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: $r(t) = \cos(2\pi ft + \phi)$, estimate ϕ
  - Record speech signal, estimate/decide what word is pronounced

Estimation vs Decision

- Consider the following estimation: $r(t) = A + noise$, 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

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

Likelihood function

- In an estimation problem:
  - $\mathbf{r}$ is known
  - Θ is unknown
- We want to estimate Θ based on $\mathbf{r}$, so we are interested in the following function:

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

- This is the likelihood (distribution / probability) of Θ, for a given known $\mathbf{r}$

Bayes rule

- In general, we can use the Bayes rule

$$L(\Theta) = w(\Theta|\mathbf{r}) = \frac{w(\mathbf{r}|\Theta) \cdot w(\Theta)}{w(\mathbf{r})}$$

- Explanation of the terms:
  - Θ is the unknown parameter
  - $\mathbf{r}$ are the observations that we have
  - $L(\Theta) = w(\Theta|\mathbf{r})$ is the likelihood of Θ, given our current observations;
  - $w(\mathbf{r}|\Theta)$ is the "normal" probability of $\mathbf{r}$ for a given Θ, given by the noise distribution
  - $w(\Theta)$ is the **prior** distribution of Θ, i.e. what we know about Θ even in the absence of evidence
  - $w(\mathbf{r})$ is the prior distribution of $\mathbf{r}$, it is assumed constant

Bayes rule

- The previous relation is rather complex
- It shows that our estimation of Θ depends on two things:
    1. The observations that we have, via the term $w(\mathbf{r}|\Theta)$
    2. The prior knowledge (or prior belief) about Θ, via the term $w(\Theta)$

(the third term $w(\mathbf{r})$ is considered a constant, and plays no significant role)

Two 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
   - We treat $w(\Theta)$ as a constant
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* (or *prior*) 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)**
- We treat $w(\Theta$ as a constant, so that the likelihood function becomes:

$$L(\Theta) = w(\mathbf{r}|\Theta) \cdot constant$$

Maximum Likelihood definition

Maximum Likelihood (ML) Estimation:

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

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

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

How to solve

- How to solve the maximization problem?
  - i.e. how to find the estimate Θ which maximizes $L(\Theta)$
- 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")

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

Solving procedure

Solving procedure:

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

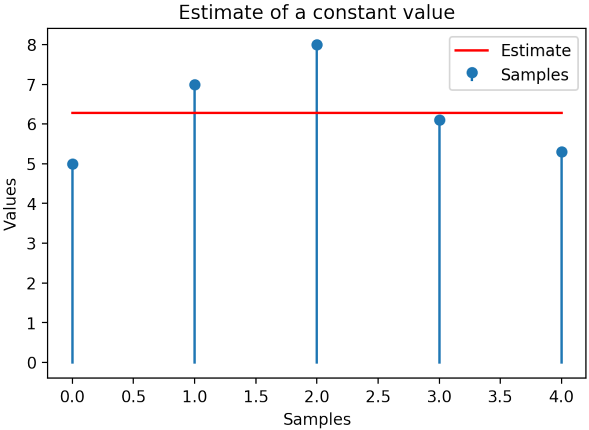
Examples:

Estimating a constant signal in gaussian noise:

- Find the ML estimate of a constant value A from 5 noisy measurements $r_i = A + \textit{noise}$ 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

```
/home/ncleju/.local/bin/pweave:5: UserWarning: In Matplotlib
individual lines on a stem plot will be added as a LineCollec
instead of individual lines. This significantly improves the
performance of a stem plot. To remove this warning and switch
new behaviour, set the "use_line_collection" keyword argument
```



Curve fitting

- Estimation = curve fitting
- From the previous graphical example:
  - we have some data **r**
  - we know the shape of the signal = a line (constant A)
  - we're fitting the best line through the data

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

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

General signal in AWGN

- 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

- The maximum of the function = the minimum of the exponent

$$\hat{\Theta} = \arg\max_{\Theta} w(r; \Theta) = \arg\min \sum (r_i - s_\Theta(t_i))^2$$

- The term $\sum (r_i - s_\Theta(t_i))^2$ is the **squared distance** $d(\mathbf{r}, s_\Theta)$

$$d(\mathbf{r}, s_\Theta) = \sqrt{\sum (r_i - s_\Theta(t_i))^2}$$

$$(d(\mathbf{r}, s_\Theta))^2 = \sum (r_i - s_\Theta(t_i))^2$$

General signal in AWGN

- ML estimation can be rewritten as:

$$\hat{\Theta} = \arg\max_{\Theta} w(r;\Theta) = \arg\min d(\mathbf{r}, \mathbf{s}_\Theta)^2$$

- ML estimate $\hat{\Theta}$ = the value that makes $s_\Theta(t_i)$ **closest to the received values r**
  - closer = more likely
  - closest = most likely = maximum likelihood
- ML estimation = minimization of distance
- True for all kinds of vector spaces
  - vectors with N elements, continous signals, etc
  - just change the definition of the distance function

General signal in AWGN

- Find maximum by setting derivative to 0

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

means

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

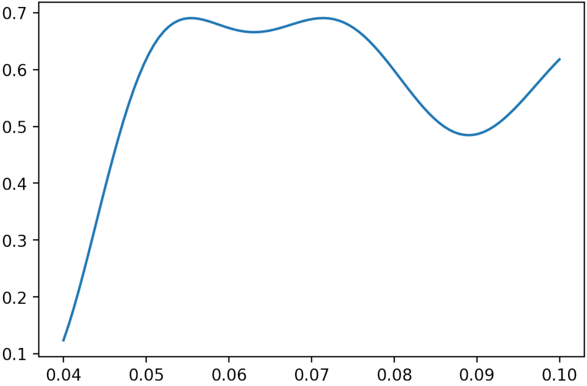
Numerical simulation

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.

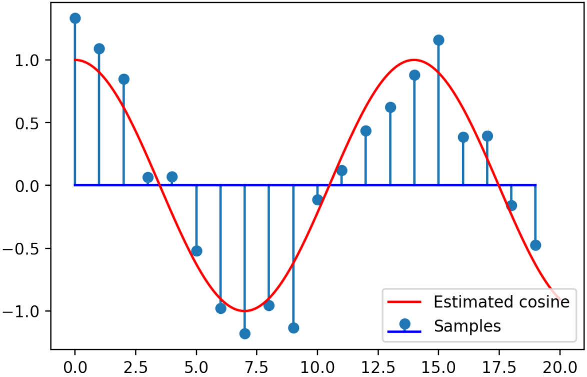
Numerical simulation

The likelihood function is:



Numerical simulation

```
/home/ncleju/.local/bin/pweave:15: UserWarning: In Matplotlib
individual lines on a stem plot will be added as a LineCollec
instead of individual lines. This significantly improves the
performance of a stem plot. To remove this warning and switch
new behaviour, set the "use_line_collection" keyword argument
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

Loss function

- The distance $d(\mathbf{r}, \mathbf{s}_\Theta)$ is known as the "**loss function**" in machine learning terminology
  - the Euclidean distance = the "**Mean Squared Error**" (MSE) loss function
- For a given $\mathbf{r}$, the MSE loss $= \frac{1}{N} d(\mathbf{r}, \mathbf{s}_\Theta)$
- Other loss functions are used in different scenarios

Multiple parameters

- What if we have more than one parameter?
  - e.g. unknown parameters are the amplitude, frequency and the initial phase of a cosine:
$$s_{\hat{}}(t) = A\cos(2\pi ft + \phi)$$
- We can consider the parameter Θ to be a vector:
$$\boldsymbol{\Theta} = [\Theta_1, \Theta_2, ...\Theta_M]$$
  - e.g. $\boldsymbol{\Theta} = [\Theta_1, \Theta_2, \Theta_3] = [A, f, \phi]$

Gradient Descent

- How to estimate the parameters $\mathbf{\Theta}$ in complicated cases?
  - e.g. in real life applications
  - usually there are many parameters ($\mathbf{\Theta}$ is a vector)
- Typically it is impossible to get the optimal values directly
- Improve them iteratively with **Gradient Descent** algorithm or its variations

Gradient Descent procedure

1. Start with some random parameter values $\mathbf{\Theta}^{(0)}$
2. Repeat for each iteration k:
    2.1 Compute loss value $L(\mathbf{\Theta}^{(k)})$
    2.2 Compute derivative $\frac{\partial L}{\partial \Theta_i^{(k)}}$ for each Θ_i
    2.3 Update all values Θ_i by subtracting the derivative

$$\Theta_i^{(k+1)} = \Theta_i^{(k)} - \mu \frac{\partial L}{\partial \Theta_i^{(k)}}$$

    - or, in vector form:

$$\mathbf{\Theta}^{(k+1)} = \mathbf{\Theta}^{k} - \mu \frac{\partial L}{\partial \mathbf{\Theta}^{(k)}}$$

3. Until termination criterion (e.g. parameters don't change much)

Gradient Descent explained

- Explanations at blackboard
- Simple example: logistic regression on 2D-data
    - maybe do example at blackboard

Neural Networks

- The most prominent example is **Artificial Neural Networks** (a.k.a. Neural Networks, Deep Learning, etc.)
  - Can be regarded as ML estimation
  - Use loss function (typically not MSE, but others)
  - Use Gradient Descent to update parameters
  - State-of-the-art applications: image classification/recognition, automated driving etc.
- More info on neural networks / machine learning:
  - look up online courses, books (e.g. prof. Iulian Ciocoiu's book)
  - join the IASI AI Meetup

II.3 Bayesian estimation

Prior distribution

- Suppose we know beforehand a distribution of Θ, $w(\Theta)$
  - we know beforehand how likely it is to have a certain value
  - known as *a priori* distribution or *prior* distribution
- The estimation must take it into account
  - the estimate will be slightly "moved" towards more likely values
- Known as "Bayesian estimation"
  - Thomas Bayes = discovered the Bayes rule
  - Stuff related to Bayes rule are often named "Bayesian"

Cost function

- The **estimation error** is the difference between the estimate $\hat{\Theta}$ and the true value Θ

$$\epsilon = \hat{\Theta} = \Theta$$

- The **cost function** $C(\epsilon)$ assigns a cost to each possible estimation error
  - when $\epsilon = 0$, the cost $C(0) = 0$
  - small errors ϵ have small costs
  - large errors ϵ have large costs

- Usual types of cost functions:
  - Quadratic: $C(\epsilon) = \epsilon^2 = \left(\hat{\Theta} - \Theta\right)^2$
  - Uniform ("hit or miss"): $C(\epsilon) = \begin{cases} 0, & \text{if } |\epsilon| = |\hat{\Theta} - \Theta| \leq E \\ 1, & \text{if } |\epsilon| = |\hat{\Theta} - \Theta| > E \end{cases}$
  - Linear: $C(\epsilon) = |\epsilon| = |\hat{\Theta} - \Theta|$
  - draw them at whiteboard

The Bayesian risk

- For each pair of values $\mathbf{r}$ and Θ, $w(\mathbf{r};\Theta)$ tells us how likely it is to have them
- Multiplying with $C(\epsilon$ gives us the cost, for each $\mathbf{r}$ and Θ

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

- Integrating over Θ gives the cost for a certain $\mathbf{r}$

$$\int_{-\infty}^{\infty} C(\epsilon)w(\mathbf{r};\Theta)d\Theta$$

- Further integrating also over $\mathbf{r}$ gives the global cost for all $\mathbf{r}$ and all Θ

$$R = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} C(\epsilon)w(\mathbf{r};\Theta)d\Theta d\mathbf{r}$$

Minimizing the risk

- We want to minimize the risk R
- Bayes rule: $w(\mathbf{r};\Theta) = w(\Theta|\mathbf{r})w(\mathbf{r})$
- Replacing in R, we obtain

$$
\begin{aligned}
R &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} C(\epsilon)w(\Theta|\mathbf{r})w(\mathbf{r})d\Theta d\mathbf{r} \\
&= \int_{-\infty}^{\infty} w(\mathbf{r})\left[\int_{-\infty}^{\infty} C(\epsilon)w(\Theta|\mathbf{r})d\Theta\right]d\mathbf{r}
\end{aligned}
$$

- Since $w(\mathbf{r}) \geq 0$, minimizing the inner integral will minimize R

$$
I = \int_{-\infty}^{\infty} C(\epsilon)w(\Theta|\mathbf{r})d\Theta
$$

- Next, we'll replace $C(\epsilon)$ with its definition and derivate over $\hat{\Theta}$
  - Attention: $\hat{\Theta}$, not Θ!

MMSE estimator

- When the cost function is quadratic $C(\epsilon) = \epsilon^2 = \left(\hat{\Theta} - \Theta\right)^2$

$$I = \int_{-\infty}^{\infty} (\hat{\Theta} - \Theta)^2 w(\Theta|\mathbf{r}) d\Theta$$

- We want the $\hat{\Theta}$ that minimizes I, so we derivate

$$\frac{dI}{d\hat{\Theta}} = 2 \int_{-\infty}^{\infty} (\hat{\Theta} - \Theta) w(\Theta|\mathbf{r}) d\Theta = 0$$

- Equivalent to

$$\hat{\Theta} \underbrace{\int_{-\infty}^{\infty} w(\Theta|\mathbf{r})}_{1} d\Theta = \int_{-\infty}^{\infty} \Theta w(\Theta|\mathbf{r}) d\Theta$$

- The **Minimum Mean Squared Error (MMSE)** estimator is

$$\hat{\Theta} = \int_{-\infty}^{\infty} \Theta \cdot w(\Theta|\mathbf{r}) d\Theta$$

Interpretation

- $w(\Theta|\mathbf{r})$ is the **posterior** (or **a posteriori**) distribution
    - it is the distribution of Θ after we know the data we received
    - the prior distribution $w(\Theta)$ is the one before knowing any data
- The MMSE estimation is the **average value** of the posterior distribution

The MAP estimator

- When the cost function is uniform
  $$C(\epsilon) = \begin{cases} 0, & \text{if } |\epsilon| = |\hat{\Theta} - \Theta| \leq E \\ 1, & \text{if } |\epsilon| = |\hat{\Theta} - \Theta| > E \end{cases} \qquad \text{\$\$}\backslash\text{begin}\{\text{split}\}$$
- Keep in mind that $\Theta = \hat{\Theta} - \epsilon$
- We obtain

$$I = \int_{-\infty}^{\hat{\Theta}-E} w(\Theta|\mathbf{r})d\Theta + \int_{Th\hat{e}ta+E}^{\infty} w(\Theta|\mathbf{r})d\Theta$$

$$I = 1 - \int_{\hat{\Theta}-E}^{\hat{\Theta}+E} w(\Theta|\mathbf{r})d\Theta$$

The MAP estimator

- To minimize I, we must maximize $\int_{\hat{\Theta}-E}^{\hat{\Theta}+E} w(\Theta|\mathbf{r}) d\Theta$, the integral around point $\hat{\Theta}$
- For E a very small, the function $w(\Theta|\mathbf{r})$ is approximately constant, so we pick the point where the function is maximum
- The **Maximum A Posteriori (MAP)** estimator is

$$\hat{\Theta} = \arg\max w(\Theta|\mathbf{r})$$

- $\arg\max$ = "the value which maximizes the function"
  - $\max f(x)$ = the maximum value of a function
  - $\arg\max f(x)$ = the x for which the function reaches its maximum

Interpretation

- The MAP estimator chooses Θ as the value where the posterior distribution is maximum
- The MMSE estimator chooses Θ as average value of the posterior distribution

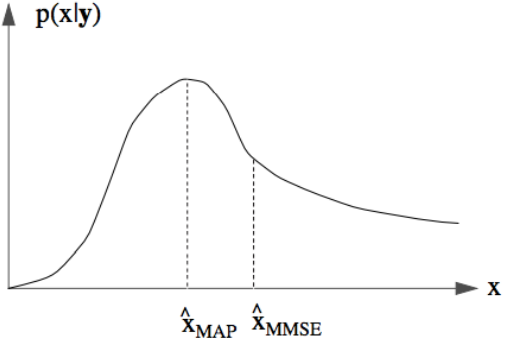


Figure 1: MAP vs MMSE estimators

Finding the posterior distribution

- That's cool, but how do we find this posterior distribution $w(\Theta|\mathbf{r})$?
- Use the Bayes rule

$$w(\Theta|\mathbf{r}) = \frac{w(\mathbf{r};\Theta)}{w(\mathbf{r})} = \frac{w(\mathbf{r}|\Theta) \cdot w(\Theta)}{w(\mathbf{r})}$$

- Since $w(\mathbf{r})$ is constant for a given $\mathbf{r}$ the MAP estimator is

$$\hat{\Theta} = \arg\max w(\Theta|\mathbf{r}) = \arg\max w(\mathbf{r}|\Theta) w(\Theta)$$

- The MAP estimator is the one which **maximizes** the likelihood of the observed data, **but multiplying with the prior distribution** $w(\Theta)$
- The MMSE estimator is the **average** of the same thing

Relation with Maximum Likelihood Estimator

- The ML estimator was just $\arg\max w(\mathbf{r}|\Theta)$
- The MAP estimator $=$ like the ML estimator but multiplied with the prior distribution $w(\Theta)$
- If $w(\Theta)$ is a constant, the MAP estimator reduces to ML
  - $w(\Theta) =$ constant means all values Θ are equally likely
  - i.e. we don't have a clue where the real Θ might be
- The MMSE estimator $=$ like MAP, but don't take the *argmax* of the function, but its average value

Relation with Detection

- The minimum probability of error criterion $\frac{w(r|H_1)}{w(r|H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \frac{P(H_0)}{P(H_1)}$

- It can be rewritten as $w(r|H_1) \cdot P(H_1) \underset{H_0}{\overset{H_1}{\gtrless}} w(r|H_0)P(H_0)$
  - i.e. choose the hypothesis where $w(r|H) \cdot P(H)$ is maximum
  - $w(r|H_1)$, $w(r|H_0)$ are the likelihood of observed data
  - $P(H_1)$, $P(H_0)$ are the prior probabilities (known beforehand)

- The MAP estimator is where $w(\mathbf{r}|\Theta)w(\Theta)$ is maximum
  - $w(\mathbf{r}|\Theta)$ is the likelihood of observed data
  - $w(\Theta)$ is the prior distribution (known beforehand)

- 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 of the whole function

2018-2019 Exam

- Chapter ends here for 2018-2019 exam. Following slides not needed.

Exercise

Exercise: constant value, 3 measurement, Gaussian same σ

- We want to estimate today's temperature in Sahara
- Our thermometer reads 40 degrees, but the value was affected by Gaussian noise $\mathcal{N}(0, \sigma^2 = 2)$ (crappy thermometer)
- We know that this time of the year, the temperature is around 35 degrees, with a Gaussian distribution $\mathcal{N}(35, \sigma^2 = 2)$.
- Estimate the true temperature using ML, MAP and MMSE estimators

Exercise

Exercise: constant value, 3 measurements, Gaussian same σ

- What if he have three thermometers, showing 40, 38, 41 degrees

Exercise: constant value, 3 measurements, Gaussian different σ

- What if the temperature this time of the year has Gaussian distribution $\mathcal{N}(35, \sigma_2^2 = 3)$
  - different variance, $\sigma_2 \neq \sigma$

General signal in AWGN

- Consider that the true underlying signal is $s_\Theta(t)$
- Consider AWGN noise $\mathcal{N}(\mu = 0, \sigma^2)$.
- As in Maximum Likelihood function, overall likelihood function

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

- But now this function is also **multiplied with** $w(\Theta)$

$$w(\mathbf{r}|\Theta) \cdot w(\Theta)$$

General signal in AWGN

- MAP estimator is the argument that maximizes this product

$$\hat{\Theta}_{MAP} = \arg\max w(\mathbf{r}|\Theta) w(\Theta)$$

- Taking logarithm

$$\begin{aligned}\hat{\Theta}_{MAP} &= \arg\max \ln\left(w(\mathbf{r}|\Theta)\right) + \ln\left(w(\Theta)\right)\\ &= \arg\max -\frac{\sum(r_i - s_{\Theta}(t_i))^2}{2\sigma^2} + \ln\left(w(\Theta)\right)\end{aligned}$$

Gaussian prior

- If the prior distribution is also Gaussian $\mathcal{N}(\mu_\Theta, \sigma_\Theta^2)$

$$\ln\left(w(\Theta)\right) = -\frac{\sum(\Theta - \mu_\Theta)^2}{2\sigma_\Theta^2}$$

- MAP estimation becomes

$$\hat{\Theta}_{MAP} = \arg\min \frac{\sum(r_i - s_\Theta(t_i))^2}{2\sigma^2} + \frac{\sum(\Theta - \mu_\Theta)^2}{2\sigma_\Theta^2}$$

- Can be rewritten as

$$\hat{\Theta}_{MAP} = \arg\min d(\mathbf{r}, s_\Theta)^2 + \underbrace{\frac{\sigma^2}{\sigma_\Theta^2}}_{\lambda} \cdot d(\Theta, \mu_\Theta)^2$$

Interpretation

- MAP estimator with Gaussian noise and Gaussian prior

$$\hat{\Theta}_{MAP} = \arg\min d(\mathbf{r}, s_\Theta)^2 + \underbrace{\frac{\sigma^2}{\sigma_\Theta^2}}_{\lambda} \cdot d(\Theta, \mu_\Theta)^2$$

- $\hat{\Theta}_{MAP}$ is close to its expected value μ_Θ and it makes the true signal close to received data **r**
    - Example: "search for a house that is close to job and close to the Mall"
    - λ controls the relative importance of the two terms
- Particular cases
    - σ_Θ very small = the prior is very specific (narrow) = λ large = second term very important = $\hat{\Theta}_{MAP}$ close to μ_Θ
    - σ_Θ very large = the prior is very unspecific = λ small = first term very important = $\hat{\Theta}_{MAP}$ close to ML estimation

Applications

- In general, practical applications:
  - can use various prior distributions
  - estimate **multiple parameters** (a vector of parameters)
- Applications
  - denoising of signals
  - signal restoration
  - signal compression

Estimator bias

- How good is an estimator?
  - Many ways to characterize
- An estimator $\hat{\Theta}$ is a **random variable**
  - can have different values, because it is computed based on the received samples, which depend on noise
  - example: in lab, try on multiple computers $=>$ slightly different results
- As a random variable, it has:
  - an average value (expected value): $E\left\{\hat{\Theta}\right\}$
  - a variance: $E\left\{(\hat{\Theta}-\Theta)^2\right\}$

Estimator bias

- **Unbiased** estimator = if the average value of the estimator is the true value of Θ
$$E\left\{\hat{\Theta}\right\} = \Theta$$
- **Biased** estimator = if the average value of the estimator is different from the true value Θ
  - the difference $E\left\{\hat{\Theta}\right\} - \Theta$ is called **the bias** of the estimator

Estimator bias

- Example: for constant signal A with AWGN noise (zero-mean), ML estimator is $\hat{A}_{ML} = \frac{1}{N}\sum_i r_i$
- Then:

$$
\begin{aligned}
E\left\{\hat{A}_{ML}\right\} &= \frac{1}{N} E\left\{\sum_i r_i\right\} \\
&= \frac{1}{N}\sum_{i=1}^{N} E\left\{r_i\right\} \\
&= \frac{1}{N}\sum_{i=1}^{N} E\left\{A + noise\right\} \\
&= \frac{1}{N}\sum_{i=1}^{N} A \\
&= A
\end{aligned}
$$

- This estimator in unbiased

Estimator variance

- Unbiased estimators are good, but if the **variance** of the estimator is large, then estimated values can be far from the true value
- We prefer estimators with **small variance**, even if maybe slightly biased