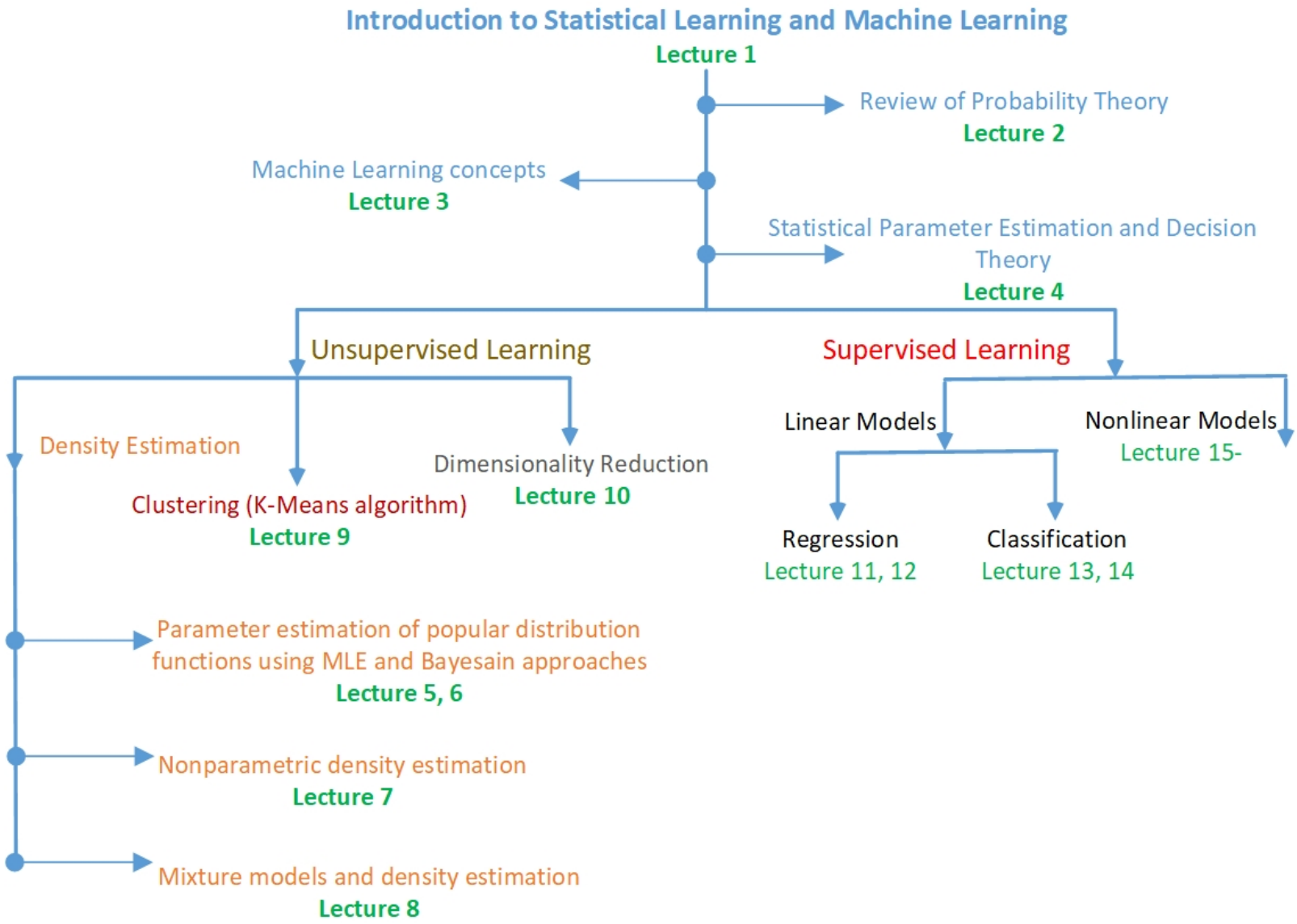
Statistical Learning and Machine Learning

Lecture 4 - Decision Theory and Review of Statistical Estimation Theory

August, 2025

Course overview and where do we stand



Objectives of the lecture

- Introduction to the key ideas of decision theory that will lay foundations for the classification problems studied in the course
- Introducing basic concepts of statistical parameter estimation with emphasis on:
  1. Maximum likelihood estimation
  2. Bayesian estimation
- Example of parameter estimation for Gaussian distribution

Decision Theory

Probability theory: models uncertainty in data
Decision theory: helps make optimal decisions amid uncertainties

Consider that we have training data consisting of multiple inputs $\{x_i\}_{i=1}^N$ with corresponding target vector $\{t\}_{i=1}^N$; *our goal is to predict* $\tilde{t}$ *given a new value of* $\tilde{x}$.

- $p(x, t)$ provides the complete information of the uncertainty associated with these two variables
- estimating $p(x, t)$ given a set of training data $\{x_n, t_n\}$, $n = 1, \ldots, N$ is called **inference** – a task in probability/estimation theory
- In practical applications, we must often make predictions of t, given x, or take a specific action/decision based on the predicted value of t – a goal of decision theory.

Example: medical diagnosis of cancer based on X-ray image of a patient.

Decision Theory

Suppose that in a given classification problem, the target t correspond to K number of classes/labels $\{\mathcal{C}_k\}_{k=1}^{K}$:

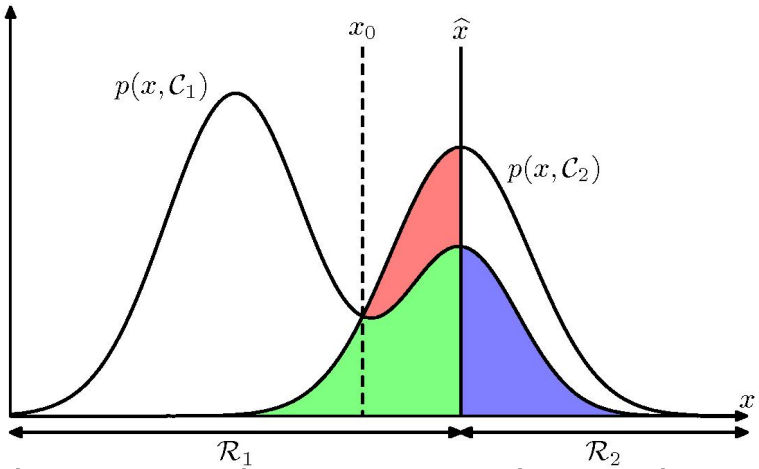
- We can define a classification rule using the Bayes' theorem:

$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x})}$$

- All quantities in the Bayes' theorem can be obtained from $p(\mathbf{x}, \mathcal{C}_k)$
- posterior $\propto$ likelihood $\times$ class prior
- Finally, we can choose the class $\mathcal{C}_k$ with the highest posterior probability. *This would minimize the chance of assigning x to a wrong class.*

Minimizing misclassification error

- Consider a classification problem where input x is real-valued and target t corresponds to two classes, $\mathcal{C}_1$ and $\mathcal{C}_2$.
- Goal: Finding a rule/function that assigns x to a particular class based on an informed decision
- Divide input space into decision regions $\mathcal{R}_1$ and $\mathcal{R}_2$, such that $\mathcal{R}_k$ is assigned to $\mathcal{C}_k$. How to create such decision boundary?



$$
\begin{aligned}
p(\text{mistake}) &= p(\mathbf{x} \in \mathcal{R}_1, \mathcal{C}_2) + p(\mathbf{x} \in \mathcal{R}_2, \mathcal{C}_1) \\
&= \int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2) d\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1) d\mathbf{x}
\end{aligned}
$$

Maximizing the probability of correct classifications

For $K > 2$ classes it is easier to maximize the probability of correct classification:

$$\begin{aligned} p(\text{correct}) &= \sum_{k=1}^{K} p(\mathbf{x} \in \mathcal{R}_k, \mathcal{C}_k) \\ &= \sum_{k=1}^{K} \int_{\mathcal{R}_k} p(\mathbf{x}, \mathcal{C}_k) d\mathbf{x} \end{aligned}$$

Using the product rule:

$$p(\mathbf{x}, \mathcal{C}_k) = p(\mathcal{C}_k|\mathbf{x}) p(\mathbf{x})$$

Since $p(\mathbf{x})$ is common for all values of k, each point $\mathbf{x}$ should be assigned to the class having the largest posterior probability $p(\mathcal{C}_k|\mathbf{x})$.

Minimizing the expected loss

- In many applications, our objective is more complex than just minimizing the number of misclassifications e.g. cancer diagnosis.
- Defining 'loss function' and minimizing average loss:
  - When a data point belongs to $\mathcal{C}_k$ and we classify it to $\mathcal{C}_j$ we incur some loss denoted by L_{kj}. The set of loss values forms the *loss matrix* $\mathbf{L}$. For example:

$$\mathbf{L} = \begin{bmatrix} 0 & 1000 \\ 1 & 0 \end{bmatrix}$$

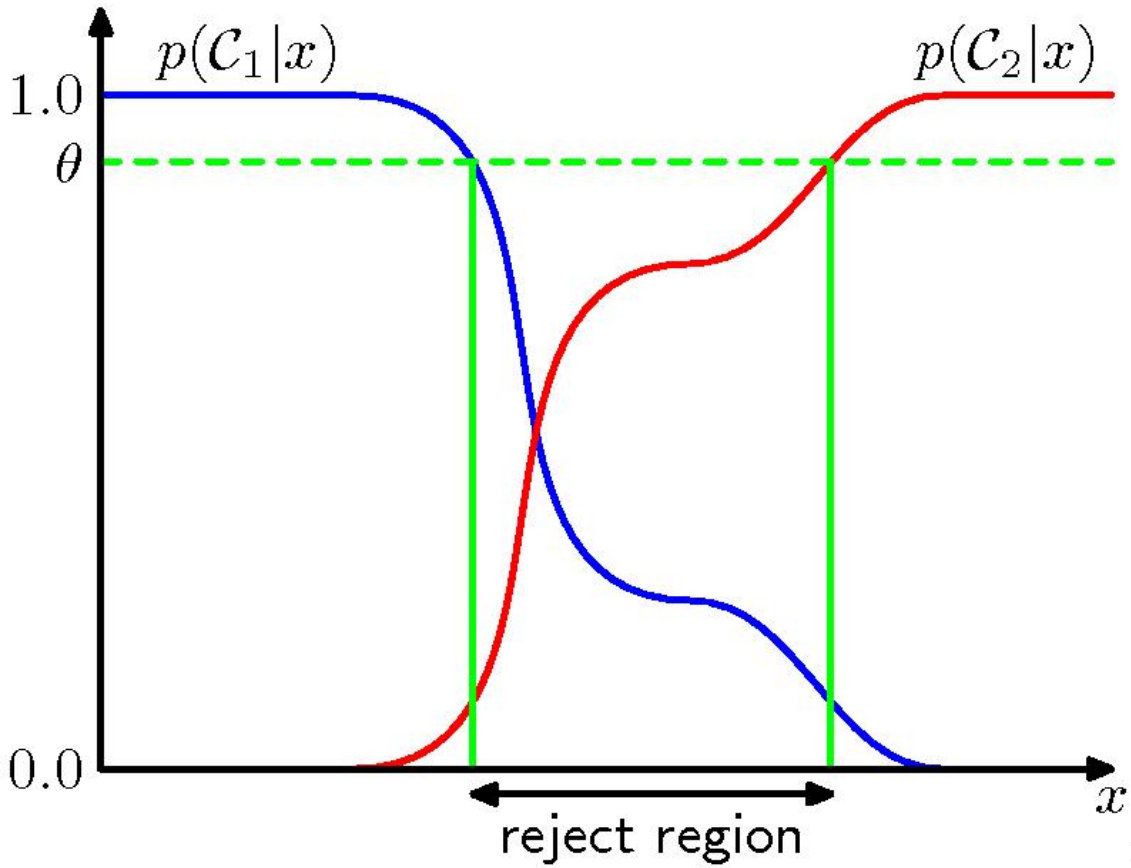
  - We seek to minimize the average loss:

$$\mathbb{E}[L] = \int_{\mathcal{R}_1} L_{21} p(\mathbf{x}, \mathcal{C}_2) d\mathbf{x} + \int_{\mathcal{R}_2} L_{12} p(\mathbf{x}, \mathcal{C}_1) d\mathbf{x}$$

  - The minimization of the expected loss is the one that assigns a data points $\mathbf{x}$ to the class j for which the quantity: $\sum_k L_{kj} p(\mathcal{C}_k|\mathbf{x})$ is minimized.

Reject region

- Classification errors arise from the regions of input space where joint distributions $p(\boldsymbol{x}, \mathcal{C}_k)$ have comparable values. These are the regions where we are uncertain about the class membership.
- In some cases, it would be appropriate not to make decisions on those difficult cases.



Inference and decision I

1. Inference followed by detection: First, solve the inference problem of determining the class-conditional densities $p(\mathbf{x}|\mathcal{C}_k)$ or joint distributions $p(\boldsymbol{x}, \mathcal{C}_k)$. Then obtain posterior densities:

$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x})} = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{\sum_j p(\mathbf{x}|\mathcal{C}_j)p(\mathcal{C}_j)}$$

Finally, decision theory is used to determine class membership for new $\boldsymbol{x}$.

Pros: Outlier detection is possible since we know $p(x)$.

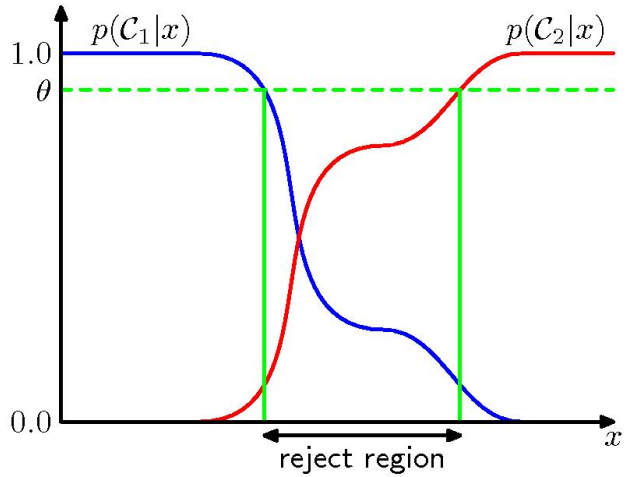
Cons:

- Computationally demanding especially for $\boldsymbol{x}$ with large dimensions. Lots of training data may be needed to obtain the class-conditional densities $p(\mathbf{x}|\mathcal{C}_k)$ with reasonable accuracy.
- Inefficient if classification decision is the main priority.

Inference and decision II

2. Obtaining posterior class probabilities followed by discrimination:
   Solve the inference problem of determining the posterior probabilities $p(\mathcal{C}_k|\mathbf{x})$, and use decision theory to assign a new $\mathbf{x}_*$ to one of the classes. This corresponds to a *discriminative model*.
   Pros:
   - Posterior probabilities are needed in case loss function is involved in minimizing risk
   - Posterior probabilities are needed to implement 'reject option'.

   Cons: No outlier detection possible



Inference and decision III

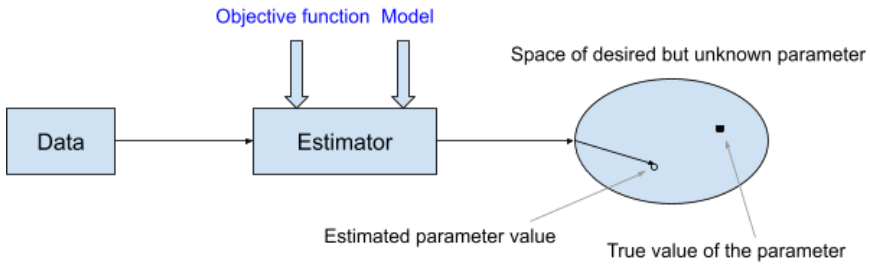
3. Find a function $f(\mathbf{x})$ called discriminant function which maps a new $\mathbf{x}_*$ onto a class label.
   Pros: Most efficient in the where case decision-making is the main priority
   Cons:
   - No outlier detection possible
   - Lack of posterior class probabilities limits the use of reject option and using flexible loss functions.

Statistical Parameter Estimation I

- Systematically inferring unobserved hidden variables from observed data by taking into account process uncertainties.
- Specifically, consider discrete data samples $x[0], x[1], \ldots, x[N-1]$ which depend on unknown parameter θ. **An estimator is a rule/function that estimates θ from a given realization of data.**

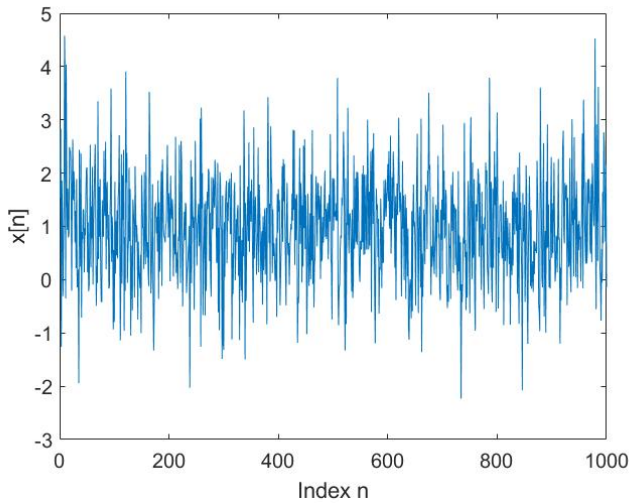
$$\hat{\theta} = g(x[0], \ldots, x[N-1]).$$

where g is some function. This is *parameter estimation*.



Statistical Parameter Estimation II

- Example (DC level in noise): Considering the data shown below, one can expect that $x[n]$ consists of DC level in noise.



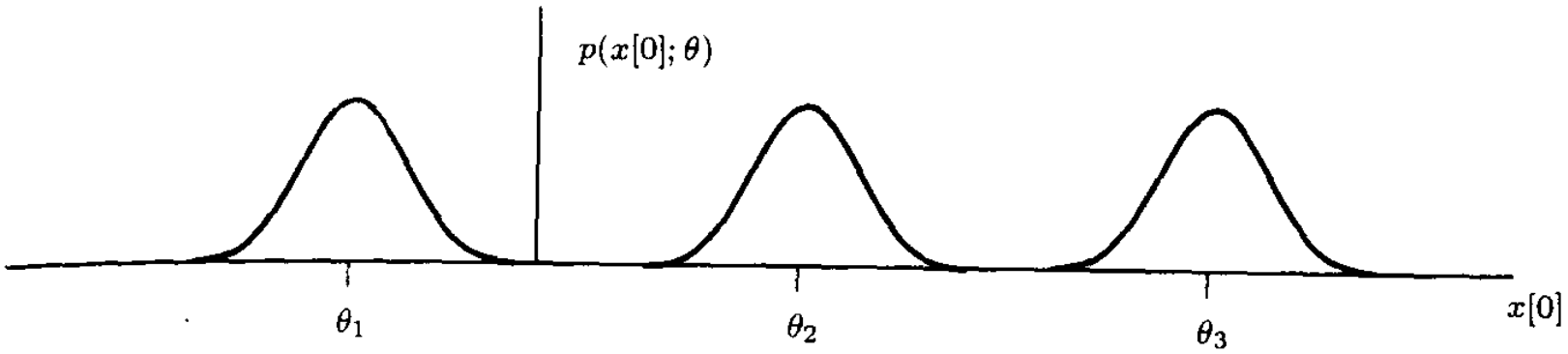
1. The model can be defined as

$$x[n] = A + w[n], \qquad \textbf{MODEL}$$

where $w[n]$ denotes zero-mean white Gaussian noise process.

Statistical Parameter Estimation III

2. **Specifying PDF of data**: This is a key step in estimation problem as it models randomness in data. The joint PDF of observations $\boldsymbol{x} = \{x[0], x[1], \ldots, x[N-1]\}$ is modelled as $p(x[0], \ldots, x[N-1]; \theta)$ or $p(\boldsymbol{x}; \theta)$.



In the previous example, $\theta = A$ and if we consider $N = 1$, the PDF is:

$$p[x[0]; A] = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x[0]-A)^2}{2\sigma^2}}.$$

The key point: Since the probability of $x[0]$ depends on θ, we can *infer* the value of θ from the observed value of $x[0]$.

Statistical Parameter Estimation IV

3. **Objective Function**: must be defined as a measure to 'drive' our estimate $\hat{\theta}$ close to the actual value θ_0; *that leads to ill-posed problem in most cases.* Why?

   Estimator: $\hat{\theta} = \frac{1}{N} \sum_{n=1}^{N} x[n]$.

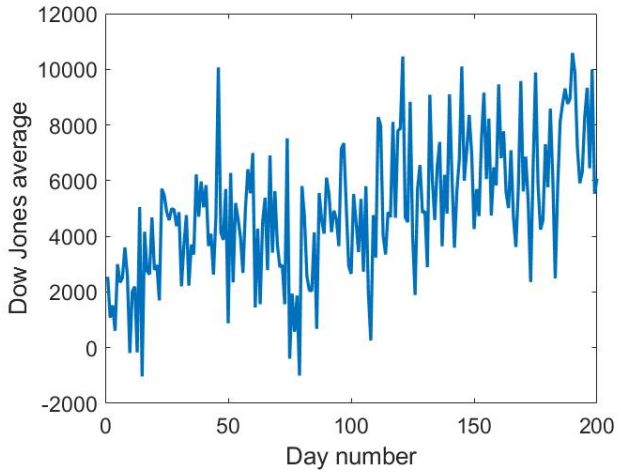
   - What happens when we change the objective function?
   - What happens when we change model?

Statistical Parameter Estimation V

4. Another example (Hypothetical Dow-Jones industrial average): The model in this case could be:

$$x[n] = A + Bn + w[n],$$

where $w[n]$ is zero-mean white Gaussian noise process. What will be the PDF of observations in this case?

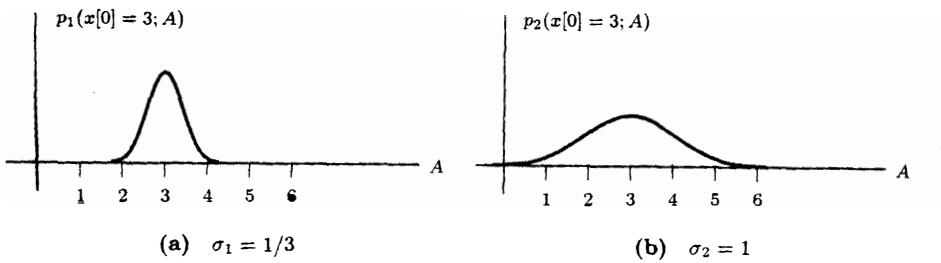


Likelihood Function I

Likelihood function: The PDF of data when viewed as a function of unknown parameter θ (with $\boldsymbol{x}$ fixed) is called the likelihood function.

Example (DC level within Noise): Let $x[0] = A + w[0]$ where $w[0] \sim \mathcal{N}(0, \sigma^2)$ and it is desired to estimate A. The liklihood function can be written as

$$p(x[0]; A) = \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{\frac{-(x[0]-A)^2}{2\sigma_i^2}}$$



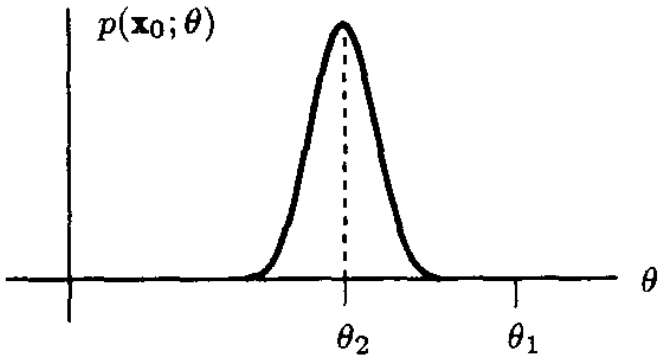
(a) $\sigma_1 = 1/3$ (b) $\sigma_2 = 1$

Rationale for MLE

- The MLE (for scalar parameter θ) gives the value of θ that maximizes the likelihood function. Assuming a differentiable likelihood function, the MLE is found from

$$\frac{\partial \ln p(\boldsymbol{x};\theta)}{\partial \theta} = 0$$

- What is the physical intuition behind this?



Example: Estimating mean for DC in WGN process

Consider the following model

$$x[n] = A + w[n] \quad n = 0, \ldots, N-1$$

where A is the parameter to be estimated via MLE and $w[n]$ is the WGN process with known variance σ^2.

The MLE for A in this case is:

$$\hat{A} = \frac{1}{N} \sum_{n=0}^{N-1} x[n].$$

Bayesian Approach to Parameter Estimation I

- **Main assumption**: The parameter of interest θ is considered a random variable whose particular realization we must estimate.
- **Advantages**:
  1. Prior information about parameter of interest can be incorporated in the estimator via its prior PDF; generally, prior uncertainty (of θ) will be high and it will be reduced after observing data.
  2. If optimal estimator does not exist or cannot be found, then by assigning a PDF to θ, we can devise strategies to find such approximately optimal estimator.

Bayesian Approach to Parameter Estimation II

- Specifying joint PDF of parameter and observed data: Using the conditional PDF relations, we get

$$\begin{aligned} p(\boldsymbol{x};\theta) &= p(\boldsymbol{x}|\theta)p(\theta) \\ &= p(\theta|\boldsymbol{x})p(\boldsymbol{x}). \end{aligned}$$

where $p(\boldsymbol{x}|\theta)$ is the **likelihood function** and $p(\theta)$ is termed the **prior PDF** of parameter of interest.
Using the above relations, we get

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

where $p(\boldsymbol{x})$ is independent of θ and thus can be considered a constant, leading to

$$p(\theta|\boldsymbol{x}) = c \times p(\boldsymbol{x}|\theta)p(\theta)$$

The $p(\theta|\boldsymbol{x})$ is called the **posterior PDF**.

Reading Assignment

- We have already covered a significant part of **section 1.2.4** of the text book; please read that section.
- Go through the **section 1.2.5** of the textbook.