Decision and Estimation in Data Processing

Chapter III. Elements of Estimation Theory

III.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
    - Even complicated problems:
        - Record speech signal, estimate/decide what word is pronounced

Estimation vs Decision

- Consider the following estimation problem:

  We receive a signal $r(t) = A + noise$, 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 $r(t) = s_\Theta(t) + \text{noise}$
    - it is affected by noise
    - it depends on the unknown parameter Θ
- We consider **N samples** from $r(t)$, taken at some sample times t_i

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

- The samples depend on the value of Θ

Available data

- 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 **r** is a N-dimensional random variable that depends on Θ (and the noise)

  - It has a N-dimensional distribution that depends on Θ

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

  - Equal to the product of all $w_i(r_i|\Theta)$

  $$w(\mathbf{r}|\Theta) = w_1(r_1|\Theta) \cdot w_2(r_2|\Theta) \cdot ... \cdot w_N(r_N|\Theta)$$

Two types of estimation

▶ We consider two types of estimation:

1. **Maximum Likelihood Estimation (MLE)**: Besides **r**, nothing else is known about the parameter Θ, except maybe some allowed range (e.g. $\Theta > 0$)

2. **Bayesian Estimation**: Besides **r**, we know a **prior** distribution $p(\Theta)$ for Θ, which tells us the values of Θ that are more likely than others

II.2 Maximum Likelihood estimation

Maximum Likelihood definition

- When no distribution is known except **r**, we use a method known as **Maximum Likelihood estimation (MLE)**

- We define the **likelihood** of a parameter value Θ, given the available observations **r** as:

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

- $L(\Theta|\mathbf{r})$ is the likelihood function

- "The plausibility of a parameter value Θ given some measurements $\mathbf{r} =$ $=$ the probability density of generating **r** if the true value would be Θ"

- Compare with formula in Chapter 2, slide 20
    - it is the same
    - here we try to "guess" Θ, there we "guessed" H_i

Maximum Likelihood (ML) Estimation:

- The estimate $\hat{\Theta}_{ML}$ is **the value that maximizes the likelihood, given the observed data r**

  - i.e. the value that maximizes $L(\Theta|\mathbf{r})$, i.e. maximize $w(\mathbf{r}|\Theta)$

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

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

▶ General mathematical notations:

    ▶ $\arg\max_x f(x) =$ "the value x which maximizes the function f(x)"

    ▶ $\max_x f(x) =$ "the maximum value of the function f(x)"

Maximum Likelihood estimation vs decision

- ▶ Very similar with decision problem!
- ▶ ML decision criterion:
    - ▶ "pick the hypothesis with a higher likelihood":

$$\frac{L(H_1|r)}{L(H_0|r)} = \frac{w(r|H_1)}{w(r|H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} 1$$

- ▶ ML estimation
    - ▶ "pick the value which maximizes the likelihood"

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

- How to solve the maximization problem?
  - i.e. how to find the estimate $\hat{\Theta}_{ML}$ which maximizes $L(\Theta|\mathbf{r})$
- Find maximum by setting derivative to 0

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

- We can also maximize the **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

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

2. Set the condition that derivative of $L(\Theta|\mathbf{r})$ or $\ln((L(\Theta|\mathbf{r}))$ is 0

$$\frac{dL(\Theta|\mathbf{r})}{d\Theta} = 0, \text{ or } \frac{d\ln(L(\Theta))}{d\Theta} = 0$$

3. Solve and find the value $\hat{\Theta}_{ML}$

4. Check that second derivative at point $\hat{\Theta}_{ML}$ is negative, to check that point is a maximum

   ▶ because derivative $= 0$ for both maximum and minimum points
   ▶ we'll sometimes skip this, for brevity

Examples:

▶ Estimating a constant signal in gaussian noise:

Find the ML estimate of a constant value $s_\Theta(t) = A$ 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 $\hat{A}_{ML}$ is the average value of the samples

    ▶ not surprisingly, what other value would have been more likely?
    ▶ that's literally what "expected value" means

Estimate of a constant value
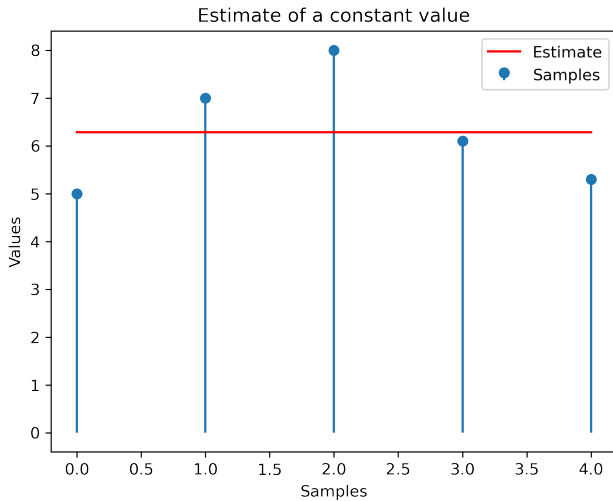
- **Estimation = curve fitting**
    - we're finding the best fitting of $s_\Theta(t)$ through the data **r**
- From the previous graphical example:
    - we have some data **r** = some points
    - we know the shape of the signal = a line (constant A)
    - we're fitting the best line through the data

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 value $\mu = s_\Theta(t_i)$ and variance σ^2
- Overall likelihood function = product of likelihoods for each sample r_i

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

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

- The log-likelihood is

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

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

$$\hat{\Theta}_{ML} = \arg \max_{\Theta} L(\Theta|\mathbf{r}) = \arg \min \sum (r_i - s_\Theta(t_i))^2$$

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

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

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

- ML estimation can be rewritten as:

$$\hat{\Theta}_{ML} = \arg\max_{\Theta} L(\Theta|\mathbf{r}) = \arg\min_{\Theta} d(\mathbf{r}, \mathbf{s}_{\Theta})^2$$

- ML estimate $\hat{\Theta}_{ML}$ = the value that makes $s_{\Theta}(t_i)$ **closest to the received values r**

    - closer = beter fit = more likely
    - closest = best fit = most likely = maximum likelihood

- ML estimation in AWGN noise = **minimization of distance**

- Hey, we had the same interpretation with ML decision!

    - but for decision, we choose the minimum out of 2 options
    - here, we choose the minimum out of all possible options

- Same interpretation applies 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

Procedure for ML estimation in AWGN noise:

1. Write the expression for the (squared) distance:

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

2. We want it minimal, so set derivative to 0:

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

3. Solve and find the value $\hat{\Theta}_{ML}$

4. Check that second derivative at point $\hat{\Theta}_{ML}$ is positive, to check that point is a minimum

   ▶ we'll sometimes skip this, for brevity
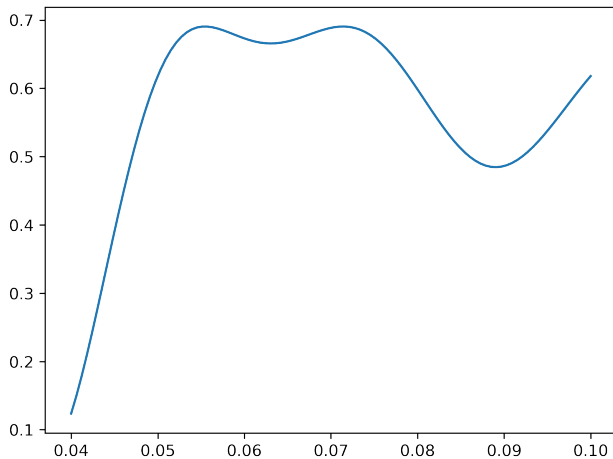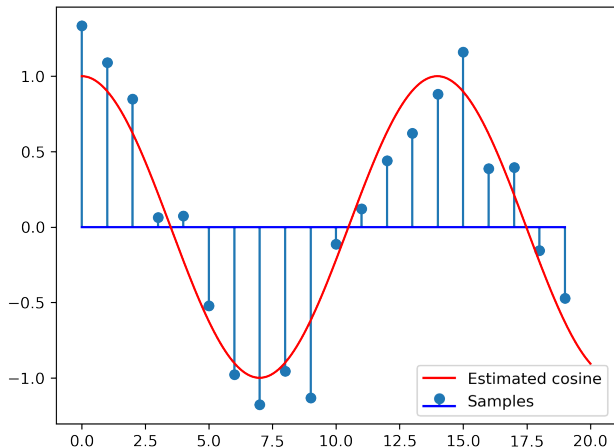
Estimating the frequency f of a cosine signal

- ▶ Find the Maximum Likelihood estimate of the frequency f of a cosine signal $s_\Theta(t) = cos(2\pi f t_i)$, 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.

The likelihood function is:

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:
    $$s^{\cdot}(t) = A\cos(2\pi f t + \phi)$$

- ▶ We can consider the parameter Θ to be a vector:

$$\boldsymbol{\Theta} = [\Theta_1, \Theta_2, ...\Theta_M]$$

  - ▶ e.g. $\boldsymbol{\Theta} = [\Theta_1, \Theta_2, \Theta_3] = [A, f, \phi]$

Multiple parameters

- We solve with the same procedure, but instead of one derivative, we have M derivatives

- We solve the system:

$$\begin{cases} \frac{\partial L}{\partial \Theta_1} = 0 \\ \frac{\partial L}{\partial \Theta_2} = 0 \\ \dots \\ \frac{\partial L}{\partial \Theta_M} = 0 \end{cases}$$

  - sometimes difficult to solve

Gradient Descent

▶ How to estimate the parameters Θ in complicated cases?

  ▶ e.g. in real life applications
  ▶ usually there are many parameters (Θ is a vector)

▶ Typically it is impossible to get the optimal values directly by solving the system

▶ 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 function $L(\mathbf{\Theta}^{(k)}|\mathbf{r})$

    2.2 Compute derivatives $\frac{\partial L}{\partial \Theta_i^{(k)}}$ for each Θ_i (**"gradient"**)

    2.3 Update all values Θ_i by subtracting the derivative ("**descent**")

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

    ▶ 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 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
  - ▶ join the IASI AI Meetup

Estimator bias and variance

▶ How good is an estimator?

▶ 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: $E\left\{(\hat{\Theta} - \Theta)^2\right\}$

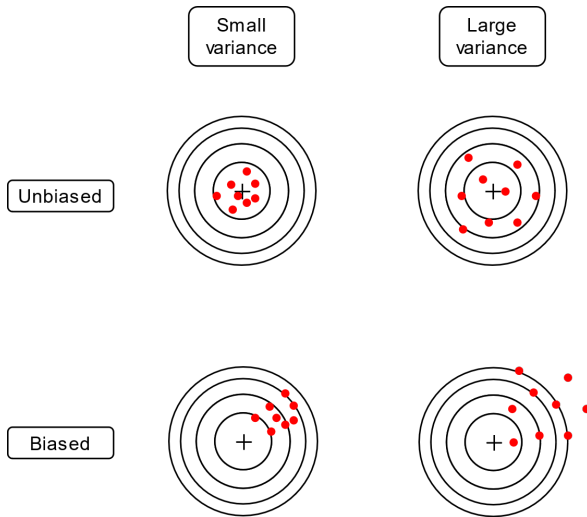# Estimator bias and variance

Figure 1: Estimator bias and variance

Estimator bias

▶ The **bias** of an estimator $\hat{\Theta}$ = difference between the estimator's average value and the true value

$$Bias = E\left\{\hat{\Theta}\right\} - \Theta$$

▶ Estimator is **unbiased** = the average value of the estimator is the true value of Θ

$$E\left\{\hat{\Theta}\right\} = \Theta$$

▶ Estimator is **biased** = the average value of the estimator is different from the true value Θ

  ▶ the difference $E\left\{\hat{\Theta}\right\} - \Theta$ is **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

- The **variance** of an estimator measures the "spread" of the estimator around its average
  - that's the definition of 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

II.3 Bayesian estimation

Bayesian estimation

- **Bayesian estimation** considers extra factors alongside $w(\mathbf{r}|\Theta)$ in the estimation:
  - a prior distribution $w(\Theta)$
  - possibly some cost function
- This makes it the estimation version of the MPE and MR decision criteria

▶ Conceptually, Bayesian estimation consists of two major steps:

1. Finding the **posterior distribution** $w(\Theta|\mathbf{r})$
2. Estimating a value from the distribution, based on a **cost function**

Bayesian estimation

▶ We define the **posterior** probability density of Θ, given the known observations **r**, using the **Bayes rule**:

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

▶ Explanation of the terms:

  ▶ Θ is the unknown parameter
  ▶ **r** are the observations that we have
  ▶ $w(\Theta|\mathbf{r})$ is the probability of a certain value Θ to be the correct one, given our current observations **r**;
  ▶ $w(\mathbf{r}|\Theta)$ is the likelihood function
  ▶ $w(\Theta)$ is the **prior distribution** of Θ, i.e. what we know about Θ even in the absence of evidence
  ▶ $w(\mathbf{r})$ is a scaling constant, which makes the integral of the resulting function be 1 (like for any distribution)

Bayesian estimation

- With MLE estimation, we only have the term $w(\mathbf{r}|\Theta)$. When viewed as a function of Θ, this is not a distribution of Θ. It's just something we want to maximize.

- Bayesian estimation, however, uses $w(\Theta|\mathbf{r})$, which **is** the actual probability distribution of the possible values of Θ

Bayes rule

- The Bayes rule shows that the probability of a value 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 role)

- Known as "Bayesian estimation"

  - Thomas Bayes = discovered the Bayes rule
  - Stuff related to Bayes rule are often named "Bayesian"

The prior distribution

- The role of the prior distribution $w(\Theta)$ is to express what we know beforehand about Θ
  - we know beforehand how likely it is to have a certain value
  - known as *a priori* distribution or *prior* distribution
- Bayesian estimation takes the prior information into account, alongside the measurements
  - the estimate will be slightly "moved" towards more likely values

The MAP estimator

▶ Suppose we know $w(\Theta|\mathbf{r})$. What is our estimate?

▶ Let's pick the value with the highest probability

▶ The **Maximum A Posteriori (MAP)** estimator:

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

▶ The MAP estimator chooses Θ as the value where the posterior distribution $w(\Theta|\mathbf{r})$ is maximum

▶ The MAP estimator maximizes the likelihood of the observed data **but multiplied with the prior distribution** $w(\Theta)$

Image example here

Relation with Maximum Likelihood Estimator

▶ The ML estimator:

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

▶ The MAP estimator:

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

▶ The ML estimator is a particular case of MAP when $w(\Theta)$ is a constant

    ▶ $w(\Theta) =$ constant means all values Θ are equally likely

    ▶ i.e. we don't have a clue where the real Θ might be

Relation with Detection

- The MPE 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 $w(r|H_i) \cdot P(H_i)$ is maximum

- **MPE decision criterion**: pick hypothesis which maximizes $w(r|H_i) \cdot P(H_i)$

  - out of the two possible hypotheses

- **The MAP estimator**: pick value which maximizes $w(\mathbf{r}|\Theta) \cdot w(\Theta)$

  - out of all possible values of Θ

- Same principle!

Cost function

▶ Let's find an equivalent for the Minimum Risk criterion. We need an equivalent for the costs C_{ij}

▶ The **estimation error** = 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

Cost function

- Usual types of cost functions:
    - Quadratic:
    $$C(\epsilon) = \epsilon^2 = \left(\hat{\Theta} - \Theta\right)^2$$

    - Uniform ("hit or miss"):
    $$C(\epsilon) = \begin{cases} 0, & \text{if } |\epsilon| = |\hat{\Theta} - \Theta| \le E \\ 1, & \text{if } |\epsilon| = |\hat{\Theta} - \Theta| > E \end{cases}$$

    - Linear:
    $$C(\epsilon) = |\epsilon| = |\hat{\Theta} - \Theta|$$

- Draw them at whiteboard

▶ The cost function $C(\epsilon)$ is the equivalent of the costs C_{ij} at detection

  ▶ for detection we only had 4 costs: C_{00}, C_{01}, C_{10}, C_{11}
  ▶ now we have a cost for all possible estimation errors ϵ

▶ The cost function guides which value to choose from $w(\Theta|\mathbf{r})$

The importance of the cost function

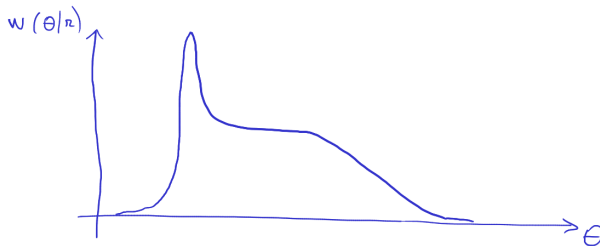▶ Consider the following posterior distribution

Figure 2: Unbalanced posterior distribution

▶ Which is the MAP estimate?

▶ Supposing we have the following cost function, does your estimate change ?:

  ▶ if your estimate $\hat{\Theta}$ is $<$ then the real Θ, you pay 1000$
  ▶ if your estimate $\hat{\Theta}$ is $>$ then the real Θ, you pay 1$

The importance of the cost function

- Choosing one particular value $\hat{\Theta}$ from the distribution of possible values is driven by the cost function
- The most probable value is not always the best
- The best value is the one which leads to the smallest average cost

The Bayesian risk

- The posterior distribution $w(\Theta|\mathbf{r})$ tells us the probability of a certain value $\hat{\Theta}$ to be the correct one of Θ
- Picking a certain estimate value $\hat{\Theta}$ implies a certain error ϵ
- The error implies a certain cost $C(\epsilon)$
- The **risk** = the average cost = $C(\epsilon) \times$ the probability:

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

The Bayes estimator

▶ We need to pick the value $\hat{\Theta}$ which **minimizes the expected cost** R

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

▶ To find it, replace $C(\epsilon)$ with its definition and derivate over $\hat{\Theta}$

    ▶ Attention: derivate with respect to $\hat{\Theta}$, not Θ!

MMSE estimator

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

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

▶ We want the $\hat{\Theta}$ that minimizes R, so we derivate

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

▶ Equivalent to

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

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

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

Interpretation

- **The MMSE estimator**: the estimator $\hat{\Theta}$ is the **average value** of the posterior distribution $w(\Theta|\mathbf{r})$

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

  - MMSE = "Minimum Mean Squared Error"
  - average value = sum (integral) of every Θ times its probability $w(\Theta|\mathbf{r})$

- The MMSE estimator is obtained from the posterior distribution $w(\Theta|\mathbf{r})$ considering the quadratic cost function

The MAP estimator

▶ When the cost function is uniform:

$$C(\epsilon) = \begin{cases} 0, \text{ if } |\epsilon| = |\hat{\Theta} - \Theta| \leq E \\ 1, \text{ if } |\epsilon| = |\hat{\Theta} - \Theta| > E \end{cases}$$

▶ Keep in mind that $\Theta = \hat{\Theta} - \epsilon$

▶ We obtain

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

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

The MAP estimator

- To minimize C, 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** = the value $\hat{\Theta}$ which maximizes $w(\Theta|\mathbf{r})$

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

Interpretation

▶ The MAP estimator chooses Θ as the value where the posterior distribution is maximum

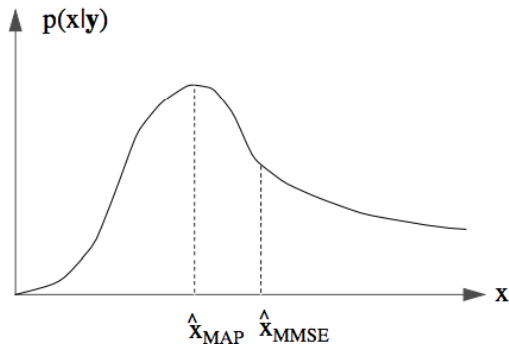▶ The MMSE estimator chooses Θ as average value of the posterior distribution

Figure 3: MAP vs MMSE estimators

Relationship between MAP and MMSE

- The MAP estimator = minimizing the average cost, using the uniform cost function
  - similar with the MPE decision criteria = MR when all costs are same
- The MMSE estimator = minimizing the average cost, using the quadratic cost function
  - similar to MR decision criteria, but more general

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: 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, $\sigma_2 \neq \sigma$

- 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** $w(\Theta)$

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

▶ MAP estimator is the argument that maximizes this product

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

▶ Taking logarithm

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

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 the expected value μ_Θ **and** it makes the true signal close to received data $\mathbf{r}$
  - 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

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

Sample applications

1. Single object tracking with Kalman filtering

▶ estimating an object's position through successive noisy measurements (e.g. consecutive frames in a video)

▶ at every new measurement, we have two distributions of the position:
   ▶ one given by the measurement itself, $w(r|\Theta)$
   ▶ one predicted based on position and speed from last moment
   ▶ both are presumed Gaussian, described only through average value and variance

▶ the two are combined via the Bayes rule $=>$ a more precise distribution $w(\Theta|r)$, also Gaussian

▶ the exact position is estimated with MMSE (average value of $w(\Theta|r)$

▶ $w(\Theta|r)$ + speed is used to predict the position at the next time moment

Single object tracking

Sample applications

2. Constrained Least Squares (CLS) image restoration

► We have an image I corrupted by noise (additive noise, missing pixels, blurring)

$$I_{noisy} = I_{true} + Z$$

► We can estimate the original image by solving:

$$\hat{I_{true}} = argmin_I \|I - I_{zg}\|_2 + \lambda \cdot \|HighPass\{I\}\|_2$$

► Examples:

   ► https://www.mathworks.com/help/images/deblurring-images-using-a-regularized-filter.html

   ► https://demonstrations.wolfram.com/ImageRestorationForDegradedImages

   ► Google it

Constrained Least Squares (CLS) image restoration