

430.714

Estimation Theory

Prof. Songhwai Oh

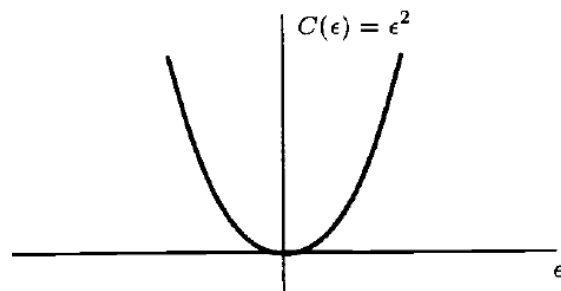
ECE, SNU

General Bayesian Estimators

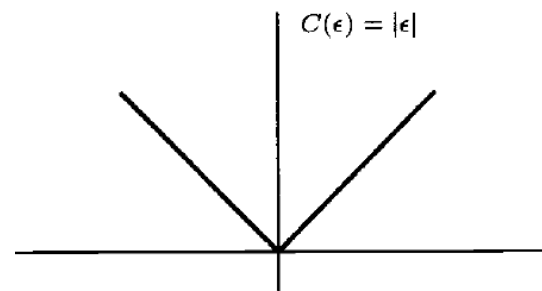
Bayes Risk

$$\mathcal{R} = E[\mathcal{C}(\epsilon)] \quad \epsilon = \theta - \hat{\theta}$$

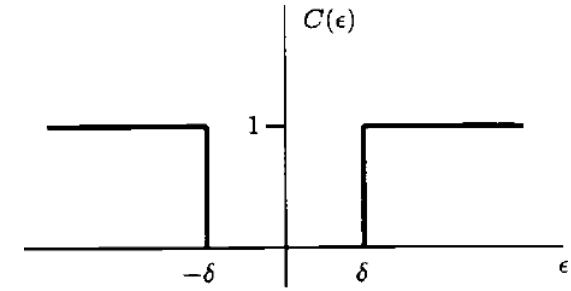
Various Cost Function



(a) Quadratic error



(b) Absolute error



(c) Hit-or-miss error

$$C(\epsilon) = \begin{cases} 0 & |\epsilon| < \delta \\ 1 & |\epsilon| > \delta \end{cases}$$

$$\begin{aligned} \mathcal{R} &= E[\mathcal{C}(\epsilon)] \\ &= \int \int \mathcal{C}(\theta - \hat{\theta}) p(\mathbf{x}, \theta) d\mathbf{x} d\theta \\ &= \int \left[\int \mathcal{C}(\theta - \hat{\theta}) p(\theta|\mathbf{x}) d\theta \right] p(\mathbf{x}) d\mathbf{x} \end{aligned}$$

Bayes Risk

Quadratic error: $C(e)=e^2$

MMSE estimator: $\hat{\theta} = E(\theta|\mathbf{x})$

Absolute error: $C(e)=|e|$

$$\begin{aligned} g(\hat{\theta}) &= \int |\theta - \hat{\theta}| p(\theta|\mathbf{x}) d\theta \\ &= \int_{-\infty}^{\hat{\theta}} (\hat{\theta} - \theta) p(\theta|\mathbf{x}) d\theta + \int_{\hat{\theta}}^{\infty} (\theta - \hat{\theta}) p(\theta|\mathbf{x}) d\theta. \end{aligned}$$

$$\begin{aligned} \frac{dg(\hat{\theta})}{d\hat{\theta}} &= \int_{-\infty}^{\hat{\theta}} p(\theta|\mathbf{x}) d\theta - \int_{\hat{\theta}}^{\infty} p(\theta|\mathbf{x}) d\theta = 0 \\ \int_{-\infty}^{\hat{\theta}} p(\theta|\mathbf{x}) d\theta &= \int_{\hat{\theta}}^{\infty} p(\theta|\mathbf{x}) d\theta \end{aligned}$$

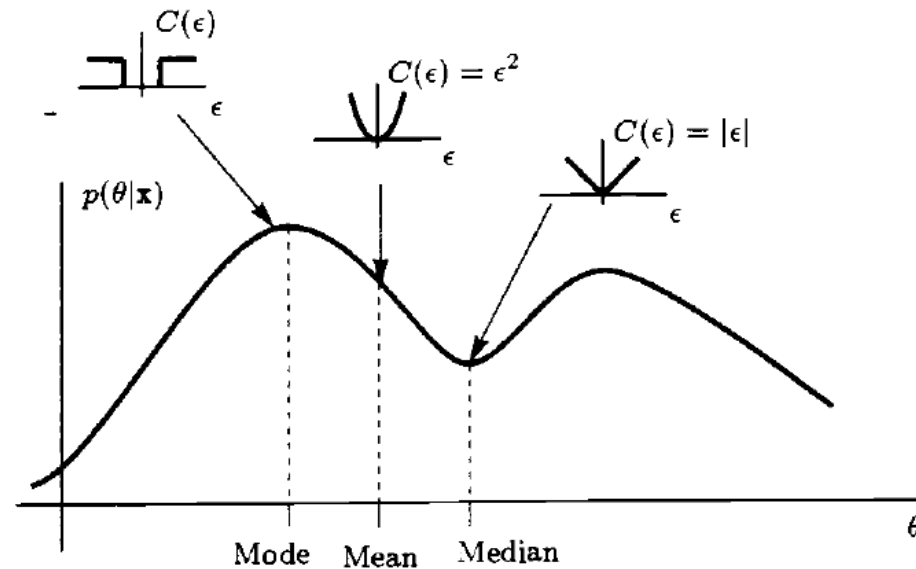
$\hat{\theta}$ is the *median* of the posterior PDF

Hit-or-miss error:

$$\begin{aligned} g(\hat{\theta}) &= \int_{-\infty}^{\hat{\theta}-\delta} 1 \cdot p(\theta|\mathbf{x}) d\theta + \int_{\hat{\theta}+\delta}^{\infty} 1 \cdot p(\theta|\mathbf{x}) d\theta \\ g(\hat{\theta}) &= 1 - \int_{\hat{\theta}-\delta}^{\hat{\theta}+\delta} p(\theta|\mathbf{x}) d\theta \end{aligned}$$

For sufficiently small δ , $\hat{\theta} = \arg \max p(\theta|\mathbf{x})$, i.e., the *mode* of the posterior PDF. This is known as the *maximum a posteriori* (MAP) estimator.

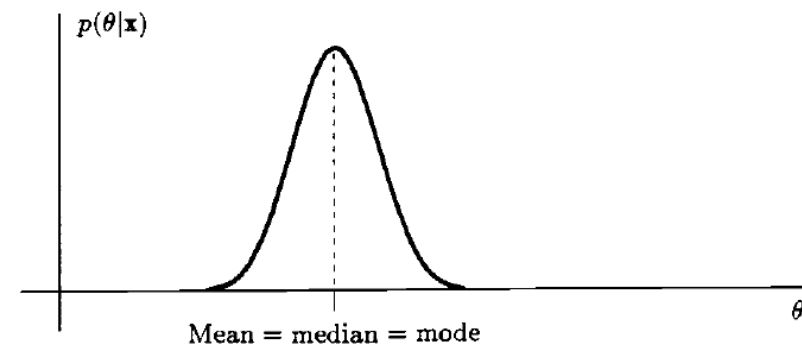
Bayes Risk



(a) General posterior PDF

	Quadratic error	Absolute error	Hit-or-miss error
Posterior's	Mean	Median	Mode

Gaussian:
$$p(\theta|\mathbf{x}) = \frac{1}{\sqrt{2\pi\sigma_{\theta|x}^2}} \exp \left[-\frac{1}{2\sigma_{\theta|x}^2} (\theta - \mu_{\theta|x})^2 \right]$$



(b) Gaussian posterior PDF

Bayesian Linear Model

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w} \quad \boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}_\theta, \mathbf{C}_\theta), \mathbf{w} \sim \mathcal{N}(0, \mathbf{C}_w)$$

$$E(\boldsymbol{\theta}|\mathbf{x}) = \boldsymbol{\mu}_\theta + \mathbf{C}_\theta \mathbf{H}^T (\mathbf{H} \mathbf{C}_\theta \mathbf{H}^T + \mathbf{C}_w)^{-1} (\mathbf{x} - \mathbf{H} \boldsymbol{\mu}_\theta) = \boldsymbol{\mu}_\theta + (\mathbf{C}_\theta^{-1} + \mathbf{H}^T \mathbf{C}_w^{-1} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{C}_w^{-1} (\mathbf{x} - \mathbf{H} \boldsymbol{\mu}_\theta)$$

$$\mathbf{C}_{\theta|x} = \mathbf{C}_\theta - \mathbf{C}_\theta \mathbf{H}^T (\mathbf{H} \mathbf{C}_\theta \mathbf{H}^T + \mathbf{C}_w)^{-1} \mathbf{H} \mathbf{C}_\theta = (\mathbf{C}_\theta^{-1} + \mathbf{H}^T \mathbf{C}_w^{-1} \mathbf{H})^{-1}$$

Matrix inversion lemma:

$$\begin{aligned} (A_{11} + A_{12} A_{22}^{-1} A_{21})^{-1} &= A_{11}^{-1} - A_{11}^{-1} A_{12} (A_{22} + A_{21} A_{11}^{-1} A_{12})^{-1} A_{21} A_{11}^{-1} \\ (A_{11} + A_{12} A_{22}^{-1} A_{21})^{-1} A_{12} A_{22}^{-1} &= A_{11}^{-1} A_{12} (A_{22} + A_{21} A_{11}^{-1} A_{12})^{-1} \end{aligned}$$

$$\mathbf{z} = [\mathbf{x}^T \boldsymbol{\theta}^T]^T$$

$$E(\mathbf{x}) = E(\mathbf{H}\boldsymbol{\theta} + \mathbf{w}) = \mathbf{H}E(\boldsymbol{\theta}) = \mathbf{H}\boldsymbol{\mu}_\theta$$

$$E(\mathbf{y}) = E(\boldsymbol{\theta}) = \boldsymbol{\mu}_\theta$$

$$\begin{aligned} \mathbf{z} &= \begin{bmatrix} \mathbf{H}\boldsymbol{\theta} + \mathbf{w} \\ \boldsymbol{\theta} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{H} & \mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\theta} \\ \mathbf{w} \end{bmatrix} \end{aligned}$$

$$\mathbf{C}_{xx} = E[(\mathbf{x} - E(\mathbf{x}))(\mathbf{x} - E(\mathbf{x}))^T]$$

$$= E[(\mathbf{H}\boldsymbol{\theta} + \mathbf{w} - \mathbf{H}\boldsymbol{\mu}_\theta)(\mathbf{H}\boldsymbol{\theta} + \mathbf{w} - \mathbf{H}\boldsymbol{\mu}_\theta)^T]$$

$$= E[(\mathbf{H}(\boldsymbol{\theta} - \boldsymbol{\mu}_\theta) + \mathbf{w})(\mathbf{H}(\boldsymbol{\theta} - \boldsymbol{\mu}_\theta) + \mathbf{w})^T]$$

$$= \mathbf{H}E[(\boldsymbol{\theta} - \boldsymbol{\mu}_\theta)(\boldsymbol{\theta} - \boldsymbol{\mu}_\theta)^T] \mathbf{H}^T + E(\mathbf{w}\mathbf{w}^T)$$

$$= \mathbf{H} \mathbf{C}_\theta \mathbf{H}^T + \mathbf{C}_w$$

$$\begin{aligned} \mathbf{C}_{yx} &= E[(\mathbf{y} - E(\mathbf{y}))(\mathbf{x} - E(\mathbf{x}))^T] \\ &= E[(\boldsymbol{\theta} - \boldsymbol{\mu}_\theta)(\mathbf{H}(\boldsymbol{\theta} - \boldsymbol{\mu}_\theta) + \mathbf{w})^T] \\ &= E[(\boldsymbol{\theta} - \boldsymbol{\mu}_\theta)(\mathbf{H}(\boldsymbol{\theta} - \boldsymbol{\mu}_\theta))^T] \\ &= \mathbf{C}_\theta \mathbf{H}^T. \end{aligned}$$

In the absence of prior knowledge,

$$\hat{\boldsymbol{\theta}} = E(\boldsymbol{\theta}|\mathbf{x}) = \boldsymbol{\mu}_\theta + (\mathbf{C}_\theta^{-1} + \mathbf{H}^T \mathbf{C}_w^{-1} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{C}_w^{-1} (\mathbf{x} - \mathbf{H} \boldsymbol{\mu}_\theta)$$

$$\mathbf{C}_\theta^{-1} \rightarrow \mathbf{0} \quad \hat{\boldsymbol{\theta}} \rightarrow (\mathbf{H}^T \mathbf{C}_w^{-1} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{C}_w^{-1} \mathbf{x} \quad (\text{MVU estimator})$$

Maximum A Posteriori (MAP) Estimators

$$\hat{\theta} = \arg \max_{\theta} p(\theta|\mathbf{x})$$

$$p(\theta|\mathbf{x}) = \frac{p(\mathbf{x}|\theta)p(\theta)}{p(\mathbf{x})}$$

$$\hat{\theta} = \arg \max_{\theta} p(\theta|\mathbf{x})$$

$$= \arg \max_{\theta} p(\mathbf{x}|\theta)p(\theta)$$

$$= \arg \max_{\theta} [\log p(\mathbf{x}|\theta) + \log p(\theta)]$$

- Easy to compute (cf. MMSE)
- Not necessarily the best for all cases

Example: Exponential PDF

$$p(x[n]|\theta) = \begin{cases} \theta \exp(-\theta x[n]) & x[n] > 0 \\ 0 & x[n] < 0 \end{cases}$$

$$p(\theta) = \begin{cases} \lambda \exp(-\lambda \theta) & \theta > 0 \\ 0 & \theta < 0 \end{cases}$$

For N i.i.d. data points, $p(\mathbf{x}|\theta) = \prod_{n=0}^{N-1} p(x[n]|\theta)$

$$g(\theta) = \ln p(\mathbf{x}|\theta) + \ln p(\theta)$$

$$= \ln \left[\theta^N \exp \left(-\theta \sum_{n=0}^{N-1} x[n] \right) \right] + \ln [\lambda \exp(-\lambda \theta)]$$

$$= N \ln \theta - N\theta \bar{x} + \ln \lambda - \lambda \theta$$

$$\frac{dg(\theta)}{d\theta} = \frac{N}{\theta} - N\bar{x} - \lambda$$

MAP estimator: $\hat{\theta} = \frac{1}{\bar{x} + \frac{\lambda}{N}}$

As $N \rightarrow \infty$, $\hat{\theta} \rightarrow 1/\bar{x}$.

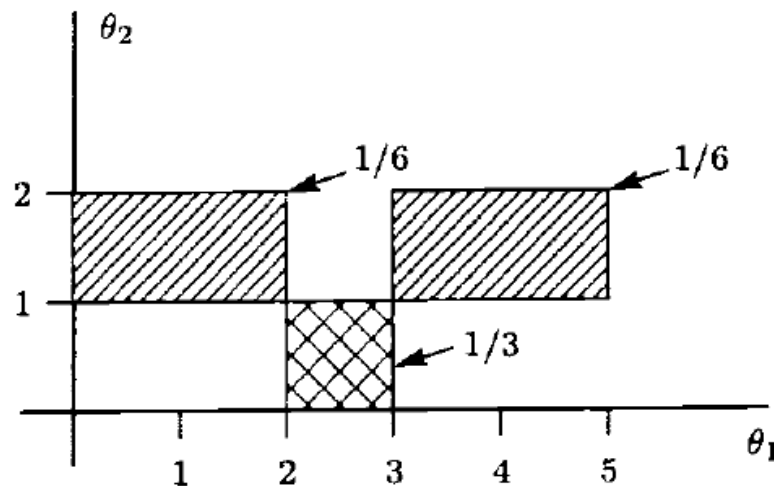
As $\lambda \rightarrow 0$, $\hat{\theta} \rightarrow 1/\bar{x}$ (MLE).

Scalar MAP Estimator vs. Vector MAP Estimator

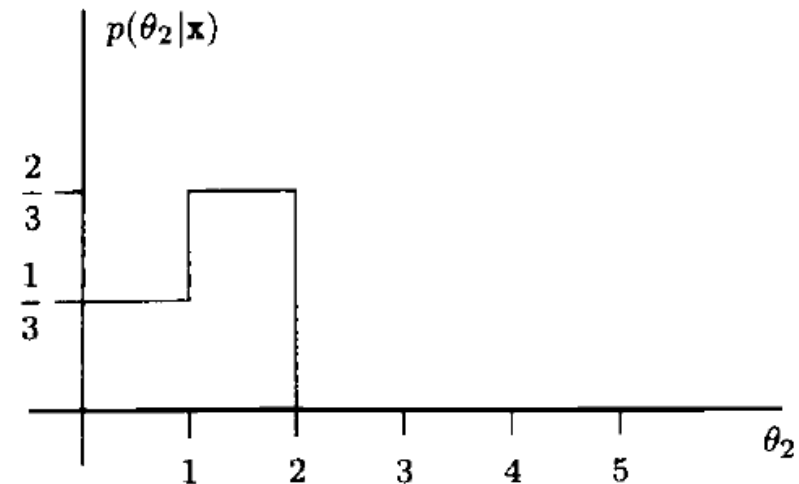
Scalar MAP estimator: $\hat{\theta}_i = \arg \max_{\theta_i} p(\theta_i | \mathbf{x}) \quad i = 1, 2, \dots, p.$

$$p(\theta_1 | \mathbf{x}) = \int \cdots \int p(\boldsymbol{\theta} | \mathbf{x}) d\theta_2 \cdots d\theta_p$$

Vector MAP estimator: $\hat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta}} p(\boldsymbol{\theta} | \mathbf{x})$



(a) Posterior PDF $p(\theta_1, \theta_2 | \mathbf{x})$



(b) Posterior PDF $p(\theta_2 | \mathbf{x})$

MAP estimator: $2 < \hat{\theta}_1 < 3, 0 < \hat{\theta}_2 < 1$

MAP estimator: $1 < \hat{\theta}_2 < 2$

Not the same !!!

Theorem 7.2 (Invariance Property of the MLE) *The MLE of the parameter $\alpha = g(\theta)$, where the PDF $p(\mathbf{x}; \theta)$ is parameterized by θ , is given by*

$$\hat{\alpha} = g(\hat{\theta})$$

where $\hat{\theta}$ is the MLE of θ . The MLE of $\hat{\theta}$ is obtained by maximizing $p(\mathbf{x}; \theta)$. If g is not a one-to-one function, then $\hat{\alpha}$ maximizes the modified likelihood function $\bar{p}_T(\mathbf{x}; \alpha)$, defined as

$$\bar{p}_T(\mathbf{x}; \alpha) = \max_{\{\theta: \alpha = g(\theta)\}} p(\mathbf{x}; \theta).$$

$$x[n] = A + w[n], \quad n = 0, 1, \dots, N-1, \quad \text{where } w[n] \sim \mathcal{N}(0, \sigma^2)$$

When $\alpha = \exp(A)$, find the MLE of α .

$$p(\mathbf{x}; A) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp \left[-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - A)^2 \right] \quad -\infty < A < \infty$$

$$p_T(\mathbf{x}; \alpha) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp \left[-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - \ln \alpha)^2 \right] \quad \alpha > 0$$

$$\dots \sum_{n=0}^{N-1} (x[n] - \ln \hat{\alpha}) \frac{1}{\hat{\alpha}} = 0 \quad \hat{\alpha} = \exp(\bar{x}) \quad \hat{\alpha} = \exp(\hat{A}) = \exp(\hat{\theta})$$

Invariance Property Does Not Hold for the MAP Estimator

Example: Exponential PDF

$$p(x[n]|\theta) = \begin{cases} \theta \exp(-\theta x[n]) & x[n] > 0 \\ 0 & x[n] < 0 \end{cases}$$

$$p(\theta) = \begin{cases} \lambda \exp(-\lambda \theta) & \theta > 0 \\ 0 & \theta < 0 \end{cases}$$

For N iid data pts, $p(\mathbf{x}|\theta) = \prod_{n=0}^{N-1} p(x[n]|\theta)$

MAP estimator: $\hat{\theta} = \frac{1}{\bar{x} + \frac{\lambda}{N}}$

Suppose we want to estimate $\alpha = 1/\theta$. $\hat{\alpha} = \frac{1}{\hat{\theta}} \quad \hat{\alpha} = \bar{x} + \frac{\lambda}{N} \quad ???$

$$p(x[n]|\alpha) = \begin{cases} \frac{1}{\alpha} \exp\left(-\frac{x[n]}{\alpha}\right) & x[n] > 0 \\ 0 & x[n] < 0 \end{cases}$$

$$p_{\alpha}(\alpha) = \frac{p_{\theta}(\theta(\alpha))}{\left|\frac{d\alpha}{d\theta}\right|} = \begin{cases} \frac{\lambda \exp(-\lambda/\alpha)}{\alpha^2} & \alpha > 0 \\ 0 & \alpha < 0 \end{cases}$$

$$g(\alpha) = \ln p(\mathbf{x}|\alpha) + \ln p(\alpha)$$

$$= \ln \left[\left(\frac{1}{\alpha}\right)^N \exp\left(-\frac{1}{\alpha} \sum_{n=0}^{N-1} x[n]\right) \right] + \ln \frac{\lambda \exp(-\lambda/\alpha)}{\alpha^2}$$

$$= -N \ln \alpha - N \frac{\bar{x}}{\alpha} + \ln \lambda - \frac{\lambda}{\alpha} - 2 \ln \alpha$$

$$= -(N+2) \ln \alpha - \frac{N\bar{x} + \lambda}{\alpha} + \ln \lambda.$$

$$\frac{dg}{d\alpha} = -\frac{N+2}{\alpha} + \frac{N\bar{x} + \lambda}{\alpha^2}$$

MAP estimator: $\hat{\alpha} = \frac{N\bar{x} + \lambda}{N+2}$

Performance Description

- $p(\hat{\theta}|\theta)$, different PDF of the estimator for each realization of θ .
- Good estimator: close to θ for every possible value of θ , i.e., $\epsilon = \theta - \hat{\theta}$ is small.
- Good estimator $\hat{\theta}$: $p(\hat{\theta}|\theta)$ should be concentrated about θ for all possible θ .
- Such estimator is the MMSE estimator $\hat{\theta} = \mathbb{E}(\theta|\mathbf{x})$, which minimizes $\mathbb{E}_{x,\theta}((\theta - \hat{\theta})^2)$.

$$\epsilon = \theta - E(\theta|\mathbf{x})$$

$$\begin{aligned} E_{x,\theta}(\epsilon) &= E_{x,\theta}[\theta - E(\theta|\mathbf{x})] \\ &= E_x[E_{\theta|x}(\theta) - E_{\theta|x}(\theta|\mathbf{x})] \\ &= E_x[E(\theta|\mathbf{x}) - E(\theta|\mathbf{x})] = 0 \end{aligned}$$

$$\begin{aligned} \text{var}(\epsilon) &= E_{x,\theta}(\epsilon^2) \\ &= E_{x,\theta}[(\theta - \hat{\theta})^2] \end{aligned}$$

Bayesian MSE

if ϵ is Gaussian, we have that $\epsilon \sim \mathcal{N}(0, \text{Bmse}(\hat{\theta}))$

DC level in WGN – Gaussian prior

$$\hat{A} = \frac{\sigma_A^2}{\sigma_A^2 + \frac{\sigma^2}{N}} \bar{x} + \frac{\frac{\sigma^2}{N}}{\sigma_A^2 + \frac{\sigma^2}{N}} \mu_A$$

$$\text{Bmse}(\hat{A}) = \frac{1}{\frac{N}{\sigma^2} + \frac{1}{\sigma_A^2}}$$

$$\epsilon \sim \mathcal{N}\left(0, \frac{1}{\frac{N}{\sigma^2} + \frac{1}{\sigma_A^2}}\right)$$

Consistent in the Bayesian sense, i.e., for large data the estimator will always be close to the realization of the parameter.