STAT 576 Bayesian Analysis

Lecture 4: Asymptotic Properties of Bayesian Inference

Chencheng Cai

Washington State University

Normal Approximation to the Posterior Distribution

- Let $\hat{\boldsymbol{\theta}}$ be the maximize-a-posteriori (MAP) estimator, that is

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \Theta}{\arg\max} \; p(\boldsymbol{\theta} \mid y)$$

Normal Approximation to the Posterior Distribution

- Let $\hat{\boldsymbol{\theta}}$ be the maximize-a-posteriori (MAP) estimator, that is

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \Theta} \; p(\boldsymbol{\theta} \mid y)$$

- Consider a Taylor expansion of the $\log p(\boldsymbol{\theta} \mid y)$ at its mode $\hat{\boldsymbol{\theta}}$:

$$\log p(\boldsymbol{\theta} \mid y) = \log p(\hat{\boldsymbol{\theta}} \mid y) + \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \left[\frac{d^2}{d\boldsymbol{\theta}^2} \log p(\boldsymbol{\theta} \mid y) \right]_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + o(\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|^2)$$

Normal Approximation to the Posterior Distribution

- Let $\hat{\boldsymbol{\theta}}$ be the maximize-a-posteriori (MAP) estimator, that is

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \Theta} \; p(\boldsymbol{\theta} \mid y)$$

- Consider a Taylor expansion of the $\log p(\boldsymbol{\theta} \mid y)$ at its mode $\hat{\boldsymbol{\theta}}$:

$$\log p(\boldsymbol{\theta} \mid y) = \log p(\hat{\boldsymbol{\theta}} \mid y) + \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \left[\frac{d^2}{d\boldsymbol{\theta}^2} \log p(\boldsymbol{\theta} \mid y) \right]_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + o(\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|^2)$$

- The linear term is omitted because

$$\left[\frac{d}{d\boldsymbol{\theta}} \log p(\boldsymbol{\theta} \mid y) \right]_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}} = \mathbf{0}$$

Normal Approximation to the Posterior Distribution

- With the second approximation of the log-density around the mode:

$$\log p(\boldsymbol{\theta} \mid y) \approx \log p(\hat{\boldsymbol{\theta}} \mid y) + \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \left[\frac{d^2}{d\boldsymbol{\theta}^2} \log p(\boldsymbol{\theta} \mid y)\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})$$

Normal Approximation to the Posterior Distribution

- With the second approximation of the log-density around the mode:

$$\log p(\boldsymbol{\theta} \mid y) \approx \log p(\hat{\boldsymbol{\theta}} \mid y) + \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \left[\frac{d^2}{d\boldsymbol{\theta}^2} \log p(\boldsymbol{\theta} \mid y)\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})$$

- we have the normal approximation of the posterior by

$$p(\boldsymbol{\theta} \mid y) \approx \mathcal{N}\left(\hat{\boldsymbol{\theta}}, \boldsymbol{J}(\hat{\boldsymbol{\theta}})^{-1}\right)$$

  where

$$\boldsymbol{J}(\boldsymbol{\theta}) = -\frac{d^2}{d\boldsymbol{\theta}^2} \log p(\boldsymbol{\theta} \mid y)$$

  is the **observed information matrix**.

Normal Approximation to the Posterior Distribution

$$p(\boldsymbol{\theta} \mid y) \approx \mathcal{N}\left(\hat{\boldsymbol{\theta}}, \boldsymbol{J}(\hat{\boldsymbol{\theta}})^{-1}\right)$$

- The normal approximation works for any distribution of θ (with mode $\hat{\theta}$) when

Normal Approximation to the Posterior Distribution

$$p(\boldsymbol{\theta} \mid y) \approx \mathcal{N}\left(\hat{\boldsymbol{\theta}}, \boldsymbol{J}(\hat{\boldsymbol{\theta}})^{-1}\right)$$

- The normal approximation works for any distribution of θ (with mode $\hat{\theta}$) when
  - $\hat{\boldsymbol{\theta}}$ is an inner point of Θ.
  - $\log p(\boldsymbol{\theta} \mid y)$ is second-order differentiable at $\hat{\boldsymbol{\theta}}$.
  - $\boldsymbol{J}(\hat{\boldsymbol{\theta}})$ is positive-definite / non-singular.

Normal Approximation to the Posterior Distribution

$$p(\boldsymbol{\theta} \mid y) \approx \mathcal{N}\left(\hat{\boldsymbol{\theta}}, \boldsymbol{J}(\hat{\boldsymbol{\theta}})^{-1}\right)$$

- The normal approximation works for any distribution of θ (with mode $\hat{\theta}$) when
  - $\hat{\boldsymbol{\theta}}$ is an inner point of Θ.
  - $\log p(\boldsymbol{\theta} \mid y)$ is second-order differentiable at $\hat{\boldsymbol{\theta}}$.
  - $\boldsymbol{J}(\hat{\boldsymbol{\theta}})$ is positive-definite / non-singular.
- Using Bayes' rule, we have

$$\boldsymbol{J}(\boldsymbol{\theta}) = -\frac{d^2}{d\boldsymbol{\theta}^2}\log p(\boldsymbol{\theta} \mid y) = \underbrace{-\frac{d^2}{d\boldsymbol{\theta}^2}\log p(y \mid \boldsymbol{\theta})}_{\text{info. from observations}} \quad \underbrace{-\frac{d^2}{d\boldsymbol{\theta}^2}\log p(\boldsymbol{\theta})}_{\text{info. from prior}}$$

Information Matrix

- Suppose we have i.i.d. observations $y = (y_1, \dots, y_n)$ from a distribution $F_{\boldsymbol{\theta}}$ from a parametric family $\{F_{\boldsymbol{\theta}_0} : \boldsymbol{\theta} \in \Theta\}$ with true parameter $\boldsymbol{\theta}_0$.

Information Matrix

- Suppose we have i.i.d. observations $y = (y_1, \ldots, y_n)$ from a distribution $F_{\boldsymbol{\theta}}$ from a parametric family $\{F_{\boldsymbol{\theta}_0} : \boldsymbol{\theta} \in \Theta\}$ with true parameter $\boldsymbol{\theta}_0$.
- Then the observed information matrix is

$$\boldsymbol{J}_n(\boldsymbol{\theta}) = -\sum_{i=1}^{n} \frac{d^2}{d\boldsymbol{\theta}^2} \log p(y_i \mid \boldsymbol{\theta}) - \frac{d^2}{d\boldsymbol{\theta}^2} \log p(\boldsymbol{\theta})$$

Information Matrix

- Suppose we have i.i.d. observations $y = (y_1, \ldots, y_n)$ from a distribution $F_{\boldsymbol{\theta}}$ from a parametric family $\{F_{\boldsymbol{\theta}_0} : \boldsymbol{\theta} \in \Theta\}$ with true parameter $\boldsymbol{\theta}_0$.
- Then the observed information matrix is

$$\boldsymbol{J}_n(\boldsymbol{\theta}) = -\sum_{i=1}^{n} \frac{d^2}{d\boldsymbol{\theta}^2} \log p(y_i \mid \boldsymbol{\theta}) - \frac{d^2}{d\boldsymbol{\theta}^2} \log p(\boldsymbol{\theta})$$

- With Law of Large Numbers, we know

$$-\frac{1}{n} \sum_{i=1}^{n} \frac{d^2}{d\boldsymbol{\theta}^2} \log p(y_i \mid \boldsymbol{\theta}) \xrightarrow{F_{\boldsymbol{\theta}_0}} \mathbb{E}_{\boldsymbol{\theta}_0}\left[-\frac{d^2}{d\boldsymbol{\theta}^2} \log p(y_i \mid \boldsymbol{\theta})\right]$$

Information Matrix

- Suppose we have i.i.d. observations $y = (y_1, \ldots, y_n)$ from a distribution $F_{\boldsymbol{\theta}}$ from a parametric family $\{F_{\boldsymbol{\theta}_0} : \boldsymbol{\theta} \in \Theta\}$ with true parameter $\boldsymbol{\theta}_0$.
- Then the observed information matrix is

$$\boldsymbol{J}_n(\boldsymbol{\theta}) = -\sum_{i=1}^{n} \frac{d^2}{d\boldsymbol{\theta}^2} \log p(y_i \mid \boldsymbol{\theta}) - \frac{d^2}{d\boldsymbol{\theta}^2} \log p(\boldsymbol{\theta})$$

- With Law of Large Numbers, we know

$$-\frac{1}{n} \sum_{i=1}^{n} \frac{d^2}{d\boldsymbol{\theta}^2} \log p(y_i \mid \boldsymbol{\theta}) \xrightarrow{F_{\boldsymbol{\theta}_0}} \mathbb{E}_{\boldsymbol{\theta}_0}\left[-\frac{d^2}{d\boldsymbol{\theta}^2} \log p(y_i \mid \boldsymbol{\theta})\right]$$

- Note: This is **NOT** the Fisher's information matrix because the expectation is taken under the true parameter $\boldsymbol{\theta}_0$.

Normal Approximation to the Posterior Distribution

- With the approximation from previous slide, we can revise the Taylor expansion of $\log p(\boldsymbol{\theta} \mid y)$ to

$$\log p(\boldsymbol{\theta} \mid y) = \log p(\hat{\boldsymbol{\theta}} \mid y) + \frac{n}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \mathbb{E}_{\boldsymbol{\theta}_0}\left[-\frac{d^2}{d\boldsymbol{\theta}^2}\log p(y_i \mid \boldsymbol{\theta})\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + o_P(n\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|^2)$$

Normal Approximation to the Posterior Distribution

- With the approximation from previous slide, we can revise the Taylor expansion of $\log p(\boldsymbol{\theta} \mid y)$ to

$$\log p(\boldsymbol{\theta} \mid y) = \log p(\hat{\boldsymbol{\theta}} \mid y) + \frac{n}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \mathbb{E}_{\boldsymbol{\theta}_0}\left[-\frac{d^2}{d\boldsymbol{\theta}^2}\log p(y_i \mid \boldsymbol{\theta})\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + o_P(n\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|^2)$$

- Then the posterior can be approximated by

$$p(\boldsymbol{\theta} \mid y) \approx \mathcal{N}\left(\boldsymbol{\theta} \;\middle|\; \hat{\boldsymbol{\theta}}, \frac{1}{n}\mathbb{E}_{\boldsymbol{\theta}_0}^{-1}\left[-\frac{d^2}{d\boldsymbol{\theta}^2}\log p(y_i \mid \boldsymbol{\theta})\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right)$$

Normal Approximation to the Posterior Distribution

- With the approximation from previous slide, we can revise the Taylor expansion of $\log p(\boldsymbol{\theta} \mid y)$ to

$$\log p(\boldsymbol{\theta} \mid y) = \log p(\hat{\boldsymbol{\theta}} \mid y) + \frac{n}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \mathbb{E}_{\boldsymbol{\theta}_0}\left[-\frac{d^2}{d\boldsymbol{\theta}^2}\log p(y_i \mid \boldsymbol{\theta})\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + o_P(n\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|^2)$$

- Then the posterior can be approximated by

$$p(\boldsymbol{\theta} \mid y) \approx \mathcal{N}\left(\boldsymbol{\theta} \,\middle|\, \hat{\boldsymbol{\theta}}, \frac{1}{n}\mathbb{E}^{-1}_{\boldsymbol{\theta}_0}\left[-\frac{d^2}{d\boldsymbol{\theta}^2}\log p(y_i \mid \boldsymbol{\theta})\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right)$$

- Or the rescaled version:

$$p(\sqrt{n}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \mid y) \approx \mathcal{N}\left(\boldsymbol{h} \,\middle|\, \mathbf{0}, \mathbb{E}^{-1}_{\boldsymbol{\theta}_0}\left[-\frac{d^2}{d\boldsymbol{\theta}^2}\log p(y_i \mid \boldsymbol{\theta})\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right)$$

where $\boldsymbol{h} = \sqrt{n}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})$ is called the **local parameter** to $\hat{\boldsymbol{\theta}}$.

Normal Approximation to the Posterior Distribution

The approximation

$$p(\sqrt{n}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \mid y) \approx \mathcal{N}\left(\boldsymbol{h} \,\middle|\, \mathbf{0}, \mathbb{E}^{-1}_{\boldsymbol{\theta}_0}\left[-\frac{d^2}{d\boldsymbol{\theta}^2} \log p(y_i \mid \boldsymbol{\theta})\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right)$$

is still not satisfying as an asymptotic result because

Normal Approximation to the Posterior Distribution

The approximation

$$p(\sqrt{n}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \mid y) \approx \mathcal{N}\left(\boldsymbol{h} \,\middle|\, \mathbf{0}, \mathbb{E}_{\boldsymbol{\theta}_0}^{-1}\left[-\frac{d^2}{d\boldsymbol{\theta}^2} \log p(y_i \mid \boldsymbol{\theta})\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right)$$

is still not satisfying as an asymptotic result because

- It is a finite sample approximation.
- The variance depends on the true parameter $\boldsymbol{\theta}_0$ and thus infeasible.
- The variance is random by involving $\hat{\boldsymbol{\theta}}$ in the formula.

Normal Approximation to the Posterior Distribution

The approximation

$$p(\sqrt{n}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \mid y) \approx \mathcal{N}\left(\boldsymbol{h} \,\middle|\, \mathbf{0}, \mathbb{E}^{-1}_{\boldsymbol{\theta}_0}\left[-\frac{d^2}{d\boldsymbol{\theta}^2}\log p(y_i \mid \boldsymbol{\theta})\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right)$$

is still not satisfying as an asymptotic result because

- It is a finite sample approximation.
- The variance depends on the true parameter $\boldsymbol{\theta}_0$ and thus infeasible.
- The variance is random by involving $\hat{\boldsymbol{\theta}}$ in the formula.

Therefore, we need first to investigate the asymptotic behavior of $\hat{\boldsymbol{\theta}}$ itself.

Asymptotic Equivalence of MAP and MLE

- Maximize-a-posteriori estimator:

$$\hat{\boldsymbol{\theta}}_n^{(map)} = \arg\max \ \log p(\boldsymbol{\theta} \mid y) = \arg\max \ \underbrace{\frac{1}{n}\sum_{i=1}^{n} \log p(y_i \mid \boldsymbol{\theta}) + \frac{1}{n}\log p(\boldsymbol{\theta})}_{f_n(\boldsymbol{\theta})}$$

Asymptotic Equivalence of MAP and MLE

- Maximize-a-posteriori estimator:

$$\hat{\boldsymbol{\theta}}_n^{(map)} = \arg\max \; \log p(\boldsymbol{\theta} \mid y) = \arg\max \; \underbrace{\frac{1}{n}\sum_{i=1}^{n} \log p(y_i \mid \boldsymbol{\theta}) + \frac{1}{n}\log p(\boldsymbol{\theta})}_{f_n(\boldsymbol{\theta})}$$

- Maximum Likelihood Estimator:

$$\hat{\boldsymbol{\theta}}_n^{(mle)} = \arg\max \; \log p(y \mid \boldsymbol{\theta}) = \arg\max \; \underbrace{\frac{1}{n}\sum_{i=1}^{n} \log p(y_i \mid \boldsymbol{\theta})}_{g_n(\boldsymbol{\theta})}$$

Asymptotic Equivalence of MAP and MLE

- Maximize-a-posteriori estimator:

$$\hat{\boldsymbol{\theta}}_n^{(map)} = \arg\max \ \log p(\boldsymbol{\theta} \mid y) = \arg\max \ \underbrace{\frac{1}{n}\sum_{i=1}^{n} \log p(y_i \mid \boldsymbol{\theta}) + \frac{1}{n}\log p(\boldsymbol{\theta})}_{f_n(\boldsymbol{\theta})}$$

- Maximum Likelihood Estimator:

$$\hat{\boldsymbol{\theta}}_n^{(mle)} = \arg\max \ \log p(y \mid \boldsymbol{\theta}) = \arg\max \ \underbrace{\frac{1}{n}\sum_{i=1}^{n} \log p(y_i \mid \boldsymbol{\theta})}_{g_n(\boldsymbol{\theta})}$$

- The difference $f_n(\boldsymbol{\theta}) - g_n(\boldsymbol{\theta})$ does not uniformly converge to zero.

Asymptotic Equivalence of MAP and MLE

- Maximize-a-posteriori estimator:

$$\hat{\boldsymbol{\theta}}_n^{(map)} = \arg\max \ \log p(\boldsymbol{\theta} \mid y) = \arg\max \ \underbrace{\frac{1}{n}\sum_{i=1}^{n} \log p(y_i \mid \boldsymbol{\theta}) + \frac{1}{n}\log p(\boldsymbol{\theta})}_{f_n(\boldsymbol{\theta})}$$

- Maximum Likelihood Estimator:

$$\hat{\boldsymbol{\theta}}_n^{(mle)} = \arg\max \ \log p(y \mid \boldsymbol{\theta}) = \arg\max \ \underbrace{\frac{1}{n}\sum_{i=1}^{n} \log p(y_i \mid \boldsymbol{\theta})}_{g_n(\boldsymbol{\theta})}$$

- The difference $f_n(\boldsymbol{\theta}) - g_n(\boldsymbol{\theta})$ does not uniformly converge to zero.
- But since $p(\hat{\boldsymbol{\theta}}_n^{(map)}) \geq p(\hat{\boldsymbol{\theta}}_n^{(mle)})$, as long as $\hat{\boldsymbol{\theta}}_n^{(mle)} \in \{\boldsymbol{\theta} \in \Theta : p(\boldsymbol{\theta}) > 0\}$, we only need to consider the subset with positive prior density.

Asymptotic Equivalence of MAP and MLE

- Maximize-a-posteriori estimator:

$$\hat{\boldsymbol{\theta}}_n^{(map)} = \arg\max \ \log p(\boldsymbol{\theta} \mid y) = \arg\max \ \underbrace{\frac{1}{n}\sum_{i=1}^{n} \log p(y_i \mid \boldsymbol{\theta}) + \frac{1}{n}\log p(\boldsymbol{\theta})}_{f_n(\boldsymbol{\theta})}$$

- Maximum Likelihood Estimator:

$$\hat{\boldsymbol{\theta}}_n^{(mle)} = \arg\max \ \log p(y \mid \boldsymbol{\theta}) = \arg\max \ \underbrace{\frac{1}{n}\sum_{i=1}^{n} \log p(y_i \mid \boldsymbol{\theta})}_{g_n(\boldsymbol{\theta})}$$

- The difference $f_n(\boldsymbol{\theta}) - g_n(\boldsymbol{\theta})$ does not uniformly converge to zero.
- But since $p(\hat{\boldsymbol{\theta}}_n^{(map)}) \geq p(\hat{\boldsymbol{\theta}}_n^{(mle)})$, as long as $\hat{\boldsymbol{\theta}}_n^{(mle)} \in \{\boldsymbol{\theta} \in \Theta : p(\boldsymbol{\theta}) > 0\}$, we only need to consider the subset with positive prior density.
- A sufficient condition is (1) $\hat{\boldsymbol{\theta}}_n^{(mle)}$ is consistent for $\boldsymbol{\theta}_0$, and (2) $p(\boldsymbol{\theta})$ is strictly positive in a neighbor of $\boldsymbol{\theta}_0$.

Normal Approximation to the Posterior Distribution

- Under regularity conditions on the prevoius slide, we have

$$\hat{\theta}_n^{(map)} \xrightarrow{P} \boldsymbol{\theta}_0$$

Normal Approximation to the Posterior Distribution

- Under regularity conditions on the prevoius slide, we have

$$\hat{\theta}_n^{(map)} \xrightarrow{P} \boldsymbol{\theta}_0$$

- Therefore,

$$\mathbb{E}_{\boldsymbol{\theta}_0}\left[-\frac{d^2}{d\boldsymbol{\theta}^2}\log p(y_i \mid \boldsymbol{\theta})\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_n} \xrightarrow{P} \mathbb{E}_{\boldsymbol{\theta}_0}\left[-\frac{d^2}{d\boldsymbol{\theta}^2}\log p(y_i \mid \boldsymbol{\theta})\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \mathcal{I}(\boldsymbol{\theta}_0)$$

Normal Approximation to the Posterior Distribution

- Under regularity conditions on the prevoius slide, we have

$$\hat{\theta}_n^{(map)} \xrightarrow{P} \boldsymbol{\theta}_0$$

- Therefore,

$$\mathbb{E}_{\boldsymbol{\theta}_0}\left[-\frac{d^2}{d\boldsymbol{\theta}^2}\log p(y_i \mid \boldsymbol{\theta})\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_n} \xrightarrow{P} \mathbb{E}_{\boldsymbol{\theta}_0}\left[-\frac{d^2}{d\boldsymbol{\theta}^2}\log p(y_i \mid \boldsymbol{\theta})\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \mathcal{I}(\boldsymbol{\theta}_0)$$

- In this case, the approximation of the posterior is

$$p(\sqrt{n}(\boldsymbol{\theta}-\boldsymbol{\theta}_0) \mid y) \approx \mathcal{N}\left(\boldsymbol{h} \,\middle|\, \mathbf{0}, \mathcal{I}^{-1}(\boldsymbol{\theta}_0)\right)$$

with $\boldsymbol{h} = \sqrt{n}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)$ the **local parameter**.

Normal Approximation to the Posterior Distribution

- Under regularity conditions on the prevoius slide, we have

$$\hat{\theta}_n^{(map)} \xrightarrow{P} \boldsymbol{\theta}_0$$

- Therefore,

$$\mathbb{E}_{\boldsymbol{\theta}_0}\left[-\frac{d^2}{d\boldsymbol{\theta}^2}\log p(y_i \mid \boldsymbol{\theta})\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_n} \xrightarrow{P} \mathbb{E}_{\boldsymbol{\theta}_0}\left[-\frac{d^2}{d\boldsymbol{\theta}^2}\log p(y_i \mid \boldsymbol{\theta})\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \mathcal{I}(\boldsymbol{\theta}_0)$$

- In this case, the approximation of the posterior is

$$p(\sqrt{n}(\boldsymbol{\theta}-\boldsymbol{\theta}_0) \mid y) \approx \mathcal{N}\left(\boldsymbol{h} \,\middle|\, \mathbf{0}, \mathcal{I}^{-1}(\boldsymbol{\theta}_0)\right)$$

  with $\boldsymbol{h} = \sqrt{n}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)$ the **local parameter**.
- The unnormalized version is the distribution that is degenerate at $\boldsymbol{\theta}_0$.

$$p(\boldsymbol{\theta} \mid y) \approx \delta_{\boldsymbol{\theta}_0}$$

Bayes Estimator

Besides the MAP estimator, we can define a general Bayes estimator based on any loss function L.

Bayes Estimator

Besides the MAP estimator, we can define a general Bayes estimator based on any loss function L.

- $L(\theta, \delta)$ is the **loss** in utitlity when the true parameter is θ while the estimator is δ.
  - Squared loss: $L(\theta, \delta) = (\theta - \delta)^2$
  - Misclassification loss: $L(y, \hat{y}) = \mathbb{I}\{y \neq haty\}$.

Bayes Estimator

Besides the MAP estimator, we can define a general Bayes estimator based on any loss function L.

- $L(\theta, \delta)$ is the **loss** in utitlity when the true parameter is θ while the estimator is δ.
  - Squared loss: $L(\theta, \delta) = (\theta - \delta)^2$
  - Misclassification loss: $L(y, \hat{y}) = \mathbb{I}\{y \neq haty\}$.
- The **risk** of an estimator δ is given by

$$R(\theta, \delta) = \mathbb{E}_\theta[L(\theta, \delta)]$$

Bayes Estimator

Besides the MAP estimator, we can define a general Bayes estimator based on any loss function L.

- $L(\theta, \delta)$ is the **loss** in utitlity when the true parameter is θ while the estimator is δ.
  - Squared loss: $L(\theta, \delta) = (\theta - \delta)^2$
  - Misclassification loss: $L(y, \hat{y}) = \mathbb{I}\{y \neq haty\}$.
- The **risk** of an estimator δ is given by

$$R(\theta, \delta) = \mathbb{E}_\theta[L(\theta, \delta)]$$

- The **Bayes risk** of an estimator δ is

$$R(\delta) = \mathbb{E}_{p(\theta)}[R(\theta, \delta)] = \mathbb{E}[L(\theta, \delta)]$$

Bayes Estimator

Besides the MAP estimator, we can define a general Bayes estimator based on any loss function L.

- $L(\theta, \delta)$ is the **loss** in utitlity when the true parameter is θ while the estimator is δ.
  - Squared loss: $L(\theta, \delta) = (\theta - \delta)^2$
  - Misclassification loss: $L(y, \hat{y}) = \mathbb{I}\{y \neq haty\}$.
- The **risk** of an estimator δ is given by

$$R(\theta, \delta) = \mathbb{E}_\theta[L(\theta, \delta)]$$

- The **Bayes risk** of an estimator δ is

$$R(\delta) = \mathbb{E}_{p(\theta)}[R(\theta, \delta)] = \mathbb{E}[L(\theta, \delta)]$$

- The **Bayes estimator** is the estimator $\hat{\theta}$ that minimizes the Bayes risk:

$$\hat{\theta}_n = \arg\min_{\delta \in \Theta} \; R(\delta)$$

Bayes Estimator

- Note that $R(\delta) = \mathbb{E}[\mathbb{E}_{p(\theta|y)}[L(\theta, \delta) \mid y]]$

Bayes Estimator

- Note that $R(\delta) = \mathbb{E}[\mathbb{E}_{p(\theta|y)}[L(\theta, \delta) \mid y]]$
- The Bayes estimator turns our to be the conditional optimizer:

$$\hat{\theta}_n(y) = \arg\min_{\delta \in \Theta} \mathbb{E}_{p(\theta|y)}[L(\theta, \delta) \mid y] = \arg\min_{\delta \in \Theta} \int L(\theta, \delta) p(\theta \mid y) d\mu$$

Bayes Estimator

- Note that $R(\delta) = \mathbb{E}[\mathbb{E}_{p(\theta|y)}[L(\theta, \delta) \mid y]]$
- The Bayes estimator turns our to be the conditional optimizer:

$$\hat{\theta}_n(y) = \arg\min_{\delta \in \Theta} \ \mathbb{E}_{p(\theta|y)}[L(\theta, \delta) \mid y] = \arg\min_{\delta \in \Theta} \int L(\theta, \delta) p(\theta \mid y) d\mu$$

- Examples:
  - under squared loss: $\hat{\theta}_n$ is the posterior mean.
  - under absolute loss: $\hat{\theta}_n$ is the posterior median.
  - under cross entropy loss: $\hat{\theta}_n$ is the one with minimum Kullback-Leibler divergence.

Bayes Estimator

- Note that $R(\delta) = \mathbb{E}[\mathbb{E}_{p(\theta|y)}[L(\theta, \delta) \mid y]]$
- The Bayes estimator turns our to be the conditional optimizer:

$$\hat{\theta}_n(y) = \arg\min_{\delta \in \Theta} \mathbb{E}_{p(\theta|y)}[L(\theta, \delta) \mid y] = \arg\min_{\delta \in \Theta} \int L(\theta, \delta) p(\theta \mid y) d\mu$$

- Examples:
  - under squared loss: $\hat{\theta}_n$ is the posterior mean.
  - under absolute loss: $\hat{\theta}_n$ is the posterior median.
  - under cross entropy loss: $\hat{\theta}_n$ is the one with minimum Kullback-Leibler divergence.
- Do we still have the consistency result for Bayes estimators other than the MAP?

Bayes Estimator

- Note that $R(\delta) = \mathbb{E}[\mathbb{E}_{p(\theta|y)}[L(\theta, \delta) \mid y]]$
- The Bayes estimator turns our to be the conditional optimizer:

$$\hat{\theta}_n(y) = \arg\min_{\delta \in \Theta} \mathbb{E}_{p(\theta|y)}[L(\theta, \delta) \mid y] = \arg\min_{\delta \in \Theta} \int L(\theta, \delta) p(\theta \mid y) d\mu$$

- Examples:
  - under squared loss: $\hat{\theta}_n$ is the posterior mean.
  - under absolute loss: $\hat{\theta}_n$ is the posterior median.
  - under cross entropy loss: $\hat{\theta}_n$ is the one with minimum Kullback-Leibler divergence.
- Do we still have the consistency result for Bayes estimators other than the MAP? **Yes. Doob's Consistency Theorem.**

Bayes Estimator

- Note that $R(\delta) = \mathbb{E}[\mathbb{E}_{p(\theta|y)}[L(\theta, \delta) \mid y]]$
- The Bayes estimator turns our to be the conditional optimizer:

$$\hat{\theta}_n(y) = \arg\min_{\delta \in \Theta} \mathbb{E}_{p(\theta|y)}[L(\theta, \delta) \mid y] = \arg\min_{\delta \in \Theta} \int L(\theta, \delta) p(\theta \mid y) d\mu$$

- Examples:
  - under squared loss: $\hat{\theta}_n$ is the posterior mean.
  - under absolute loss: $\hat{\theta}_n$ is the posterior median.
  - under cross entropy loss: $\hat{\theta}_n$ is the one with minimum Kullback-Leibler divergence.
- Do we still have the consistency result for Bayes estimators other than the MAP? **Yes. Doob's Consistency Theorem.**
- Do we still have the normal approximation for the posterior without utilizing the MAP?

Bayes Estimator

- Note that $R(\delta) = \mathbb{E}[\mathbb{E}_{p(\theta|y)}[L(\theta, \delta) \mid y]]$
- The Bayes estimator turns our to be the conditional optimizer:

$$\hat{\theta}_n(y) = \arg\min_{\delta \in \Theta} \mathbb{E}_{p(\theta|y)}[L(\theta, \delta) \mid y] = \arg\min_{\delta \in \Theta} \int L(\theta, \delta) p(\theta \mid y) d\mu$$

- Examples:
  - under squared loss: $\hat{\theta}_n$ is the posterior mean.
  - under absolute loss: $\hat{\theta}_n$ is the posterior median.
  - under cross entropy loss: $\hat{\theta}_n$ is the one with minimum Kullback-Leibler divergence.
- Do we still have the consistency result for Bayes estimators other than the MAP?
  Yes. Doob's Consistency Theorem.
- Do we still have the normal approximation for the posterior without utilizing the MAP?
  Yes. Berstein-Von Mises Theorem.

Counter-Examples

Before we move on to the Doob's Theorem and the Berstein-Von Mises Theorem. We first look at the a few counter-examples that are related to the key assumptions so far.

Counter-Examples

Before we move on to the Doob's Theorem and the Berstein-Von Mises Theorem. We first look at the a few counter-examples that are related to the key assumptions so far.

- Unidentifiable Models:
  Only observe the values of u for

$$\begin{pmatrix} u \\ v \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right)$$

Counter-Examples

Before we move on to the Doob's Theorem and the Berstein-Von Mises Theorem. We first look at the a few counter-examples that are related to the key assumptions so far.

- Unidentifiable Models:
  Only observe the values of u for

$$\begin{pmatrix} u \\ v \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right)$$

- Non-fixed Number of Parameters:

$$y_i \sim \mathcal{N}(\theta_i, 1)$$

Counter-Examples

Before we move on to the Doob's Theorem and the Berstein-Von Mises Theorem. We first look at the a few counter-examples that are related to the key assumptions so far.

- Unidentifiable Models:
  Only observe the values of u for

$$\begin{pmatrix} u \\ v \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right)$$

- Non-fixed Number of Parameters:

$$y_i \sim \mathcal{N}(\theta_i, 1)$$

- Zero prior density at $\boldsymbol{\theta}_0$.
- Converge to the edge of the parameter space.