ECE 275A: Parameter Estimation I
Bayesian Estimation – Part I

Florian Meyer

Electrical and Computer Engineering Department
University of California San Diego

UC San Diego
JACOBS SCHOOL OF ENGINEERING

Loss Functions

- Loss functions $\ell\left(\hat{\boldsymbol{\theta}}(\boldsymbol{y}), \boldsymbol{\theta}\right)$ have the following properties

$$\ell\left(\hat{\boldsymbol{\theta}}(\boldsymbol{y}), \boldsymbol{\theta}\right) \geq 0 \quad \text{with} \quad \ell\left(\hat{\boldsymbol{\theta}}(\boldsymbol{y}), \boldsymbol{\theta}\right) = 0 \Leftrightarrow \hat{\boldsymbol{\theta}}(\boldsymbol{y}) = \boldsymbol{\theta}$$

- Possible choices of $\ell(\hat{\boldsymbol{\theta}}(\boldsymbol{y}), \boldsymbol{\theta})$ include

1. Quadratic error: $\ell(\hat{\boldsymbol{\theta}}(\boldsymbol{y}), \boldsymbol{\theta}) = \|\hat{\boldsymbol{\theta}}(\boldsymbol{y}) - \boldsymbol{\theta}\|^2$

2. Hit-or-miss error: $\ell(\hat{\boldsymbol{\theta}}(\boldsymbol{y}), \boldsymbol{\theta}) = \lim_{\delta \to 0} \ell_\delta(\hat{\boldsymbol{\theta}}(\boldsymbol{y}), \boldsymbol{\theta})$ with

$$\ell_\delta(\hat{\boldsymbol{\theta}}(\boldsymbol{y}), \boldsymbol{\theta}) = \begin{cases} 0 & \|\hat{\boldsymbol{\theta}}(\boldsymbol{y}) - \boldsymbol{\theta}\| \leq \delta \\ 1 & \text{others} \end{cases}$$

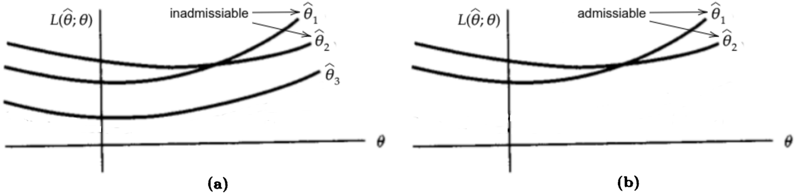
3. Absolute error: $\ell(\hat{\boldsymbol{\theta}}(\boldsymbol{y}), \boldsymbol{\theta}) = \|\hat{\boldsymbol{\theta}}(\boldsymbol{y}) - \boldsymbol{\theta}\|_1$

Average Loss and Admissibility of Estimators

- The average loss is given by $L(\hat{\boldsymbol{\theta}}; \boldsymbol{\theta}) = E_{\mathbf{y}}\big(\ell(\hat{\boldsymbol{\theta}}(\mathbf{y}), \boldsymbol{\theta})\big) \geq 0$
- If the quadratic error is used, the average loss is the mean square error $L(\hat{\boldsymbol{\theta}}; \boldsymbol{\theta}) = \text{mse}_{\boldsymbol{\theta}}\big(\hat{\boldsymbol{\theta}}(\mathbf{y})\big)$
- An estimator $\hat{\boldsymbol{\theta}}(\mathbf{y})$ is inadmissible if another estimator $\hat{\boldsymbol{\theta}}'(\mathbf{y})$ never has greater average loss than $\hat{\boldsymbol{\theta}}(\mathbf{y})$ but sometimes has strictly lower average loss
- In other words, $\hat{\boldsymbol{\theta}}(\mathbf{y})$ is inadmissible if there exists an $\hat{\boldsymbol{\theta}}'(\mathbf{y})$ such that $L(\hat{\boldsymbol{\theta}}'; \boldsymbol{\theta}) \leqslant L(\hat{\boldsymbol{\theta}}; \boldsymbol{\theta})$ for all $\boldsymbol{\theta}$ and $L(\hat{\boldsymbol{\theta}}'; \boldsymbol{\theta}') < L(\hat{\boldsymbol{\theta}}; \boldsymbol{\theta}')$ for some $\boldsymbol{\theta}'$

Admissible Estimators – Classical Frequentist Statistics

- In classical frequentist statistics $\boldsymbol{\theta}$ is deterministic
- To find the optimal estimator we need to find all admissible estimators
- Recall that an estimator $\hat{\boldsymbol{\theta}}$ that is optimal $\forall \boldsymbol{\theta}$ is typically not realizable



- In (a), estimator $\hat{\boldsymbol{\theta}}_3$ is admissible and optimal for all $\boldsymbol{\theta}$
- In (b), both $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$ are admissible, and there is no estimator that is optimal for all $\boldsymbol{\theta}$

Bayesian Statistics

- In Bayesian statistics, both $\mathbf{y}$ and $\boldsymbol{\theta}$ are random
- The joint statistics $p(\boldsymbol{y}, \boldsymbol{\theta}) = p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ are assumed known
- The Bayes loss $L(\hat{\boldsymbol{\theta}})$ is defined as

$$\begin{aligned} L(\hat{\boldsymbol{\theta}}) &= E_{\mathbf{y},\boldsymbol{\theta}}\big(\ell(\hat{\boldsymbol{\theta}}(\mathbf{y}), \boldsymbol{\theta})\big) \\ &= E_{\boldsymbol{\theta}}\Big(E_{\mathbf{y}|\boldsymbol{\theta}}\big(\ell(\hat{\boldsymbol{\theta}}(\mathbf{y}), \theta)\,\big|\,\boldsymbol{\theta} = \theta\big)\Big) \\ &= E_{\boldsymbol{\theta}}\big(L(\hat{\boldsymbol{\theta}}|\theta)\big) \\ &= \int L(\hat{\boldsymbol{\theta}}|\theta)\,p(\theta)\,\mathrm{d}\theta \end{aligned}$$

  where $L(\hat{\boldsymbol{\theta}}|\theta) = E_{\mathbf{y}|\boldsymbol{\theta}}\big(\ell(\hat{\boldsymbol{\theta}}(\mathbf{y}), \theta)\,\big|\,\boldsymbol{\theta} = \theta\big)$

- The Bayes optimal estimator minimizes the Bayes loss, i.e., $\hat{\boldsymbol{\theta}}_* = \arg\min_{\hat{\boldsymbol{\theta}}} L(\hat{\boldsymbol{\theta}})$.

Bayesian Statistics

- The Bayes optimal estimation has the following two properties:
  1. $\forall\, p(\theta), \exists$ an admissible $\hat{\boldsymbol{\theta}}$ that is Bayes optimal
  2. $\forall$ admissible $\hat{\boldsymbol{\theta}}, \exists\, p(\theta)$ for which $\hat{\boldsymbol{\theta}}$ is Bayes optimal
- Note that $\hat{\boldsymbol{\theta}}_* = \arg\min_{\hat{\boldsymbol{\theta}}} L(\hat{\boldsymbol{\theta}})$ is an optimization in function space
- How to do this tractably?
- $L(\hat{\boldsymbol{\theta}})$ can also be written as

$$\begin{aligned} L(\hat{\boldsymbol{\theta}}) &= E_{\mathbf{y},\boldsymbol{\theta}}\big(\ell(\hat{\boldsymbol{\theta}}(\mathbf{y}), \boldsymbol{\theta})\big) \\ &= E_{\mathbf{y}}\Big(E_{\boldsymbol{\theta}|\mathbf{y}}\big(\ell(\hat{\boldsymbol{\theta}}(\boldsymbol{y}), \boldsymbol{\theta})|\mathbf{y} = \boldsymbol{y}\big)\Big) \\ &= E_{\mathbf{y}}\big(L(\hat{\boldsymbol{\theta}}|\boldsymbol{y})\big) \\ &= \int L(\hat{\boldsymbol{\theta}}|\boldsymbol{y})p(\boldsymbol{y})d\boldsymbol{y} \end{aligned}$$

where $L(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) = E_{\boldsymbol{\theta}|\mathbf{y}}\big(\ell(\hat{\boldsymbol{\theta}}(\boldsymbol{y}), \theta)|\mathbf{y} = \boldsymbol{y}\big)$

$$L(\hat{\boldsymbol{\theta}}) = E_{\boldsymbol{\theta}}\big(L(\hat{\boldsymbol{\theta}}|\theta)\big) = E_{\mathbf{y}}\big(L(\hat{\boldsymbol{\theta}}|\boldsymbol{y})\big)$$

$$L(\hat{\boldsymbol{\theta}}) = \int \underbrace{L(\hat{\boldsymbol{\theta}}|\boldsymbol{\theta})}_{\geq 0}\,\underbrace{p(\boldsymbol{\theta})}_{\geq 0}\,\mathrm{d}\boldsymbol{\theta} = \int \underbrace{L(\hat{\boldsymbol{\theta}}|\boldsymbol{y})}_{\geq 0}\,\underbrace{p(\boldsymbol{y})}_{\geq 0}\,\mathrm{d}\boldsymbol{y}$$

- Either minimizing $L(\hat{\boldsymbol{\theta}}|\theta)$ or $L(\hat{\boldsymbol{\theta}}|\boldsymbol{y})$ minimizes $L(\hat{\boldsymbol{\theta}})$
- Recall that $\hat{\boldsymbol{\theta}}_* = \arg\min_{\hat{\boldsymbol{\theta}}} L(\hat{\boldsymbol{\theta}})$ requires optimization in function space, but if we know the value of $\hat{\boldsymbol{\theta}}_*(\boldsymbol{y})$ for each $\boldsymbol{y}$, then we know the function $\hat{\boldsymbol{\theta}}(\cdot)$
- Let $\hat{\boldsymbol{\theta}}_*(\boldsymbol{y}) = \arg\min_{\hat{\boldsymbol{\theta}}} L(\hat{\boldsymbol{\theta}}|\boldsymbol{y})$, which is a regular vector optimization for each $\boldsymbol{y} \in \mathbb{Y}$ and $\hat{\boldsymbol{\theta}}_*(\boldsymbol{y}) \in \mathbb{R}^p$, whereas $\hat{\boldsymbol{\theta}}_* \in \left\{f : \mathbb{Y} \subset \mathbb{R}^m \to \mathbb{R}^p\right\}$

Quadratic Error

- If the loss function is quadratic error (i.e., $\ell\big(\hat{\theta}(y), \theta\big) = \|\hat{\theta}(y) - \theta\|^2$), the Bayes loss $L(\hat{\boldsymbol{\theta}})$ becomes the Bayesian mean square error, i.e.,

$$\begin{aligned}\text{bmse}(\hat{\boldsymbol{\theta}}) &= E_{\mathbf{y},\boldsymbol{\theta}}\big(\|(\hat{\boldsymbol{\theta}}(\mathbf{y}) - \boldsymbol{\theta}\|^2\big) \\ &= E_{\mathbf{y},\boldsymbol{\theta}}\big(\|\tilde{\boldsymbol{\theta}}(\mathbf{y})\|^2\big) \\ &= E_{\mathbf{y}}\Big(\underbrace{E_{\boldsymbol{\theta}|\mathbf{y}}\big(\|\tilde{\boldsymbol{\theta}}(y)\|^2|\mathbf{y} = y\big)}_{L(\hat{\boldsymbol{\theta}}|y)}\Big)\end{aligned}$$

- We want $\hat{\boldsymbol{\theta}}_*(y) = \arg\min_{\hat{\boldsymbol{\theta}}} L(\hat{\boldsymbol{\theta}}|y) = \arg\min_{\hat{\theta}} E_{\boldsymbol{\theta}|\mathbf{y}}\big(\|\tilde{\boldsymbol{\theta}}(y)\|^2|\mathbf{y} = y\big)$

Quadratic Error

- To minimize $L(\hat{\boldsymbol{\theta}}|\boldsymbol{y})$, we take the derivative with respect to $\hat{\boldsymbol{\theta}}$ and set it to 0,

$$\begin{aligned}\nabla_{\hat{\boldsymbol{\theta}}} L(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) &= E_{\boldsymbol{\theta}|\mathbf{y}}\left(\nabla_{\hat{\boldsymbol{\theta}}} \|(\tilde{\boldsymbol{\theta}}(\boldsymbol{y})\|^2 | \mathbf{y} = \boldsymbol{y}\right) \\ &= 2E_{\boldsymbol{\theta}|\mathbf{y}}\left(\hat{\boldsymbol{\theta}}(\boldsymbol{y}) - \boldsymbol{\theta}|\mathbf{y} = \boldsymbol{y}\right) \stackrel{\text{set}}{=} 0\end{aligned}$$

  We get $\hat{\boldsymbol{\theta}}_*(\boldsymbol{y}) = E_{\boldsymbol{\theta}|\mathbf{y}}\left(\boldsymbol{\theta}|\boldsymbol{y}\right)$

- The Hessian $\nabla^2_{\hat{\boldsymbol{\theta}}} L(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) = 2\boldsymbol{I} \Rightarrow \hat{\boldsymbol{\theta}}_*(\boldsymbol{y})$ is optimal

- Besides $E_{\mathbf{y}}(\hat{\boldsymbol{\theta}}_*(\boldsymbol{y})) = E_{\mathbf{y}}\left(E_{\boldsymbol{\theta}|\mathbf{y}}\left(\boldsymbol{\theta}|\boldsymbol{y}\right)\right) = E_{\boldsymbol{\theta}}(\boldsymbol{\theta})$, hence it is unbiased in a Bayesian sense

Example: Quadratic Error

- Consider a scalar observation $y = x + n$, where $x \sim \mathcal{N}(0, \sigma_x^2)$ is the unknown parameter and $n \sim \mathcal{N}(0, \sigma_n^2)$
- It is assumed that x and n are independent $\Rightarrow p(y|x) = \mathcal{N}(x, \sigma_n^2)$
- To find the minimum mean square error (MMSE) estimator $\hat{x}_{\text{MMSE}} = E(x|y)$, we first calculate the posterior distribution $p(x|y)$

$$\begin{aligned}
p(x|y) &\propto p(y|x)p(x) \\
&\propto \exp\left(-\frac{(y-x)^2}{2\sigma_n^2} - \frac{x^2}{2\sigma_x^2}\right) \\
&= \exp\left(-\frac{\sigma_x^2 + \sigma_n^2}{2\sigma_x^2\sigma_n^2}x^2 + \frac{1}{\sigma_n^2}yx - \frac{1}{2\sigma_n^2}y^2\right) \\
&\propto \mathcal{N}\left(\frac{\sigma_x^2}{\sigma_x^2 + \sigma_n^2}y, \frac{\sigma_x^2\sigma_n^2}{\sigma_x^2 + \sigma_n^2}\right)
\end{aligned}$$

- Thus $\hat{x}_{\text{MMSE}} = E(x|y) = \frac{\sigma_x^2}{\sigma_x^2+\sigma_n^2}y$, which is different from the ML estimator $\hat{x}_{\text{ML}} = \arg\max_x p(y|x) = y$.