ | $\sup_{\theta \in B_n^{(k)}} \mathrm{Var}\left(l(\theta, Y) - l(\pi_n\theta_0, Y)\right) = v_k$ |
| $v$ | $4A_2(D\epsilon_n^{(k-1)})^{2\beta} \propto$ **Condition C2** |

# Applying lemma 1 to lemma 3

- To use **Lemma 1** on $\mathbb{P}\left(B_n^{(k)}\right)$, we must show that the following inequality holds:

$$\mathbb{P}(B_n^{(k)}) \leq \mathbb{P}\left(\sup_{\theta \in B_n^{(k)}} (L_n(\theta) - L_n(\pi_n\theta_0)) \geq -\eta_n\right)$$

- By the definition of the Sieve estimate, $\hat{\theta}_n$ satisfies:

$$L_n(\hat{\theta}_n) \geq \sup_{\theta \in \Theta_n} L_n(\theta) - \eta_n$$

- Since $\pi_n\theta_0 \in \Theta_n$, the following holds:

$$L_n(\hat{\theta}_n) - L_n(\pi_n\theta_0) \geq -\eta_n$$

- If $\hat{\theta}_n \in B_n^{(k)}$, $\sup_{\theta \in B_n^{(k)}} L_n(\theta) \geq L_n(\hat{\theta}_n)$,

$$\sup_{\theta \in B_n^{(k)}} (L_n(\theta) - L_n(\pi_n\theta_0)) \geq L_n(\hat{\theta}_n) - L_n(\pi_n\theta_0) \geq -\eta_n$$

# Matching notation for Applying lemma 1 to lemma 3

- Recall
  - $f(Y_i) = l(\theta, Y_i) - l(\pi_n\theta_0, Y_i)$.
  - $\nu_n(f) \longleftrightarrow n^{1/2}\left(L_n(\theta) - L_n(\pi_n\theta_0) - \mathbb{E}[L_n(\theta) - L_n(\pi_n\theta_0)]\right)$.

$$
\begin{aligned}
L_n(\theta) - L_n(\pi_n\theta_0) &= \frac{1}{n}\sum_{i=1}^{n} l(\theta, Y_i) - \frac{1}{n}\sum_{i=1}^{n} l(\pi_n\theta_0, Y_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} \left(l(\theta, Y_i) - l(\pi_n\theta_0, Y_i)\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} \left(f(Y_i) \pm \mathbb{E}[f(Y_i)]\right) \\
&= \frac{1}{\sqrt{n}}\nu_n(f) + \mathbb{E}[l(\theta, Y) - l(\pi_n\theta_0, Y)]
\end{aligned}
$$

# Applying condition C1 to lemma 3

- Recall

$$\sup_{\theta \in B_n^{(k)}} \left( \frac{1}{\sqrt{n}} \nu_n(f) + \mathbb{E}[l(\theta, Y) - l(\pi_n \theta_0, Y)] \right) \geq -\eta_n \tag{15}$$

- And

$$\sup_{\theta \in B_n^{(k)}} \left( \frac{1}{\sqrt{n}} \nu_n(f) \right) \geq \inf_{\theta \in B_n^{(k)}} (\mathbb{E}[l(\pi_n \theta_0, Y) - l(\theta, Y)]) - \eta_n \tag{16}$$

- Recall

$$\mathbb{P}(B_n^{(k)}) \leq \mathbb{P} \left( \sup_{\theta \in B_n^{(k)}} (L_n(\theta) - L_n(\pi_n \theta_0)) \geq -\eta_n \right) \tag{17}$$

$$= \mathbb{P} \left( \sup_{\theta \in B_n^{(k)}} \left( \frac{1}{\sqrt{n}} \nu_n(f) \right) \geq \inf_{\theta \in B_n^{(k)}} (\mathbb{E}[l(\pi_n \theta_0, Y) - l(\theta, Y)]) - \eta_n \right) \tag{18}$$

# Applying condition C1 to lemma 3

- To use **Condition C1**

$$\mathbb{E}[l(\pi_n\theta_0, Y) - l(\theta, Y)] = \mathbb{E}[l(\theta_0, Y) - l(\theta, Y)] - \mathbb{E}[l(\theta_0, Y) - l(\pi_n\theta_0, Y)]$$

- The second term represents the **Sieve Approximation Error**, denoted as $K(\pi_n\theta_0, \theta_0)$.
- By applying **Condition C1** on $B_n^{(k)}$, the first term is bounded below:

$$\inf_{\theta \in B_n^{(k)}} \mathbb{E}[l(\theta_0, Y) - l(\theta, Y)] - K(\pi_n\theta_0, \theta_0) \geq 2A_1(D\epsilon_n^{(k)})^{2\alpha} - K(\pi_n\theta_0, \theta_0)$$

# Applying lemma 1 to lemma 3

- Thus, allocating half of the coefficient $2A_1$ easily absorbs these errors:

$$2A_1(D\epsilon_n^{(k)})^{2\alpha} - K(\pi_n\theta_0, \theta_0) - \eta_n \geq A_1(D\epsilon_n^{(k)})^{2\alpha}$$

- **Final Result**: Applying Lemma 1 with the entropy bound and variance bound $v_k$ yields:

$$\begin{aligned}
\mathbb{P}(B_n^{(k)}) &\leq \mathbb{P}\left(\frac{1}{\sqrt{n}} \sup_{\theta \in B_n^{(k)}} \nu_n(f) \geq A_1(D\epsilon_n^{(k)})^{2\alpha}\right) \\
&= \mathbb{P}\left(\sup_{\theta \in B_n^{(k)}} \nu_n(f) \geq \sqrt{n}A_1(D\epsilon_n^{(k)})^{2\alpha}\right) \\
&\leq 5\exp\left(-(1-\epsilon)\psi_1\left(A_1 n^{1/2}(D\epsilon_n^{(k)})^{2\alpha}, 4A_2(D\epsilon_n^{(k-1)})^{2\beta}, \mathcal{F}_n\right)\right)
\end{aligned}$$

# Outline

# Bernstein-von Mises Theorem

- Note that the posterior distribution is given as

$$\pi(\theta|\text{Data}) \propto \prod_{i=1}^{n} f(X_i|\theta)\pi(\theta).$$

- Let $\hat{\theta}$ be the MLE and rewrite the posterior as

$$\pi(\theta|\text{Data}) \propto \exp\left(\ell_n(\theta) - \ell_n(\hat{\theta})\right)\pi(\theta),$$

  where $\ell_n(\theta) = \sum_{i=1}^{n} \log f(X_i|\theta)$.
- Taylor expansion yields

$$\ell_n(\theta) - \ell_n(\hat{\theta}) \approx s_n(\hat{\theta})^\top(\theta - \hat{\theta}) + (\theta - \hat{\theta})^\top[\partial s_n(\theta)/\partial\theta]_{\theta=\hat{\theta}}(\theta - \hat{\theta}).$$

# Bernstein-von Mises Theorem

- Since $s_n(\hat{\theta}) = 0$, we have

$$\pi(\theta|\text{Data}) \propto \exp\left((\theta - \hat{\theta})^\top [\partial s_n(\theta)/\partial\theta]_{\theta=\hat{\theta}}(\theta - \hat{\theta})\right)\pi(\theta).$$

- Thus, we can say that

$$\sqrt{n}(\theta - \hat{\theta})|\text{Data} \xrightarrow{d} \mathcal{N}(0, \mathbb{I}^{-1}(\theta^*))$$

in probability.

# Outline

# Preliminaries

- $d(P, Q)$ : Hellinger distance for distribution $P$ and $Q$
- 데이터 $X_1, ..., X_n \sim P_0$
- $\mathcal{P}$ : 고려하는 분포들의 집합 (Model space)
- $p$ : density for $P \in \mathcal{P}$
- $\Pi_n(\cdot)$ : $\mathcal{P}$ 위에서 정의된 prior distribution
- Posterior distribution은 다음과 같이 정의됨

$$\Pi_n(B|X_1, ..., X_n) = \frac{\int_B \prod_{i=1}^n p(X_i) d\Pi_n(P)}{\int \prod_{i=1}^n p(X_i) d\Pi_n(P)}$$

# Main Theorem

## Theorem (Theorem 2.1 from Ghosal (2000))

*Suppose that for a sequence* $\epsilon_n$ *with* $\epsilon_n \to 0$ *and* $n\epsilon_n^2 \to \infty$, *a constant* $C > 0$ *and sets* $\mathcal{P}_n \subseteq \mathcal{P}$, *we have*

$$\log N(\epsilon_n, \mathcal{P}_n, d) \leq n\epsilon_n^2 \tag{19}$$

$$\Pi_n\left(P : \mathbb{E}_{P_0}\left[\log \frac{p_0(X)}{p(X)}\right] \leq \epsilon_n^2, \mathbb{E}_{P_0}\left[\log \frac{p_0(X)}{p(X)}\right]^2 \leq \epsilon_n^2\right) \geq \exp(-n\epsilon_n^2 C) \tag{20}$$

$$\Pi_n(\mathcal{P} \backslash \mathcal{P}_n) \leq \exp(-n\epsilon_n^2(C+4)) \tag{21}$$

*Then for sufficiently large* $M$, *we have that* $\Pi_n(P : d(P, P_0) \geq M\epsilon_n | X_1, ..., X_n) \to 0$ *in* $P_0^n$*-probability*

# Proof of Main Theorem

다음과 같은 2가지의 Step을 보임으로써 Main Theorem을 증명하고자 한다.

1. $\mathbb{E}_{P_0^n}[\Pi_n(\mathcal{P}\backslash\mathcal{P}_n|X_1,...,X_n)] \to 0$
2. $\mathbb{E}_{P_0^n}[\Pi_n(P \in \mathcal{P}_n : d(P,P_0) \geq M\epsilon_n|X_1,...,X_n)] \to 0$

# Proof of Main Theorem - Step 1

$$B_n = \left\{ P : \mathbb{E}_{P_0}\left[\log \frac{p_0(X)}{p(X)}\right] \leq \epsilon_n^2, \mathbb{E}_{P_0}\left[\log \frac{p_0(X)}{p(X)}\right]^2 \leq \epsilon_n^2 \right\} \tag{22}$$

$$A_n = \left\{ X_1, ..., X_n : \int \prod_{i=1}^{n} \frac{p(X_i)}{p_0(X_i)} d\Pi_n(P) \geq \exp(-2n\epsilon_n^2)\Pi_n(B_n) \right\} \tag{23}$$

이라고 할때, Lemma 8.1(Ghosal 2000) 에 의하여 $P_0^n(A_n) \to 1$ 이 성립함.

해석 : 진짜($P_0$)와 비슷한 모델들이 충분히 많다면 ($\Pi_n(B_n) \uparrow$), Likelihood ratio는 일정 하한보다 더 작아지지 않는다.

# Proof of Main Theorem - Step 1

따라서,

$$\mathbb{E}_{P_0^n}[\Pi_n(\mathcal{P}\backslash\mathcal{P}_n|X_1,...,X_n)] \tag{24}$$

$$\leq \mathbb{E}_{P_0^n}[\Pi_n(\mathcal{P}\backslash\mathcal{P}_n|X_1,...,X_n)\mathbb{I}(A_n)] + P_0^n(A_n^c) \tag{25}$$

$$= \mathbb{E}_{P_0^n}\left[\frac{\int_{\mathcal{P}\backslash\mathcal{P}_n}\prod_{i=1}^n p(X_i)d\Pi_n(P)}{\int\prod_{i=1}^n p(X_i)d\Pi_n(P)}\mathbb{I}(A_n)\right] + P_0^n(A_n^c) \tag{26}$$

$$= \mathbb{E}_{P_0^n}\left[\frac{\int_{\mathcal{P}\backslash\mathcal{P}_n}\prod_{i=1}^n (p(X_i)/p_0(X_i))d\Pi_n(P)}{\int\prod_{i=1}^n (p(X_i)/p_0(X_i))d\Pi_n(P)}\mathbb{I}(A_n)\right] + P_0^n(A_n^c) \tag{27}$$

$$\leq \mathbb{E}_{P_0^n}\left[\int_{\mathcal{P}\backslash\mathcal{P}_n}\prod_{i=1}^n (p(X_i)/p_0(X_i))d\Pi_n(P)\right]\exp(2n\epsilon_n^2)\frac{1}{\Pi_n(B_n)} + P_0^n(A_n^c) \tag{28}$$

$$= \Pi_n(\mathcal{P}\backslash\mathcal{P}_n)\exp(2n\epsilon_n^2)\frac{1}{\Pi_n(B_n)} + P_0^n(A_n^c) \to 0 \tag{29}$$

# Step 1의 결론

즉, Seive $\mathcal{P}_n$ 밖에는 posterior mass가 거의 남아있지 않다. 따라서, 우리는 Sieve에서의 posterior mass만을 조사하면 된다.

# Proof of Main Theorem - Step 2

## Lemma (Theorem 7.1 in Ghosal (2000))

*Assume that condition (19) holds, i.e.,*

$$\log N(\epsilon_n, \mathcal{P}_n, d) \leq n\epsilon_n^2.$$

*Then, there exist tests $\phi_n$ such that*

$$\mathbb{E}_{P_0^n}\phi_n \leq 2\exp(-Kn\epsilon_n^2) \tag{30}$$

$$\sup_{P\in\mathcal{P}_n : d(P,P_0)>M\epsilon_n} \mathbb{E}_{P^n}(1-\phi_n) \leq \exp(-KnM^2\epsilon_n^2), \tag{31}$$

*where $K > 0$ is a constant.*

해석 : Sieve에 있는 분포들이 적당히 적다보니, 제1종오류 및 제2종오류가 잘 control되는 test function이 존재한다.

# Proof of Main Theorem - Step 2 : Background

## Local Testing (Birgé's Lemma)

For the true distribution $P_0$ and an alternative $P_1$ with distance $d(P_0, P_1) \geq \epsilon_n$, there exists a test function $\phi$ such that:

$$E_{P_0}[\phi] \leq \exp(-Cn\epsilon_n^2), \quad E_{P_1}[1-\phi] \leq \exp(-Cn\epsilon_n^2)$$

## Global Testing

To test $P_0$ against a vast alternative space $\mathcal{P}_n = \{P : d(P, P_0) \geq \epsilon_n\}$ covered by $N$ balls, we combine local tests (Union Bound).
There exists a global test $\Phi$:

- **Global Type I Error:** $E_{P_0}[\Phi] \leq N \exp(-Cn\epsilon_n^2)$
- **Global Type II Error:** $\sup_{P \in \mathcal{P}_n} E_P[1-\Phi] \leq \exp(-Cn\epsilon_n^2)$

# Proof of Main Theorem - Step 2 : Overview

## The Birth of the Entropy Condition

For the global test to succeed, the Type I error must vanish as $n \to \infty$. The exponential decay must strictly dominate the model complexity $N$:

$$N \exp(-Cn\epsilon_n^2) \to 0 \implies \log N \leq Cn\epsilon_n^2$$

$\Rightarrow$ *This is exactly Ghosal's Condition (19)!*

# Proof of Main Theorem - Step 2

앞의 Lemma에 의하여, test function을 얻을 수 있고, 이를 이용하면 다음과 같다.

$$\mathbb{E}_{P_0}\left[\Pi_n(P \in \mathcal{P}_n : d(P, P_0) \geq M\epsilon_n | X_1, ..., X_n)\right] \tag{32}$$

$$= \mathbb{E}_{P_0}\left[\Pi_n(P \in \mathcal{P}_n : d(P, P_0) \geq M\epsilon_n | X_1, ..., X_n)\phi_n\right] \tag{33}$$

$$+ \mathbb{E}_{P_0}\left[\Pi_n(P \in \mathcal{P}_n : d(P, P_0) \geq M\epsilon_n | X_1, ..., X_n)(1 - \phi_n)\right] \tag{34}$$

식 (33) 은 Test function의 제1종오류가 지수적으로 작기 때문에 무시해도 된다.

# Proof of Main Theorem - Step 2

**Test function을 사용하는 이유**

식 (33) 의 경우 : Test function이 $P_0$에 나오질 않았다고 판단하는 경우, Type 1 error 확률로 제어

식 (34) 의 경우 : Test function이 $P_0$에서 나왔다고 판단한 경우, Type 2 error 확률로 제어

# Proof of Main Theorem - Step 2

식 (34) 의 upper bound는 다음과 같다.

$$\mathbb{E}_{P_0^n}\left[\Pi_n(P \in \mathcal{P}_n : d(P, P_0) \geq M\epsilon_n | X_1, ..., X_n)(1 - \phi_n)\right] \tag{35}$$

$$\leq \mathbb{E}_{P_0^n}\left[\frac{\int_{P \in \mathcal{P}_n : d(P,P_0) \geq M\epsilon_n} \prod_{i=1}^n \frac{p(X_i)}{p_0(X_i)} d\Pi_n(P)}{\int \prod_{i=1}^n \frac{p(X_i)}{p_0(X_i)} d\Pi_n(P)}(1 - \phi_n)\mathbb{I}(A_n)\right] + P_0^n(A_n^c) \tag{36}$$

# Proof of Main Theorem - Step 2

식 (36) 에서 첫번째 term의 upper bound는 다음과 같다.

$$\mathbb{E}_{P_0^n}\left[\frac{\int_{P\in\mathcal{P}_n:d(P,P_0)\geq M\epsilon_n}\prod_{i=1}^n\frac{p(X_i)}{p_0(X_i)}d\Pi_n(P)}{\int\prod_{i=1}^n\frac{p(X_i)}{p_0(X_i)}d\Pi_n(P)}(1-\phi_n)\mathbb{I}(A_n)\right] \tag{37}$$

$$\leq\int_{P\in\mathcal{P}_n:d(P,P_0)\geq M\epsilon_n}\mathbb{E}_{P_0^n}\left[\prod_{i=1}^n\frac{p(X_i)}{p_0(X_i)}(1-\phi_n)\mathbb{I}(A_n)\right]d\Pi_n(P)\frac{\exp(2n\epsilon_n^2)}{\Pi_n(B_n)} \tag{38}$$

$$\leq\int_{P\in\mathcal{P}_n:d(P,P_0)\geq M\epsilon_n}\mathbb{E}_{P^n}[1-\phi_n]d\Pi_n(P)\frac{\exp(2n\epsilon_n^2)}{\Pi_n(B_n)} \tag{39}$$

$$\leq\sup_{P\in\mathcal{P}_n:d(P,P_0)\geq M\epsilon_n}\mathbb{E}_{P^n}[1-\phi_n]\frac{\exp(2n\epsilon_n^2)}{\Pi_n(B_n)}\to 0 \tag{40}$$

□

# Outline

# Paper

Asymptotic Behavior of Bayes Estimates under Possibly Incorrect Models. Bunke and Milhaud. 1998.

# Preliminaries

**Notation**

- $X_1, \cdots, X_n \overset{i.i.d}{\sim} G$ and $P_\theta$, $\theta \in \Theta$ is a misspecified model.
- $g$ and $p_\theta$ are densities of each $G$ and $P_\theta$.
- $K(\theta) := \mathrm{KL}(g, p_\theta)$ and $\Theta_G = \{\theta_G \in \Theta | \theta_G = \mathrm{argmin}_{\theta \in \Theta} K(\theta)\}$.
- Let $\theta \sim \pi(\theta)$ be the prior distribution and $\pi(\theta|X_1, \cdots, X_n) \propto \pi(\theta) \prod_{i=1}^n p_\theta(X_i)$ be the posterior distribution under the misspecified model.
- For a loss function $L : \Theta \times \Theta \to [0, \infty)$,

$$\hat{\theta}_n = \mathrm{argmin}_{t \in \theta} \mathbb{E}_{\theta | X_{1:n}} [L(t, \theta) | X_1, \cdots, X_n] \tag{41}$$

  is called a pseudo-Bayes estimator.

**Goal:** Strong consistency and asymptotic normality of the pseudo-Bayes estimator.

# Conditions

## Assumption

A1. $\Theta$ is a closed convex set in $\mathbb{R}^d$ with a nonempty interior, the density $p_\theta(x)$ is bounded over $\Theta \times \mathbb{R}^k$ and its support is the same for all $\theta \in \Theta$.
A2. For all $\theta \in \Theta$, there is a sphere $S[\theta, \eta(\theta)]$ such that

$$E \sup \left\{ \left| \log \left( \frac{g(X)}{p_t(X)} \right) \right| ; t \in S[\theta, \eta(\theta)] \right\} < \infty. \tag{42}$$

A3. For all fixed $x \in \mathbb{R}^k$, the density $p_\theta(x)$ has a continuous derivative $p'_\theta(x)$ w.r.t. $\theta$ and there are positive constants $c, b_0$ with

$$\int ||[p_\theta(x)]^{-1} p'_\theta(x)||^{4(d+1)} p_\theta(x) dx < c(1 + ||\theta||^{b_0}) \tag{43}$$

for all $\theta \in \Theta$.

# Theorem

## Assumption (Continued)

A4. For some positive constant $b_1$,

$$Q(\theta) = \int [p_\theta(x)g(x)]^{1/2}dx < c||\theta||^{-b_1}, \quad \theta \in \Theta. \tag{44}$$

A5. There are positive constants $b_2, b_3$ so that for all $\theta \in \Theta$ and $r > 0$ it holds that

$$0 < \pi(S(\theta, r)) \leq cr^{b_2}(1 + (||\theta|| + r)^{b_3}). \tag{45}$$

## Theorem

*Under the assumption A1-A5, followings hold for all* $p > 0$.
$\mathbb{E}_{\theta|X_{1:n}}(d_G^p(\theta)|X_1, \cdots, X_n) \overset{a.s.}{\to} 0$ *where* $d_G(\theta) = \min\{||\theta - t|| \mid t \in \Theta_G\}$.

# Proof of Theorem

Let $Z_n(\theta) := \Pi_{i=1}^n p_\theta(X_i)/g(X_i)$ and $\epsilon, \delta, \eta > 0$ be given. Then, we can get the upper bound of $\mathbb{E}_{\theta|X_{1:n}}(d_G^p(\theta)|X_1, \cdots, X_n)$.

$$\begin{aligned}
&\mathbb{E}_{\theta|X_{1:n}}(d_G^p(\theta)|X_1, \cdots, X_n) = \int d_G^p(\theta)\pi(\theta|X_1, \cdots, X_n)d\theta \\
&= \int d_G^p(\theta)\frac{\Pi_{i=1}^n p_\theta(X_i)}{\int \Pi_{i=1}^n p_\theta(X_i)\pi(\theta)d\theta}\pi(\theta)d\theta = \int d_G^p(\theta)\frac{Z_n(\theta)}{\int Z_n(\theta)\pi(\theta)d\theta}\pi(\theta)d\theta \\
&= \frac{\mathbb{E}_\theta\left[d_G^p(\theta)Z_n(\theta)\right]}{\mathbb{E}_\theta[Z_n(\theta)]} \leq \epsilon^p + \frac{\mathbb{E}_\theta\left[d_G^p(\theta)Z_n(\theta)I(d_G^p(\theta) \geq \epsilon)\right]}{\mathbb{E}_\theta[Z_n(\theta)]} \\
&= \epsilon^p + \frac{\mathbb{E}_\theta\left[||\theta||^p Z_n(\theta)I(d_G^p(\theta) \geq \epsilon, ||\theta|| \leq \delta)\right] + \mathbb{E}_\theta\left[||\theta||^p Z_n(\theta)I(||\theta|| > \delta)\right]}{\mathbb{E}_\theta[I(||\theta|| < \eta)Z_n(\theta)]}
\end{aligned} \tag{46}$$

# Proof of Theorem (Continued)

To bound the $d_G^p(\theta)$, authors suppose $0 \in \Theta_G$ without loss of generality. From this,

$$d_G(\theta) = \min\{||\theta - t|| \mid t \in \Theta_G\} \leq ||\theta - 0|| = ||\theta||. \tag{47}$$

And then, let

$$\begin{aligned} A_n &= \mathbb{E}_\theta \left[ ||\theta||^p Z_n(\theta) I(d_G^p(\theta) \geq \epsilon, ||\theta|| \leq \delta) \right] \\ B_n &= \mathbb{E}_\theta \left[ ||\theta||^p Z_n(\theta) I(||\theta|| > \delta) \right] \\ C_n &= \mathbb{E}_\theta [I(||\theta|| < \eta) Z_n(\theta)] \end{aligned} \tag{48}$$

Finally, need to show that the followings hold for some $\alpha > 0$

1. $\exp\left(n(K(0) + 2\alpha)\right) A_n \to 0$ a.s.
2. $\exp\left(n(K(0) + \alpha)\right) B_n \to 0$ a.s.
3. $\exp\left(n(K(0) + \frac{\alpha}{2})\right) C_n \to \infty$ a.s.

# Theorem

## Assumption (Additional)

A6. Let $L : \Theta \times \Theta \to [0, \infty)$ be a measurable loss function with $L(\theta, \theta) = 0$, $c_1$, $c_2$, $c_3$, $b_4$, $b_5$ be positive constants with

$$(c_1 ||t - \theta||^{b_4}) \wedge c_2 \leq L(t, \theta) \leq c_3 ||t - \theta||^{b_5} \tag{49}$$

for all $t, \theta \in \Theta$.

## Theorem

*If $\theta_G$ is unique, it holds under the assumption A1-A6 that for all pssudo-Bayes estimators $\hat{\theta}_n$ w.r.t. the loss function L,*

$$\hat{\theta}_n \overset{a.s.}{\to} \theta_G. \tag{50}$$

# Asymptotic Normality

- Under some regularity conditions regarding the $l(x,\theta) := \log \frac{p_\theta(x)}{g(x)}$ and the loss function $L$, the author shows the asymptotic normality of the pseudo-Bayes estimators.

## Theorem

*If $\theta_G$ is unique, the pseudo-Bayes estimator $\hat{\theta}_n$ is asymptotically normal.*

$$\sqrt{n}(\hat{\theta}_n - \theta_G) \xrightarrow{d} N(0, \Sigma) \tag{51}$$

*where,* $\Sigma = L_2^{-1} L_1 M^{-1} I_G M^{-1} (L_2^{-1} L_1)^\top$, $I_G := E\left[l'(x,\theta_G) l'(x,\theta_G)^\top\right]$, $M := -El''(x,\theta_G)$, $L_1 = L^{(1,1)}(\theta_G,\theta_G)$ *and* $L_2 = L^{(2,0)}(\theta_G,\theta_G)$.
*Note that* $l'(x,\theta_G) = \frac{\partial}{\partial\theta} l(x,\theta_G)$ *and* $l''(x,\theta_G) = \frac{\partial^2}{\partial\theta^2} l(x,\theta_G)$.

# Proof

Let $r_n(t) = \mathbb{E}_{\theta|X_{1:n}}[L(t,\theta)|X_1,\cdots,X_n]$. From the definition of $\hat{\theta}_n$, $r_n'(\hat{\theta}_n) = 0$. Then, do talyor expansion $r_n'(t)$ at $t = \theta_G$,

$$0 = r_n'(\hat{\theta}_n) \approx r_n'(\theta_G) + r_n''(\theta_G)(\hat{\theta}_n - \theta_G). \tag{52}$$

From this, we can know that $\sqrt{n}(\hat{\theta}_n - \theta_G) \approx -[r_n''(\theta_G)]^{-1}\sqrt{n}r_n'(\theta_G)$.
**First,** $r_n''(\theta_G) = \mathbb{E}_{\theta|X_{1:n}}\left[L^{(2,0)}(\theta_G,\theta)|X_1,\cdots,X_n\right] \overset{a.s.}{\to} L^{(2,0)}(\theta_G,\theta_G)$ by the previous theorem.
**Second,**

$$\begin{aligned}
\sqrt{n}r_n'(\theta_G) &= \sqrt{n}\mathbb{E}_{\theta|X_{1:n}}\left[L^{(1,0)}(\theta_G,\theta)|X_1,\cdots,X_n\right] \\
&\approx \sqrt{n}\mathbb{E}_{\theta|X_{1:n}}\left[L^{(1,0)}(\theta_G,\theta_G) + L^{(1,1)}(\theta_G,\theta_G)(\theta-\theta_G)|X_1,\cdots,X_n\right] \\
&= L^{(1,1)}(\theta_G,\theta_G)\mathbb{E}_{\theta|X_{1:n}}\left[\sqrt{n}(\theta-\theta_G)|X_1,\cdots,X_n\right].
\end{aligned} \tag{53}$$

# Proof (Continued)

Now then, let $t = \sqrt{n}(\theta - \theta_G)$ and obtain the distribution of $t|X_1, \cdots, X_n$.
From change of variable, we can know that

$$\pi(t|X_1, \cdots, X_n) \propto \pi\left(\theta_G + \frac{t}{\sqrt{n}}\right) \frac{Z_n\left(\theta_G + \frac{t}{\sqrt{n}}\right)}{Z_n(\theta_G)}. \tag{54}$$

So that,

$$\log \pi(t|X_1, \cdots, X_n) \approx \log Z_n\left(\theta_G + \frac{t}{\sqrt{n}}\right) - \log Z_n(\theta_G) + \text{Constant}. \tag{55}$$

Applying second order Taylor expansion to $Z_n(\theta)$,

$$\log Z_n\left(\theta_G + \frac{t}{\sqrt{n}}\right) - \log Z_n(\theta_G) \approx \frac{t^\top}{\sqrt{n}} \sum_i^n l'(X_i, \theta) - \frac{1}{2}\frac{t^\top}{\sqrt{n}}\left[\sum_{i=1}^n -l''(X_i, \theta)\right]\frac{t}{\sqrt{n}}. \tag{56}$$

# Proof (Continued)

Then,

$$\log \pi(t|X_1, \cdots, X_n) \approx t^\top S_n - \frac{1}{2} t^\top M_n t + \text{Constant}, \tag{57}$$

where $S_n = \frac{1}{\sqrt{n}} \sum_i^n l'(X_i, \theta)$ and $M_n = \frac{1}{n} \left[\sum_{i=1}^n -l''(X_i, \theta)\right]$.
From this formula, we can obtain $\pi(t|X_1, \cdots, X_n) \approx N(M_n^{-1} S_n, M_n^{-1})$. So that,

$$\mathbb{E}_{\theta|X_{1:n}} \left[\sqrt{n}(\theta - \theta_G)|X_1, \cdots, X_n\right] \approx M_n^{-1} S_n \xrightarrow{d} M^{-1} N(0, I_G). \tag{58}$$

By combine above results,

$$\sqrt{n}(\hat{\theta}_n - \theta_G) \xrightarrow{d} \left[L^{(2,0)}(\theta_G, \theta_G)\right]^{-1} L^{(1,1)}(\theta_G, \theta_G) M^{-1} N(0, I_G). \tag{59}$$

# Outline

# Paper

Misspecification in infinite-dimensional Bayesian statistics. Kleijn and van der Vaart. 2006.

# BMN

- model $\mathcal{P}$가 true probability measure $P_0$를 포함하지 않고 prior의 support $\subseteq \mathcal{P}$라면, posterior도 inconsistent할 것.
- 이 논문에서 보인 것 : 이런 misspecified 상황에서 posterior가 $P^* \in \mathcal{P}$ 근처로 집중될 충분조건들과 수렴속도
- $P^*$ is a point that minimizes the KL divergence of $P_0$ to the model $\mathcal{P}$.

  Assume the existence of a point $P^* \in \mathcal{P}$. That is, $\mathbb{E}_{P_0} \log \frac{p_0}{p^*} < \infty$. ($P_0 \ll P^*$)

# Notations

- $P_0$: 관측된 데이터의 실제 분포 (True distribution)
- $p_0$: dominating measure $\mu$에 대한 $P_0$의 확률 밀도 함수
- $\mathcal{P}$: Model, a collection of probability distributions
- $P, p$: 모델 $\mathcal{P}$ 내의 임의의 분포와 그 밀도 함수
- $\Pi$: 모델 $\mathcal{P}$ 위에 주어진 사전 분포 (Prior distribution)
- $\Pi_n(\cdot|X_1, \ldots, X_n)$: 데이터 $n$개가 주어졌을 때의 사후 분포
- $d$ : 모델 $\mathcal{P}$ 위에서 정의된 임의의 semi-metric

# $N_t(\epsilon, \mathcal{P}, d; P_0, P^*)$

A covering number for testing under misspecification

- Define $N_t(\epsilon, \mathcal{P}, d; P_0, P^*)$ as the minimal number $N$ of **convex sets** $B_1, \ldots, B_N$ of probability measures needed to cover the set $\{P \in \mathcal{P} : \epsilon < d(P, P^*) < 2\epsilon\}$ such that, for every $B_i$,

$$\inf_{P \in B_i} \sup_{0<\alpha<1} -\log \mathbb{E}_{P_0} \left( \frac{p}{p^*} \right)^{\alpha} \geq \frac{\epsilon^2}{4}$$

- 이 조건을 만족하는 finite covering이 없는 경우에는 $N_t(\epsilon, \mathcal{P}, d; P_0, P^*) = \infty$로 정의

# Main Theorem

## Theorem

*For a given model $\mathcal{P}$, prior $\Pi$ on $\mathcal{P}$ and some $P^* \in \mathcal{P}$, assume that $-\mathbb{E}_{P_0} \log(p^*/p_0) < \infty$ and $\mathbb{E}_{P_0}(p/p^*) < \infty$ for all $P \in \mathcal{P}$. Suppose that there exist a sequence of strictly positive numbers $\epsilon_n$ with $\epsilon_n \to 0$ and $n\epsilon_n^2 \to \infty$ and a constant $L > 0$, such that, for all $n$,*

$$N_t(\epsilon, \mathcal{P}, d; P_0, P^*) \leq \exp(n\epsilon_n^2) \quad \forall \epsilon > \epsilon_n. \tag{60}$$

$$\Pi\left(P \in \mathcal{P} : \mathbb{E}_{P_0} \log(p^*/p) \leq \epsilon_n^2, \mathbb{E}_{P_0}[\log(p^*/p)]^2 \leq \epsilon_n^2\right) \geq \exp(-Ln\epsilon_n^2) \tag{61}$$

*Then for every sufficiently large constant $M$, as $n \to \infty$,*

$$\Pi_n(P \in \mathcal{P} : d(P, P^*) \geq M\epsilon_n | X_1, \ldots, X_n) \to 0 \quad \text{in } L_1(P_0^n). \tag{62}$$

# Main Theorem

- (60) is about the entropy condition. Corresponds to (19).
- (61) is about the prior mass condition. Requires that the prior measure assign a certain minimal share of its total mass to $P^*$. Coincides with (20) when $P^* = P_0$.
- 차이점은 $P^*$에 대한 기댓값이 아니고 $P_0$에 대한 기댓값이라는 점

# Test function

- We now aim to construct a test function.
- Ghosal (2000): The test function $\phi_n$ distinguishes between $P_0$ and sieve $\mathcal{P}_n$ (p.59).

# Test function

- We now aim to construct a test function.
- Ghosal (2000): The test function $\phi_n$ distinguishes between $P_0$ and sieve $\mathcal{P}_n$ (p.59).
- On the other hand, when the model is misspecified, we need to construct a test function that distinguishes between $P^*$ and $\mathcal{P}$.
- This might seem counterintuitive since the observed data are realizations from the true distribution $P_0$.
- Thus, the authors introduce a new measure $Q(P)$ as follows and consider testing $Q(P)$ versus $P_0$:

$$dQ(P) = (p_0/p^*)dP$$

- Note that the measure $Q$ may not be a valid probability measure.

# Recap : $N_t(\epsilon, \mathcal{P}, d; P_0, P^*)$

A covering number for testing under misspecification

- Define $N_t(\epsilon, \mathcal{P}, d; P_0, P^*)$ as the minimal number $N$ of **convex sets** $B_1, \ldots, B_N$ of probability measures needed to cover the set $\{P \in \mathcal{P} : \epsilon < d(P, P^*) < 2\epsilon\}$ such that, for every $B_i$,

$$\inf_{P \in B_i} \sup_{0<\alpha<1} -\log \mathbb{E}_{P_0} \left( \frac{p}{p^*} \right)^{\alpha} \geq \frac{\epsilon^2}{4} \tag{63}$$

- The definition of $N_t$ is designed to ensure the existence of a test function $\phi_n$.

# Why Convex?

- Testing $P^*$ against a set of alternatives $B_i$ requires finding a single test $\phi$ that performs well uniformly over $B_i$.
- **Minimax Theorem :** For a convex set $B \subseteq \mathcal{P}$, the minimax testing risk equals the risk of testing against the *Least Favorable Pair*

$$\inf_{\phi \in \Phi} \sup_{Q \in B} (\mathbb{E}_P \phi + \mathbb{E}_Q (1 - \phi)) = \sup_{Q \in B} \inf_{\phi \in \Phi} (\mathbb{E}_P \phi + \mathbb{E}_Q (1 - \phi))$$

- **Geometric Intuition:** For convex sets, the linear decision boundary separating $P_0$ and the *Least Favorable Distribution* $P \in B_i$ strictly separates $P_0$ from the **entire** set $B_i$.
- If $B_i$ is non-convex, the boundary might intersect the set, leading to a catastrophic Type II error.

# Why (63)?

From Pseudo-measure to $\alpha$-Affinity

- To test $P_0$ against the pseudo-measure $dQ(P) = (p_0/p^*)dP$, use the $\alpha$-affinity, which bounds the optimal testing error for a single observation $(0 < \alpha < 1)$:

$$\text{Affinity}_\alpha(P_0, Q) = \int p_0^\alpha (dQ)^{1-\alpha} = \int p_0^\alpha \left(p\frac{p_0}{p^*}\right)^{1-\alpha} dx \tag{64}$$

$$= \int p_0 \left(\frac{p}{p*}\right)^{1-\alpha} dx = \mathbb{E}_{P_0}\left[\left(\frac{p}{p^*}\right)^{1-\alpha}\right] \tag{65}$$

- By symmetry (replacing $1-\alpha$ with $\alpha$), this is exactly $\mathbb{E}_{P_0}[(p/p^*)^\alpha]$.

# Why (63)?

## Decoding (63)

The paper requires: $\inf_{P \in B_i} \sup_{0<\alpha<1} -\log \mathbb{E}_{P_0}\left[\left(\frac{p}{p^*}\right)^\alpha\right] \geq \frac{\epsilon^2}{4}$

By negating and exponentiating, for the **worst-case** $P \in B_i$, we get:

$$\inf_{0<\alpha<1} \mathbb{E}_{P_0}\left[\left(\frac{p}{p^*}\right)^\alpha\right] \leq \exp\left(-\frac{\epsilon^2}{4}\right)$$

**Conclusion:** Raising this to the $n$-th power guarantees the existence of a test $\phi_n$ with an exponentially decaying error rate $\leq \exp(-n\epsilon^2/4)$ for the convex set $B_i, \forall i$.

# Test function

## Theorem

*Suppose* $P^* \in \mathcal{P}$ *and* $\mathbb{E}_{P_0}(p/p^*) < \infty$ *for all* $P \in \mathcal{P}$. *Assume that condition (60) holds, i.e.,*

$$N_t(\epsilon, \mathcal{P}, d; P_0, P^*) \leq \exp(n\epsilon_n^2) \quad \forall \epsilon > \epsilon_n.$$

*Then, there exist tests* $\phi_n$ *such that,*

$$\mathbb{E}_{P_0^n} \phi_n \leq 2\exp(-n\epsilon_n^2/4) \tag{66}$$

$$\sup_{P \in \mathcal{P}: d(P,P^*) > M\epsilon_n} \mathbb{E}_{Q(P)^n}(1-\phi_n) \leq \exp(-nM^2\epsilon_n^2/4). \tag{67}$$

# Proof of Main Theorem

Define $A_n$ and $B_n$ similarly as in p.56. That is,

$$A_n := \{X_1, \ldots, X_n : \int \prod_{i=1}^{n} \frac{p}{p^*}(X_i) d\Pi(P) \geq \exp(-2n\epsilon_n^2)\Pi(B_n)\}$$

$$B_n := \{P \in \mathcal{P} : \mathbb{E}_{P_0}(\log p^*/p) \leq \epsilon_n^2, \mathbb{E}_{P_0}[(\log p^*/p)]^2 \leq \epsilon_n^2\}$$

Note that $B_n$ appears in (61).

# Proof of Main Theorem

We prove the main theorem by showing that each term in the following decomposition converges to zero.

$$
\begin{aligned}
&\mathbb{E}_{P_0}[\Pi_n(d(P,P^*) \geq M\epsilon_n | X_1,\ldots,X_n)] \\
&= \mathbb{E}_{P_0}[\Pi_n(d(P,P^*) \geq M\epsilon_n | X_1,\ldots,X_n)\phi_n] && (68)\\
&+ \mathbb{E}_{P_0}[\Pi_n(d(P,P^*) \geq M\epsilon_n | X_1,\ldots,X_n)(1-\phi_n)\mathbb{I}_{A_n^c}] && (69)\\
&+ \mathbb{E}_{P_0}[\Pi_n(d(P,P^*) \geq M\epsilon_n | X_1,\ldots,X_n)(1-\phi_n)\mathbb{I}_{A_n}] && (70)
\end{aligned}
$$

- The first term (68) $\leq \mathbb{E}_{P_0^n}[\phi_n] \leq 2\exp(-n\epsilon_n^2/4) \to 0$ from (66)
- The second term (69) $\leq P_0^n(A_n^c) \to 0$ by Lemma 8.1(Ghosal 2000).

# Proof of Main Theorem

- The final term (70) converges to 0 similarly as in p.65. That is,

$$
\begin{aligned}
&\mathbb{E}_{P_0^n}[\Pi_n(d(P,P^*) \geq M\epsilon_n | X_1,\ldots,X_n)(1-\phi_n)\mathbb{I}_{A_n}] \\
&= \mathbb{E}_{P_0^n}\left[\frac{\int_{P\in\mathcal{P}:d(P,P^*)\geq M\epsilon_n}\prod_{i=1}^n \frac{p}{p^*}(X_i)d\Pi_n(P)}{\int\prod_{i=1}^n \frac{p}{p^*}(X_i)d\Pi_n(P)}(1-\phi_n)\mathbb{I}(A_n)\right] \\
&\leq \int_{P\in\mathcal{P}:d(P,P^*)\geq M\epsilon_n}\mathbb{E}_{P_0^n}\left[\prod_{i=1}^n \frac{p}{p^*}(X_i)(1-\phi_n)\right]d\Pi_n(P)\frac{\exp(2n\epsilon_n^2)}{\Pi_n(B_n)} \\
&\leq \int_{P\in\mathcal{P}:d(P,P^*)\geq M\epsilon_n}\mathbb{E}_{Q(P)^n}(1-\phi_n)d\Pi_n(P)\frac{\exp(2n\epsilon_n^2)}{\Pi_n(B_n)} \\
&\leq \exp(-nM^2\epsilon_n^2/4)\frac{\exp(2n\epsilon_n^2)}{\Pi_n(B_n)} \to 0 \text{ from (67)}
\end{aligned}
$$

# Outline

# Gibbs posterior

**Notations**

- Data: $U \sim P$ (in most cases, $U = (X, Y)$), $U \in \mathcal{U}$.
- Loss function $l_\theta(u) : \mathcal{U} \to \mathbb{R}$
  - e.g. $l_\theta(u) = (y - \theta(x))^2$ for $u = (x, y)$ and a function of interest $\theta$.
- Population risk: $R(\theta) = \mathbb{E}_{U \sim P} l_\theta(U)$
- Empirical risk: $R_n(\theta) = \frac{1}{n} \sum_{i=1}^n l_\theta(U_i)$
- Population risk minimizer:

$$\theta^* \in \underset{\theta \in \Theta}{\operatorname{argmin}} R(\theta) \tag{71}$$

- Prior: $\pi(\cdot)$

# Gibbs posterior

- Gibbs posterior = Generalized posterior; update the belief via empirical risk, not likelihood (Bissiri (2016)).

## Definition (Gibbs posterior)

Given a loss function $l_\theta$ and the corresponding empirical risk $R_n$, define the Gibbs posterior as:

$$\pi_n^{(\omega)}(\theta|\mathcal{D}_n) \propto e^{-\omega n R_n(\theta)}\pi(\theta),\ \theta \in \Theta \tag{72}$$

- $\omega > 0$: called learning rate.

# Posterior Concentration

- Following Syring and Martin (2023).

## Concentration

$$\pi_n^{(\omega)}(\{\theta : d(\theta, \theta^*) > M\varepsilon_n\}|\mathcal{D}_n) \to 0 \text{ in } P^n\text{-probability as } n \to \infty \tag{73}$$

for $n\varepsilon_n^r \to \infty$ and a large constant $M > 0$.

# Conditions

- Concentration can be done, when following 2 conditions are satisfied.
- $m(\theta, \theta^*) = \mathbb{E}_{U\sim P}(l_\theta - l_{\theta^*}),\ v(\theta, \theta^*) = \mathbb{E}_{U\sim P}(l_\theta - l_{\theta^*})^2 - m(\theta, \theta^*)^2.$

## Prior concentration condition

$$\log \pi\left(\{\theta : m(\theta, \theta^*) \vee v(\theta, \theta^*) \leq \varepsilon_n^r\}\right) \geq -Cn\omega\varepsilon_n^r \tag{74}$$

## Sub-exponential loss condition

$$d(\theta, \theta^*) > \delta \Rightarrow \log \mathbb{E}_{U\sim P}[e^{-\omega(l_\theta - l_{\theta^*})}] < -K\omega\delta^r \tag{75}$$

for all sequences $0 < \omega \leq \bar{\omega}$ and all sufficiently small $\delta > 0$.

# Comparison

Table: Comparison on conditions.

| | **Gibbs conditions** |
|---|---|
| Prior<br>Sub-exponential | $\log \pi\left(\{\theta : m(\theta,\theta^*) \vee v(\theta,\theta^*) \le \varepsilon_n^r\}\right) \ge -Cn\omega\varepsilon_n^r$<br>$d(\theta,\theta^*) > \delta \Rightarrow \log \mathbb{E}_{U\sim P}[e^{-\omega\left(l_\theta - l_{\theta^*}\right)}] < -K\omega\delta^r$ |
| | **Bayesian nonparam conditions** |
| Prior<br>Test type-II | $\Pi_n\left(P : \mathbb{E}_{P_0}\left[\log \frac{p_0}{p}\right] \le \epsilon_n^2, \mathbb{E}_{P_0}\left[\log \frac{p_0}{p}\right]^2 \le \epsilon_n^2\right) \ge \exp(-n\epsilon_n^2 C)$<br>$\sup_{P\in\mathcal{P}_n : d(P,P_0) > M\epsilon_n} \mathbb{E}_{P^n}(1-\phi_n) \le \exp(-KnM^2\epsilon_n^2)$ |

# Proof Strategy

$$A_n = \{\theta : d(\theta, \theta^*) > M\varepsilon_n\} \tag{76}$$

$$N_n(A_n) = \int_{A_n} e^{-\omega n\{R_n(\theta) - R_n(\theta^*)\}} \pi(d\theta) \tag{77}$$

$$D_n = \int e^{-\omega n\{R_n(\theta) - R_n(\theta^*)\}} \pi(d\theta) \tag{78}$$

$$\pi_n^{(\omega)}(A_n|\mathcal{D}_n) = \frac{N_n(A_n)}{D_n} \tag{79}$$

- Goal: $\pi_n^{(\omega)}(A_n|\mathcal{D}_n) \to 0$
- Strategy: lower bound $D_n$ and upper bound $N_n(A_n)$.

# Lemma: $D_n$ lower bound

## Lemma

*Let* $G_n = \{\theta : m(\theta, \theta^*) \vee v(\theta, \theta^*) \leq \varepsilon_n^r\}$. *If* $\varepsilon_n \to 0$ *and* $n\varepsilon_n^r \to \infty$, *then*

$$P^n \left[ D_n > \frac{1}{2} \pi(G_n) e^{-2n\omega\varepsilon_n^r} \right] \geq 1 - 2(n\varepsilon_n^r)^{-1} \to 1. \tag{80}$$

# Main proof (with $N_n(A_n)$ upper bound)

Denote the lower bound on $D_n$ as:

$$b_n = \frac{1}{2}\pi(G_n)e^{-2\omega n\varepsilon_n^r}.$$

Then,

$$\begin{aligned}\pi_n^{(\omega)}(A_n|\mathcal{D}_n) &\leq \frac{N_n(A_n)}{D_n}\mathbb{1}(D_n > b_n) + \mathbb{1}(D_n \leq b_n) \\ &\leq b_n^{-1}N_n(A_n) + \mathbb{1}(D_n \leq b_n).\end{aligned}$$

# Main proof

By *sub-exponential loss condition* and independence of $U^n$, we get

$$\mathbb{E}_{U^n \sim P^n} N_n(A_n) = \int_{A_n} (\mathbb{E}_{U \sim P}[e^{-w(l_\theta - l_{\theta^*})}])^n \pi(d\theta) < e^{-Kn\omega(M\varepsilon_n)^r}.$$

By *prior concentration condition*, we get $\pi(G_n) \geq e^{-Cn\varepsilon_n^r}$.
By the Lemma, we get $P(D_n \leq b_n) \geq 2(n\varepsilon_n^r)^{-1}$.
Therefore,

$$\mathbb{E}_{U^n \sim P^n} \pi_n^{(\omega)}(A_n | \mathcal{D}_n) \leq 2e^{-(\omega K M^r - C - 2\omega)n\varepsilon_n^r} + 2(n\varepsilon_n^r)^{-1} \to 0.$$

# Proof of Lemma

Denote $m(\theta, \theta^*)$ and $v(\theta, \theta^*)$ as $m(\theta), v(\theta)$ for brevity. Let

$$Z_n(\theta) = \frac{\{nR_n(\theta) - nR_n(\theta^*)\} - nm(\theta)}{\{nv(\theta)\}^{1/2}}.$$

Let

$$\mathcal{Z}_n = \{(\theta, U^n) : |Z_n(\theta)| \geq (n\varepsilon_n^r)^{1/2}\}.$$

Let

$$\mathcal{Z}_n(\theta) = \{U^n : (\theta, U^n) \in \mathcal{Z}_n\} \text{ and } \mathcal{Z}_n(U^n) = \{\theta : (\theta, U^n) \in \mathcal{Z}_n\}.$$

Then,

$$nR_n(\theta) - nR_n(\theta^*) = nm(\theta) + \{nv(\theta)\}^{1/2} Z_n(\theta).$$

# Proof of Lemma

Then, we have

$$\begin{aligned} D_n &\geq \int_{G_n \cap \mathcal{Z}_n(U^n)^c} e^{-\omega n m(\theta) - \omega \{nv(\theta)\}^{1/2} Z_n(\theta)} \pi(d\theta) \\ &\geq e^{-2\omega n \varepsilon_n^r} \pi(G_n \cap \mathcal{Z}_n(U^n)^c). \end{aligned}$$

Hence,

$$\begin{aligned} & P^n \left[ D_n \leq \frac{1}{2} \pi(G_n) e^{-2\omega n \varepsilon_n^r} \right] \\ & \leq P^n \left[ e^{-2\omega n \varepsilon_n^r} \pi(G_n \cap \mathcal{Z}_n(U^n)^c) \leq \frac{1}{2} \pi(G_n) e^{-2\omega n \varepsilon_n^r} \right] \\ & \leq P^n \left[ \pi(G_n \cap \mathcal{Z}_n(U^n)) \geq \frac{1}{2} \pi(G_n) \right] \\ & \leq \frac{2\mathbb{E}_{U^n \sim P^n} [\pi(G_n \cap \mathcal{Z}_n(U^n))]}{\pi(G_n)} \quad \text{(Markov ineq.)} \end{aligned}$$

# Proof of Lemma

The expectation term in the numerator is simplified:

$$\begin{aligned}
\mathbb{E}_{U^n \sim P^n}[\pi(G_n \cap \mathcal{Z}_n(U^n))] &= \int \int \mathbb{1}\{\theta \in G_n \cap \mathcal{Z}_n(U^n)\}\pi(d\theta)P^n(dU^n) \\
&= \int \int \mathbb{1}\{\theta \in G_n\}\mathbb{1}\{\theta \in \mathcal{Z}_n(U^n)\}P^n(dU^n)\pi(d\theta) \\
&= \int_{G_n} \mathbb{E}_{U^n \sim P^n}[\mathcal{Z}_n(\theta)]\pi(d\theta) \\
&\leq (n\varepsilon_n^r)^{-1}\pi(G_n) \quad \text{(Chebyshev)}.
\end{aligned}$$

Hence,

$$P^n\left[D_n \leq \frac{1}{2}\pi(G_n)e^{-2\omega n\varepsilon_n^r}\right] \leq 2(n\varepsilon_n^r)^{-1} \to 0.$$

# References I


1. Ghosal, S., Ghosh, J. K., & Van Der Vaart, A. W. (2000). Convergence rates of posterior distributions. Annals of Statistics, 500-531.
2. Bissiri, P. G., Holmes, C. C., & Walker, S. G. (2016). A general framework for updating belief distributions. Journal of the Royal Statistical Society Series B: Statistical Methodology, 78(5), 1103-1130.
3. Syring, N., & Martin, R. (2023). Gibbs posterior concentration rates under sub-exponential type losses. Bernoulli, 29(2), 1080-1108.

# Outline

# Bounding the Error Rate with $\alpha$-Affinity

## 1. The Likelihood Ratio Test (1 vs 1)

For a single observation $x$, we want to test $P_0$ against $Q$. We define the simplest and most powerful test function $\phi$:

$$\phi(x) = \mathbf{1}\{q(x) \geq p_0(x)\}$$

*(i.e., Reject $P_0$ if the alternative $Q$ has a higher density).*

# Bounding the Error Rate with $\alpha$-Affinity

## 2. Bounding Type I & Type II Errors (for $0 < \alpha < 1$)

By introducing $\alpha$, we can artificially inflate the integrands:

- **Type I Error:** On the region $\{q \geq p_0\}$, we have $(q/p_0)^{1-\alpha} \geq 1$.

$$\mathbb{E}_{P_0}[\phi] = \int_{q \geq p_0} p_0 dx \leq \int_{q \geq p_0} p_0 \left(\frac{q}{p_0}\right)^{1-\alpha} dx = \int_{q \geq p_0} p_0^{\alpha} q^{1-\alpha} dx$$

- **Type II Error:** On the region $\{q < p_0\}$, we have $(p_0/q)^{\alpha} > 1$.

$$\mathbb{E}_{Q}[1-\phi] = \int_{q < p_0} q dx \leq \int_{q < p_0} q \left(\frac{p_0}{q}\right)^{\alpha} dx = \int_{q < p_0} p_0^{\alpha} q^{1-\alpha} dx$$

# Bounding the Error Rate with $\alpha$-Affinity

## 3. The Total Error Bound

Since the two regions perfectly partition the whole space, summing the errors yields the exact definition of $\alpha$-Affinity:

$$\text{Total Error} \leq \int_{\{q \geq p_0\} \cup \{q < p_0\}} p_0^{\alpha} q^{1-\alpha} dx = \int p_0^{\alpha} q^{1-\alpha} dx = \text{Affinity}_{\alpha}(P_0, Q)$$

# $\alpha$-Affinity

When we have a single observation $X$,

$$\text{Affinity}(P, P^*) = \mathbb{E}_P \left[ \left( \frac{p(X)}{p^*(X)} \right)^\alpha \right]$$

Total Affinity

$$\begin{aligned} \text{TotalAffinity}(P, P^*) &= \mathbb{E}_{P^n} \left[ \left( \frac{\prod_{i=1}^n p(X)}{\prod_{i=1}^n p^*(X)} \right)^\alpha \right] = \prod_{i=1}^n \mathbb{E}_P \left[ \left( \frac{p(X)}{p^*(X)} \right)^\alpha \right] \\ &= \text{Affinity}(P, P^*)^n \end{aligned}$$