Decision and Estimation in Data Processing

Chapter III. Elements of Estimation Theory

II.1 Introduction

What means "Estimation"?

- A sender transmits a signal $s_\Theta(t)$ which depends on an **unknown** parameter Θ
- The signal is affected by noise, we receive $r(t) = s_\Theta(t) + noise$
- We want to **find out** the correct value of the parameter
  - based on samples from the received signal, or the full continuous signal
  - available data is noisy => we "estimate" the parameter
- The found value is $\hat{\Theta}$, **the estimate** of Θ ("estimatul", rom)
  - there will always be some estimation error $\epsilon = \hat{\Theta} - \Theta$
- Examples:
  - Unknown amplitude of constant signal: $r(t) = A + noise$, estimate A
  - Unknown phase of sine signal: $r(t) = \cos(2\pi f t + \phi)$, estimate ϕ
  - Record speech signal, estimate/decide what word is pronounced

Estimation vs Decision

- Consider the following estimation: $r(t) = A + \mathit{noise}$, estimate A
- For detection, we have to choose between **two known values** of A:
  - i.e. A can be 0 or 5 (hypotheses H_0 and H_1)
- For estimation, A can be anything $=>$ we choose between **infinite number of options** for A:
  - A might be any value in $\mathbb{R}$, in general

Estimation vs Decision

- Detection = Estimation constrained to only a few discrete options
- Estimation = Detection with an infinite number of options available
- The statistical methods used are quite similar
  - In practice, distinction between Estimation and Detections is somewhat blurred
  - (e.g. when choosing between 1000 hypotheses, do we call it "Detection" or "Estimation"?)

Available data

- The available data is the received signal $r(t)$
  - affected by noise, and depending on the unknown Θ
- We consider **N samples** from $r(t)$, taken at some sample times t_i

$$\mathbf{r} = [r_1, r_2, ... r_N]$$

- Each sample r_i is a random variable that depends on Θ (and the noise)
  - Each sample has a distribution that depends on Θ

$$w_i(r_i; \Theta)$$

- The whole sample vector $\mathbf{r}$ is a N-dimensional random variable that depends on Θ (and the noise)
  - It has a N-dimensional distribution that depends on Θ

$$w(\mathbf{r}; \Theta)$$

Types of estimation

- We consider estimating a parameter Θ under two circumstances:

1. No distribution is known about the parameter, except maybe some allowed range (e.g. $\Theta > 0$)
    - The parameter can be any value in the allowed range, equally likely
2. We know a distribution $p(\Theta)$ for Θ, which tells us the values of Θ that are more likely than others
    - this is known as *a priori* distribution (i.e. "known beforehand")

II.2 Maximum Likelihood estimation

Maximum Likelihood definition

- When no distribution is known about the parameter, we use a method known as Maximum Likelihood Estimation (MLE)
- The distribution of the received data, $w(\mathbf{r}; \Theta)$, is known as the **likelihood function**
  - we know the vector $\mathbf{r}$ we received, so this is a constant
  - the unknown variable in this function is Θ

$$L(\Theta) = w(\mathbf{r}; \Theta)$$

- Maximum Likelihood Estimation: The estimate $\hat{Theta}$ is **the value that maximizes the likelihood of the observed data**
  - i.e. the value Θ that maximizes $w(r; \Theta)$

$$\hat{\Theta} = \arg\max_{\Theta} L(\Theta) = \arg\max_{\Theta} w(r; \Theta)$$

- If Θ is allowed to live only in a certain range, restrict the maximization only to that range.

Computations

- Find maximum by setting derivative to 0

$$\frac{dL(\Theta)}{d\Theta} = 0$$

- We can also maximize **natural logarithm** of the likelihood function ("log-likelihood function")

$$\frac{d \ln (L(\Theta))}{d\Theta} = 0$$

Computations

Method:

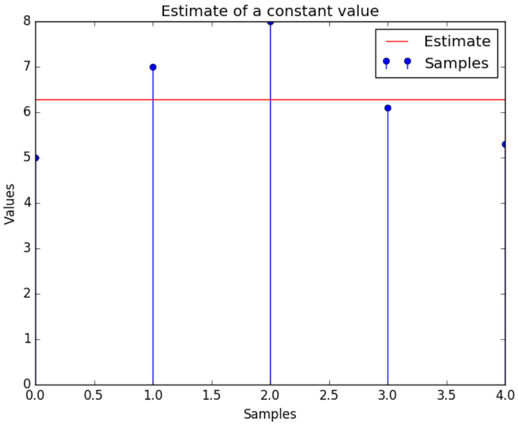
1. Find the function
$$L(\Theta) = w(\mathbf{r}; \Theta)$$
2. Set the condition that derivative of $L(\Theta)$ or $\ln((L(\Theta))$ is 0
$$\frac{dL(\Theta)}{d\Theta} = 0, \text{ or } \frac{d\ln\left(L(\Theta)\right)}{d\Theta} = 0$$
3. Solve and find the value $\hat{\Theta}$
4. Check that second derivative at point $\hat{\Theta}$ is negative, to check that point is a maximum
   - because derivative $= 0$ for both maximum and minimum points

Examples:

Estimating a constant signal in gaussian noise:

- Find the Maximum Likelihood estimate of a constant value A from 5 noisy measurements $r_i = A + noise$ with values $[5, 7, 8, 6.1, 5.3]$. The noise is AWGN $\mathcal{N}(\mu = 0, \sigma^2)$.
- Solution: at whiteboard.
- The estimate $\hat{A}$ is the average value of the samples (not surprisingly)

Numerical simulation



General signal in AWGN

- Consider that the true underlying signal is $s_\Theta(t)$
- Consider AWGN noise $\mathcal{N}(\mu = 0, \sigma^2)$.
- The samples r_i are taken at sample moments t_i
- The samples r_i have normal distribution with average $s_\Theta(t_i)$ and variance σ^2
- Overall likelihood function = product of likelihoods for each sample r_i

$$L(\Theta) = \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(r_i - s_\Theta(t_i))^2}{2\sigma^2}}$$

$$= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{\sum (r_i - s_\Theta(t_i))^2}{2\sigma^2}}$$

- The log-likelihood is

$$\ln\left(L(\Theta)\right) = \underbrace{\ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right)}_{constant} - \frac{\sum (r_i - s_\Theta(t_i))^2}{2\sigma^2}$$

General signal in AWGN

$$\frac{d\ln\left(L(\Theta)\right)}{d\Theta} = 0$$

means

$$\sum (r_i - s_\Theta(t_i))^2 \frac{ds_\Theta(t_i)}{d\Theta} = 0$$

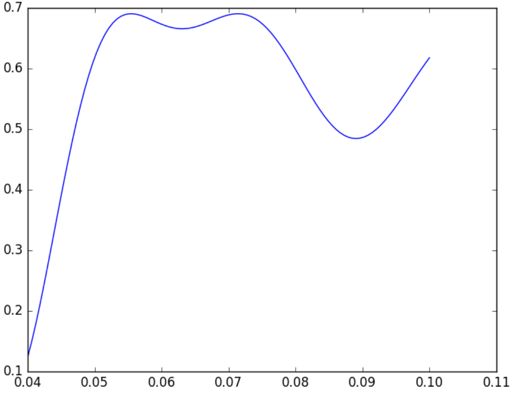
Numerical simulation

Estimating the frequency f of a cosine signal

- Find the Maximum Likelihood estimate of the frequency f of a cosine signal, from 10 noisy measurements $r_i = cos(2\pi f t_i) + noise$ with values [...]. The noise is AWGN $\mathcal{N}(\mu = 0, \sigma^2)$. The sample times $t_i = [0, 1, 2, 3, 4, 5, 6, 7, 8, 9]$
- Solution: at whiteboard.

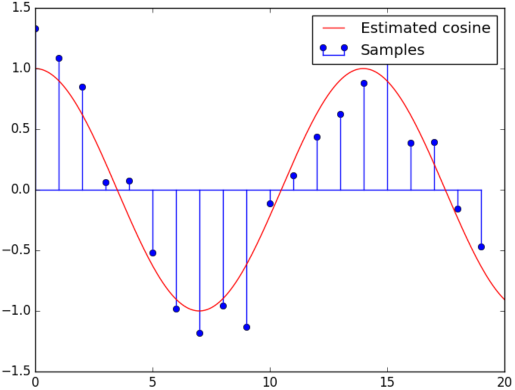
Numerical simulation

The likelihood function is:



Numerical simulation

```
True frequency = 0.070000, Estimate = 0.071515
```



ML Estimation and ML Detection

- In ML Estimation, the estimate $\hat{\Theta}$ is the value that maximizes the likelihood function
- In ML Detection, the decision criterion $\frac{w(r|H_1)}{w(r|H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} 1$ means "choose the hypothesis that maximizes the likelihood function".
- Therefore it is the same principle, merely in a different context:
  - in Detection we are restricted to a few predefined options
  - in Estimation we are unrestricted $=>$ choose the maximizing value