UNIT 1

Estimation Theory

Johannes Brandstetter
Institute for Machine Learning

JOHANNES KEPLER
UNIVERSITY LINZ

Institute for
Machine Learning

Copyright statement:

Basics of Estimation Theory

- Given: Data distribution $p(\boldsymbol{x}, \boldsymbol{w})$ parameterized by $\boldsymbol{w} \in \mathbb{R}^d$:
  - Remember $\mathbf{Z} = \begin{pmatrix} \mathbf{X} \\ \boldsymbol{y}^T \end{pmatrix}$
  - For unsupervised learning: $\mathbf{Z} = \mathbf{X}$
- Goal: Estimate $\boldsymbol{w}$ by $\hat{\boldsymbol{w}}$, given training data $\mathbf{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_l)$ so that distribution of $\mathbf{X}$ is reflected accurately.
- Unbiased estimator: $\mathrm{E}_{\mathbf{X}}(\hat{\boldsymbol{w}}) = \boldsymbol{w}$ for all possible $\boldsymbol{w}$
  - $\mathrm{E}_{\mathbf{X}}(\cdot)$ is the expectation wrt. data distribution $p(\boldsymbol{x}, \boldsymbol{w})$
  - Bias: $b(\hat{\boldsymbol{w}}) = \mathrm{E}_{\mathbf{X}}(\hat{\boldsymbol{w}}) - \boldsymbol{w}$
  - Variance: $\mathrm{var}(\hat{\boldsymbol{w}}) = \mathrm{E}_{\mathbf{X}}((\hat{\boldsymbol{w}} - \mathrm{E}_{\mathbf{X}}(\hat{\boldsymbol{w}}))^T(\hat{\boldsymbol{w}} - \mathrm{E}_{\mathbf{X}}(\hat{\boldsymbol{w}})))$
- For the rest of presentation: $\boldsymbol{w}$: vector valued, w: scalar valued ($d = 1$).

- We define the MSE as expectation over the training set:

$$
\begin{aligned}
\mathrm{mse}(\hat{\boldsymbol{w}}) &= \mathrm{E}_{\mathbf{X}}\left((\hat{\boldsymbol{w}} - \boldsymbol{w})^T (\hat{\boldsymbol{w}} - \boldsymbol{w}) \right) \\
&= \mathrm{E}_{\mathbf{X}}\Big(\big((\hat{\boldsymbol{w}} - \mathrm{E}_{\mathbf{X}}(\hat{\boldsymbol{w}})) + (\mathrm{E}_{\mathbf{X}}(\hat{\boldsymbol{w}}) - \boldsymbol{w})\big)^T \\
&\qquad \big((\hat{\boldsymbol{w}} - \mathrm{E}_{\mathbf{X}}(\hat{\boldsymbol{w}})) + (\mathrm{E}_{\mathbf{X}}(\hat{\boldsymbol{w}}) - \boldsymbol{w})\big) \Big) \\
&= \mathrm{E}_{\mathbf{X}}\Big(\big(\hat{\boldsymbol{w}} - \mathrm{E}_{\mathbf{X}}(\hat{\boldsymbol{w}})\big)^T \big(\hat{\boldsymbol{w}} - \mathrm{E}_{\mathbf{X}}(\hat{\boldsymbol{w}})\big) \\
&\qquad + 2\big(\underbrace{\hat{\boldsymbol{w}} - \mathrm{E}_{\mathbf{X}}(\hat{\boldsymbol{w}})}_{=0}\big)^T \big(\mathrm{E}_{\mathbf{X}}(\hat{\boldsymbol{w}}) - \boldsymbol{w}\big) \\
&\qquad + \big(\mathrm{E}_{\mathbf{X}}(\hat{\boldsymbol{w}}) - \boldsymbol{w}\big)^T \big(\mathrm{E}_{\mathbf{X}}(\hat{\boldsymbol{w}}) - \boldsymbol{w}\big) \Big) \\
&= \underbrace{\mathrm{E}_{\mathbf{X}}\Big(\big(\hat{\boldsymbol{w}} - \mathrm{E}_{\mathbf{X}}(\hat{\boldsymbol{w}})\big)^T \big(\hat{\boldsymbol{w}} - \mathrm{E}_{\mathbf{X}}(\hat{\boldsymbol{w}})\big) \Big)}_{\mathrm{var}(\hat{\boldsymbol{w}})} \\
&\qquad + \underbrace{\big(\mathrm{E}_{\mathbf{X}}(\hat{\boldsymbol{w}}) - \boldsymbol{w}\big)^T \big(\mathrm{E}_{\mathbf{X}}(\hat{\boldsymbol{w}}) - \boldsymbol{w}\big)}_{b^2(\hat{\boldsymbol{w}})}
\end{aligned}
$$

Example MSE of estimator

- Example: $x_i = \varepsilon_i$, $\varepsilon_i \sim \mathcal{N}(\mu, \sigma^2)$ i.i.d. (x_i is scalar-valued for simplicity)
- We consider the estimator

$$\hat{\mu} = \frac{\lambda}{l} \sum_{i=1}^{l} x_i$$

- Bias $b^2(\hat{\boldsymbol{w}})$ and variance $\mathrm{var}(\hat{\boldsymbol{w}})$ can be calculated as:

$$b^2(\hat{\mu}) = \left(\mathrm{E}_{\mathbf{x}}(\hat{\mu}) - \mu\right)^2 = \mu^2(\lambda - 1)^2$$

$$\begin{aligned}
\mathrm{var}(\hat{\mu}) &= \mathrm{E}_{\mathbf{x}}\left((\hat{\mu} - \mathrm{E}_{\mathbf{x}}(\hat{\mu}))^T(\hat{\mu} - \mathrm{E}_{\mathbf{x}}(\hat{\mu}))\right) \\
&= \mathrm{E}_{\mathbf{x}}\left((\frac{\lambda}{l}\sum_{i=1}^{l} x_i - \frac{\lambda}{l}\sum_{i=1}^{l}\mathrm{E}_{\mathbf{x}}(x_i))^T (\frac{\lambda}{l}\sum_{i=1}^{l} x_i - \frac{\lambda}{l}\sum_{i=1}^{l}\mathrm{E}_{\mathbf{x}}(x_i)) \right) \\
&= \mathrm{E}_{\mathbf{x}}\left(\frac{\lambda^2}{l^2} (\sum_{i=1}^{l} x_i - \sum_{i=1}^{l}\mathrm{E}_{\mathbf{x}}(x_i))^T (\sum_{i=1}^{l} x_i - \sum_{i=1}^{l}\mathrm{E}_{\mathbf{x}}(x_i)) \right) \\
&= \frac{\lambda^2}{l^2} \cdot l\sigma^2 = \frac{\lambda^2\sigma^2}{l}
\end{aligned}$$

- Often: multiple estimates of the same parameters, i.e., $\{\hat{w}_1, \ldots, \hat{w}_n\}$. Then they are combined to: $\hat{w} = \frac{1}{n} \sum_{j=1}^{n} \hat{w}_j$
- For unbiased and uncorrelated $\hat{w}_j$, that means:

$$\mathrm{E}_{\mathbf{X}}(\hat{w}) = w$$

$$\mathrm{Var}(\hat{w}) = \frac{1}{n^2} \sum_{j=1}^{n} \mathrm{Var}(\hat{w}_j) = \frac{\mathrm{Var}(\hat{w})}{n}$$

- The more estimates are averaged, the more the variance decreases
- But if $\mathrm{E}_{\mathbf{X}}(\hat{w}_j) = w + b^2(\hat{w}_j)$, which implies $\mathrm{E}_{\mathbf{X}}(\hat{w}) = w + b^2(\hat{w}_j)$, no matter how many estimates are averaged, $\hat{w}$ does not converge to the true w

Example unbiased estimator

- Example: $x_i = \varepsilon_i$, $\varepsilon_i \sim \mathcal{N}(\mu, \sigma^2)$ i.i.d. (x_i is scalar-valued for simplicity)
- Then, $\boldsymbol{w} = \begin{pmatrix} \mu \\ \sigma \end{pmatrix}$
- We estimate $\hat{\mu}$ as:

$$\hat{\mu} = \frac{1}{l} \sum_{i=1}^{l} x_i, \qquad \mathrm{E}(\hat{\mu}) = \frac{1}{l} \sum_{i=1}^{l} \mathrm{E}(x_i) = \mu$$

- For any value of μ, $\mathrm{E}_{\mathbf{X}}(\hat{\mu}) = \mu$

  $\Rightarrow \hat{\mu} = \frac{1}{l} \sum_{i=1}^{l} x_i$ is **(asymptotic) unbiased estimator** of μ

Example biased estimator

- Example: $x_i = \varepsilon_i$, $\varepsilon_i \sim \mathcal{N}(\mu, \sigma^2)$ i.i.d. (x_i is scalar-valued for simplicity)
- Then, $\boldsymbol{w} = \begin{pmatrix} \mu \\ \sigma \end{pmatrix}$
- We estimate $\hat{\mu}$ as:

$$\hat{\mu} = \frac{1}{2l} \sum_{i=1}^{l} x_i, \qquad \mathrm{E}(\hat{\mu}) = \frac{1}{2l} \sum_{i=1}^{l} \mathrm{E}(x_i) = \frac{1}{2}\mu$$

- For any other value than $\mu = 0$, $\mathrm{E}_{\mathbf{X}}(\hat{\mu}) \neq \mu$

  $\Rightarrow \hat{\mu} = \frac{1}{2l} \sum_{i=1}^{l} x_i$ is **biased estimator** of μ

Example biased estimator

- Example: $x_i = \varepsilon_i$, $\varepsilon_i \sim \mathcal{N}(\mu, \sigma^2)$ i.i.d. (x_i is scalar-valued)
- Then, $\hat{\sigma^2} = \frac{1}{l}\sum_{i=1}^{l}(x_i - \hat{\mu})^2$ with $\hat{\mu} = \frac{1}{l}\sum_{i=1}^{l} x_i$ is a **biased estimator**:

$$
\begin{aligned}
\mathrm{E}_{\mathbf{X}}(\hat{\sigma^2}) &= \mathrm{E}_{\mathbf{X}}\left(\frac{1}{l}\sum_{i=1}^{l}(x_i - \hat{\mu})^2\right) \\
&= \frac{1}{l}\mathrm{E}_{\mathbf{X}}\left(\sum_{i=1}^{l} x_i^2 - \sum_{i=1}^{l} 2x_i\hat{\mu} + \sum_{i=1}^{l}\hat{\mu}^2\right) \\
&= \frac{1}{l}\mathrm{E}_{\mathbf{X}}\left(\sum_{i=1}^{l} x_i^2 - 2\hat{\mu}\sum_{i=1}^{l} x_i + l\hat{\mu}^2\right) \\
&= \frac{1}{l}\mathrm{E}_{\mathbf{X}}\left(\sum_{i=1}^{l} x_i^2 - l\hat{\mu}^2\right) \\
&= \frac{1}{l}\sum_{i=1}^{l}\mathrm{E}_{\mathbf{X}}(x_i^2) - \frac{1}{l}\sum_{i=1}^{l} l\mathrm{E}_{\mathbf{X}}(\hat{\mu}^2)
\end{aligned}
$$

Example biased estimator

- Using the following properties:

$$\text{var}(x_i) = \mathrm{E}_{\mathbf{x}}(x_i^2) - [\mathrm{E}_{\mathbf{x}}(x_i)]^2 \Rightarrow \mathrm{E}_{\mathbf{x}}(x_i^2) = \sigma^2 + \mu^2$$

$$\text{var}(\hat{\mu}) = \mathrm{E}_{\mathbf{x}}(\hat{\mu}^2) - [\mathrm{E}_{\mathbf{x}}(\hat{\mu})]^2 \Rightarrow \mathrm{E}_{\mathbf{x}}(\hat{\mu}^2) = \frac{\sigma^2}{l} + \mu^2$$

- We get:

$$\begin{aligned}
\mathrm{E}_{\mathbf{x}}(\hat{\sigma^2}) &= \frac{1}{l}\sum_{i=1}^{l}\mathrm{E}_{\mathbf{x}}(x_i^2) - \frac{1}{l}\sum_{i=1}^{l} l\mathrm{E}_{\mathbf{x}}(\hat{\mu}^2) \\
&= \frac{1}{l}\left(l\sigma^2 + l\mu^2 - \sigma^2 - l\mu^2\right) \\
&= \frac{(l-1)}{l}\sigma^2 \\
&\neq \sigma^2
\end{aligned}$$

- The unbiased estimator of the sample variance is

$$\hat{\sigma^2} = \frac{1}{l-1}\sum_{i=1}^{l}(x_i - \hat{\mu})^2$$

Quality criterion, MVU estimator

- Mean squared error as quality criterion:

$$\mathrm{mse}(\hat{\boldsymbol{w}}) = \mathrm{var}(\hat{\boldsymbol{w}}) \; + \; b^2(\hat{\boldsymbol{w}})$$

- Goal: find estimator with minimal mse
- However, if bias term – and thus optimal value – depend on $\boldsymbol{w}$, this goal is not realizable
- $\Rightarrow$ Constrain bias term to be zero and find estimators that minimize variance $\Rightarrow$ **MVU (minimal variance unbiased)** estimator
- Does MVU estimator exist? In general: no

How to find MVU?

- Even if MVU estimator exists, there is no known procedure which always produces it, but we can do the following:
- We place a lower bound on the variance of all unbiased estimators. If one estimator attains this bound for all values of the parameter $\Rightarrow$ MVU estimator
- This bound is called **Cramér-Rao lower bound (CRLB)**

Maximum Likelihood Estimator (MLE)

- MVU estimator often unknown or does not exist $\Rightarrow$ How to construct estimator that approximates MVU estimator?
- This estimator should be **asymptotically efficient** and **asymptotically unbiased**
- **Likelihood** of i.i.d. data $\mathbf{X}$:

$$\mathcal{L}(\mathbf{X}; \boldsymbol{w}) = \prod_{i=1}^{l} p(\boldsymbol{x}_i; \boldsymbol{w})$$

- I.e., probability of the (data-distribution) model $p(\boldsymbol{x}; \boldsymbol{w})$ to (re-)produce the dataset
- We want to maximize the likelihood, or equivalently minimize the **negative log-likelihood**:

$$L = -\ln(\mathcal{L}) = -\sum_{i=1}^{l} \ln(p(\boldsymbol{x}_i; \boldsymbol{w}))$$

MLE of normal distribution

- Example: $x_i = \varepsilon_i$, $\varepsilon_i \sim \mathcal{N}(\mu, \sigma^2)$ i.i.d. (x_i is scalar-valued for simplicity)
- Then, $\boldsymbol{w} = \begin{pmatrix} \mu \\ \sigma \end{pmatrix}$:

$$p(x_i; \boldsymbol{w}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{1}{2} \frac{(x_i - \mu)^2}{\sigma^2} \Big)$$

- The mean μ and the variance σ^2 are the two parameters that need to be estimated.
- The likelihood function is:

$$\mathcal{L}(\mu, \sigma^2, x_1, \ldots, x_l) = \big(2\pi\sigma^2\big)^{-\frac{l}{2}} \exp\Big(-\frac{1}{2\sigma^2} \sum_{i=1}^{l} (x_i - \mu)^2 \Big)$$

MLE of normal distribution

- The negative log-likelihood function is:

$$L = -\ln(\mathcal{L})(\mu, \sigma^2, x_1, \ldots, x_l) = \frac{l}{2}\ln(2\pi) + \frac{l}{2}\ln(\sigma^2) + \frac{1}{2\sigma^2}\sum_{i=1}^{l}(x_i - \mu)^2$$

- Maximum likelihood estimate of μ:

$$\frac{\partial L}{\partial \mu} = \frac{\partial}{\partial \mu}\left(\frac{l}{2}\ln(2\pi) + \frac{l}{2}\ln(\sigma^2) + \frac{1}{2\sigma^2}\sum_{i=1}^{l}(x_i - \mu)^2\right)$$

$$= -\frac{1}{\sigma^2}\sum_{i=1}^{l}(x_i - \mu)$$

$$\Rightarrow \sum_{i=1}^{l} x_i - l\mu = 0 \Rightarrow \hat{\mu} = \frac{1}{l}\sum_{i=1}^{l} x_i$$

- Estimator $\hat{\mu} = \frac{1}{l}\sum_{i=1}^{l} x_i$ is **unbiased**

MLE of normal distribution

- The negative log-likelihood function is:

$$L = -\ln(\mathcal{L})(\mu, \sigma^2, x_1, \ldots, x_l) = \frac{l}{2}\ln(2\pi) + \frac{l}{2}\ln(\sigma^2) + \frac{1}{2\sigma^2}\sum_{i=1}^{l}(x_i - \mu)^2$$

- Maximum likelihood estimate of σ^2:

$$\frac{\partial L}{\partial \sigma^2} = \frac{\partial}{\partial \sigma^2}\left(\frac{l}{2}\ln(2\pi) + \frac{l}{2}\ln(\sigma^2) + \frac{1}{2\sigma^2}\sum_{i=1}^{l}(x_i - \mu)^2\right)$$

$$= \frac{l}{2\sigma^2} - \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{l}(x_i - \mu)^2 = \frac{1}{2\sigma^2}\left[l - \left(\frac{1}{\sigma^2}\sum_{i=1}^{l}(x_i - \mu)^2\right)\right]$$

$$\Rightarrow l - \left(\frac{1}{\sigma^2}\sum_{i=1}^{l}(x_i - \mu)^2\right) = 0 \Rightarrow \hat{\sigma^2} = \frac{1}{l}\sum_{i=1}^{l}(x_i - \mu)^2$$

- Estimator $\hat{\sigma^2} = \frac{1}{l}\sum_{i=1}^{l}(x_i - \mu)^2$ is **asymptotically unbiased**

Fisher information matrix

- The **Fisher information matrix** $\mathbf{I}_F(\boldsymbol{w})$ for a parametrized model:

$$\mathbf{I}_F(\boldsymbol{w}) : [\mathbf{I}_F(\boldsymbol{w})]_{ij} = \mathrm{E}_{p(\boldsymbol{x};\boldsymbol{w})}\left(\frac{\partial \ln p(\boldsymbol{x};\boldsymbol{w})}{\partial w_i}\frac{\partial \ln p(\boldsymbol{x};\boldsymbol{w})}{\partial w_j}\right) ,$$

  where $[\mathbf{A}]_{ij}$ selects the ijth component of a matrix $\mathbf{A}$.

- If the density function $p(\boldsymbol{x};\boldsymbol{w})$ satisfies:

$$\forall_{\boldsymbol{w}} : \mathrm{E}_{p(\boldsymbol{x};\boldsymbol{w})}\left(\underbrace{\frac{\partial \ln p(\boldsymbol{x};\boldsymbol{w})}{\partial \boldsymbol{w}}}_{\text{score}}\right) = \mathbf{0} ,$$

  then

$$\mathbf{I}_F(\boldsymbol{w}) = -\mathrm{E}_{p(\boldsymbol{x};\boldsymbol{w})}\left(\frac{\partial^2 \ln p(\boldsymbol{x};\boldsymbol{w})}{\partial \boldsymbol{w}\partial \boldsymbol{w}}\right) .$$

Fisher information matrix

$$
\begin{aligned}
\mathbf{0} &= \mathrm{E}_{p(\boldsymbol{x};\boldsymbol{w})}\left(\frac{\partial \ln p(\boldsymbol{x};\boldsymbol{w})}{\partial \boldsymbol{w}}\right) \\
&= \int_{\mathbf{X}} \frac{\partial \ln p(\boldsymbol{x};\boldsymbol{w})}{\partial \boldsymbol{w}} p(\boldsymbol{x};\boldsymbol{w}) d\boldsymbol{x} \quad \text{assuming } \mathbf{X} \text{ is distributed as } p(\boldsymbol{x};\boldsymbol{w}) \\
&\Rightarrow \frac{\partial}{\partial \boldsymbol{w}} \int_{\mathbf{X}} \frac{\partial \ln p(\boldsymbol{x};\boldsymbol{w})}{\partial \boldsymbol{w}} p(\boldsymbol{x};\boldsymbol{w}) d\boldsymbol{x} = \mathbf{0} \\
&\Rightarrow \int_{\mathbf{X}} \left(\frac{\partial^2 \ln p(\boldsymbol{x};\boldsymbol{w})}{\partial \boldsymbol{w} \partial \boldsymbol{w}} p(\boldsymbol{x};\boldsymbol{w}) + \frac{\partial \ln p(\boldsymbol{x};\boldsymbol{w})}{\partial \boldsymbol{w}} \frac{\partial p(\boldsymbol{x};\boldsymbol{w})}{\partial \boldsymbol{w}} \right) d\boldsymbol{x} \\
&\Rightarrow \int_{\mathbf{X}} \left(\frac{\partial^2 \ln p(\boldsymbol{x};\boldsymbol{w})}{\partial \boldsymbol{w} \partial \boldsymbol{w}} p(\boldsymbol{x};\boldsymbol{w}) + \frac{\partial \ln p(\boldsymbol{x};\boldsymbol{w})}{\partial \boldsymbol{w}} \frac{\partial \ln p(\boldsymbol{x};\boldsymbol{w})}{\partial \boldsymbol{w}} p(\boldsymbol{x};\boldsymbol{w}) \right) d\boldsymbol{x} \\
&\Rightarrow -\mathrm{E}_{p(\boldsymbol{x};\boldsymbol{w})}\left(\frac{\partial^2 \ln p(\boldsymbol{x};\boldsymbol{w})}{\partial \boldsymbol{w} \partial \boldsymbol{w}} \right) = \mathrm{E}_{p(\boldsymbol{x};\boldsymbol{w})}\left(\frac{\partial \ln p(\boldsymbol{x};\boldsymbol{w})}{\partial \boldsymbol{w}} \frac{\partial \ln p(\boldsymbol{x};\boldsymbol{w})}{\partial \boldsymbol{w}} \right)
\end{aligned}
$$

Fisher information matrix

- Fisher information is the average curvature of the negative log-likelihood
- Main rationale: the larger this quantity, the smaller the variance of $p(\boldsymbol{x}, \boldsymbol{w})$
- The Fisher information gives the amount of information that an observable random variable $\boldsymbol{x}$ carries about an unobservable parameter $\boldsymbol{w}$ upon which the parameterized density function $p(\boldsymbol{x}; \boldsymbol{w})$ of $\boldsymbol{x}$ depends

Cramér-Rao lower bound (CRLB)

- Let $\hat{\boldsymbol{w}}$ be an unbiased estimator of $\boldsymbol{w}$
- Let $Cov(\hat{\boldsymbol{w}})$ be the covariance matrix and
  $[\mathbf{I}_F(\boldsymbol{w})]_{ij} = -\mathrm{E}_{p(\boldsymbol{x};\boldsymbol{w})}\left(\frac{\partial^2 \ln p(\boldsymbol{x};\boldsymbol{w})}{\partial \boldsymbol{w}_i \partial \boldsymbol{w}_j}\right)$
- **CRLB**:

$$Cov(\hat{\boldsymbol{w}}) \geq \frac{1}{\mathbf{I}_F(\boldsymbol{w})}$$

- An estimator is said to be **efficient** if it reaches the CRLB. It is efficient in that it efficiently makes use of the data and extracts information to estimate the parameter
- A MVU estimator may or may not be efficient. This means it could have minimum variance but without reaching the CRLB

CRLB of normal distribution

- First we calculate $\text{var}(\hat{\mu})$ and $\text{var}(\hat{\sigma^2})$:

$$
\begin{aligned}
\text{var}(\hat{\mu}) &= \mathrm{E}_{\mathbf{x}}\left((\hat{\mu} - \mathrm{E}_{\mathbf{x}}(\hat{\mu}))^T(\hat{\mu} - \mathrm{E}_{\mathbf{x}}(\hat{\mu}))\right) \\
&= \mathrm{E}_{\mathbf{x}}\left(\left(\frac{1}{l}\sum_{i=1}^{l} x_i - \frac{1}{l}\sum_{i=1}^{l}\mathrm{E}_{\mathbf{x}}(x_i)\right)^T\left(\frac{1}{l}\sum_{i=1}^{l} x_i - \frac{1}{l}\sum_{i=1}^{l}\mathrm{E}_{\mathbf{x}}(x_i)\right)\right) \\
&= \mathrm{E}_{\mathbf{x}}\left(\frac{1}{l^2}\left(\sum_{i=1}^{l} x_i - \sum_{i=1}^{l}\mathrm{E}_{\mathbf{x}}(x_i)\right)^T\left(\sum_{i=1}^{l} x_i - \sum_{i=1}^{l}\mathrm{E}_{\mathbf{x}}(x_i)\right)\right) \\
&= \frac{1}{l^2}\cdot l\sigma^2 = \frac{\sigma^2}{l} \\
\text{var}(\hat{\sigma^2}) &= \frac{2\sigma^4}{l-1}
\end{aligned}
$$

- Derivation of $\text{var}(\hat{\sigma^2})$ via $\dfrac{(l-1)\hat{\sigma^2}}{\sigma^2} \sim \chi^2_{l-1}$

CRLB of normal distribution

- The negative log-likelihood function is:

$$L = -\ln(\mathcal{L})(\mu, \sigma^2, x_1, \ldots, x_l) = \frac{l}{2}\ln(2\pi) + \frac{l}{2}\ln(\sigma^2) + \frac{1}{2\sigma^2}\sum_{i=1}^{l}(x_i - \mu)^2$$

- $[\mathbf{I}_F(\boldsymbol{w})]_{ij} = -\mathrm{E}_{p(\boldsymbol{x};\boldsymbol{w})}\left(\frac{\partial^2 \ln p(\boldsymbol{x};\boldsymbol{w})}{\partial \boldsymbol{w}_i \partial \boldsymbol{w}_j}\right)$:

$$\Rightarrow \nabla^2_{\mu,\sigma^2} L = \begin{pmatrix} \frac{\partial^2}{\partial \mu^2} & \frac{\partial^2}{\partial \mu \partial \sigma^2} \\ \frac{\partial^2}{\partial \sigma^2 \partial \mu} & \frac{\partial^2}{\partial (\sigma^2)^2} \end{pmatrix} = \begin{pmatrix} \frac{l}{\sigma^2} & \frac{1}{\sigma^4}\sum_{i=1}^{l}(x_i - \mu) \\ \frac{1}{\sigma^4}\sum_{i=1}^{l}(x_i - \mu) & \frac{1}{(\sigma^2)^3}\sum_{i=1}^{l}(x_i - \mu)^2 - \frac{l}{2(\sigma^2)^2} \end{pmatrix}$$

$$\Rightarrow -\mathrm{E}_{p(\boldsymbol{x};\boldsymbol{w})}\left(\frac{\partial^2 \ln p(\boldsymbol{x};\boldsymbol{w})}{\partial \boldsymbol{w}_i \partial \boldsymbol{w}_j}\right) = \mathrm{E}\left(\nabla^2_{\mu,\sigma^2} L\right) = \begin{pmatrix} \frac{l}{\sigma^2} & 0 \\ 0 & \frac{l}{2(\sigma^2)^2} \end{pmatrix}$$

- **CRLB**:

$$Cov(\hat{\boldsymbol{w}}) \geq \frac{1}{\mathbf{I}_F(\boldsymbol{w})}$$

$$\begin{pmatrix} \frac{\sigma^2}{l} & 0 \\ 0 & \frac{2\sigma^4}{l-1} \end{pmatrix} \geq \begin{pmatrix} \frac{l}{\sigma^2} & 0 \\ 0 & \frac{l}{2(\sigma^2)^2} \end{pmatrix}^{-1}$$

$$\geq \begin{pmatrix} \frac{\sigma^2}{l} & 0 \\ 0 & \frac{2(\sigma^2)^2}{l} \end{pmatrix}$$

- MLE estimate $\hat{\mu} = \frac{1}{l}\sum_{i=1}^{l} x_i$ is **efficient**
- MLE estimate $\hat{\sigma^2} = \frac{1}{l}\sum_{i=1}^{l}(x_i - \mu)^2$ is **asymptotically efficient**

Properties of MLE

- Summary: MLE is a popular estimator on which almost all practical estimation tasks are based.
- The popularity stems from the fact that it can be applied to a broad range of problems and it approximates the MVU estimator for large data sets.
- MLE is asymptotically efficient and unbiased.

Properties of MLE

- **Invariance under parameter change**: Let g be a function changing the parameter $\boldsymbol{w}$ into parameter $\boldsymbol{u}$: $\boldsymbol{u} = g(\boldsymbol{w})$, then

$$\hat{\boldsymbol{u}} = g(\hat{\boldsymbol{w}}) \ ,$$

  where $\boldsymbol{u}$ and $\boldsymbol{w}$ are ML estimators. If g changes $\boldsymbol{w}$ into different $\boldsymbol{w}$ then $\hat{\boldsymbol{u}} = g(\hat{\boldsymbol{w}})$ maximizes the likelihood function: $\max\limits_{\boldsymbol{w}:\boldsymbol{u}=g(\boldsymbol{w})} p(\{\boldsymbol{x}\};\boldsymbol{w})$

- **Asymptotically unbiased**:

$$\mathrm{E}_{p(\boldsymbol{x};\boldsymbol{w})}(\hat{\boldsymbol{w}}) \overset{l\to\infty}{\to} \boldsymbol{w}$$

- **Asymptotically efficienct**:

$$Cov(\hat{\boldsymbol{w}}) \overset{l\to\infty}{\to} \text{CRLB}$$