Notes on Estimation theory

Shreyas S

Part I

Metrology

Chapter 1

Metrology and Statistics

Uncertainty is considered to be an inseparable part of all types of measurements. Any measurement result without uncertainty is an abstract number. For example, a number with an absolute value, such as 10, cannot be the result of a measurement because numerically "10" means that all digits behind the last one are zero, i.e. 10.000.... Obviously, no measurement has this accuracy and therefore the number "10" itself is an abstract (hypothetical) number. In practice, we cannot measure a quantity free from uncertainty. Consequently, in metrology, the measurement results cannot be expressed as an absolute number, because of the uncertainty of this result. [5]

Chapter 2

Estimation Theory/Bayesian Statistics

We also try to understand some related estimation theory and strictly confine ourselves to the bounds of this project.

2.1 Probability Review [21]

2.1.1 Random Variables (RV)

A random variable is a numerical description of the outcome of a statistical experiment. A random variable that may assume only a finite number or an infinite sequence of values is said to be discrete; one that may assume any value in some interval on the real number line is said to be continuous. Since we don't know how to completely determine what value will occur, we can only specify probabilities of RV values occurring.

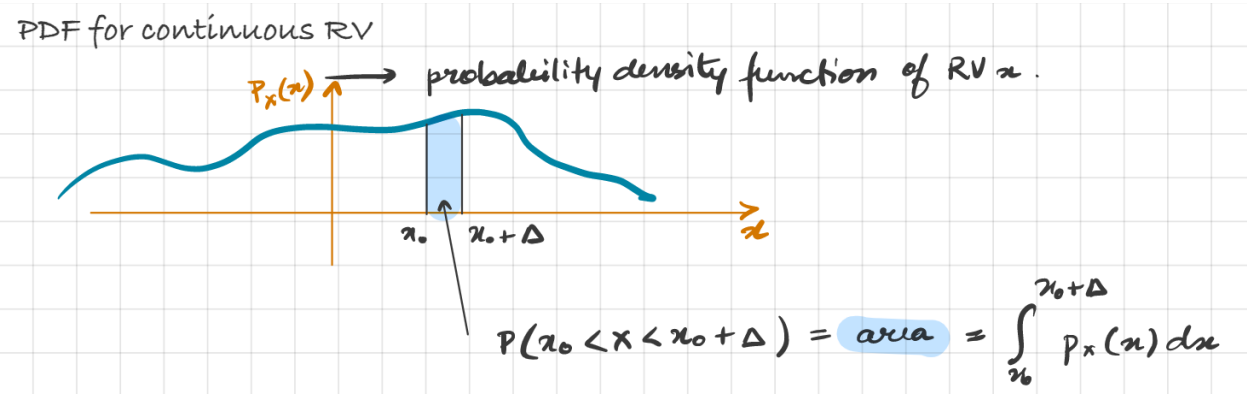


Figure 2.1: Probability of finding RV x in the interval $[x_0, x_0 + \Delta]$

2.1.2 Probability density function [14]

In probability theory, a probability density function (PDF), or density of a continuous random variable, is a function whose value at any given sample (or point) in the sample space (the set of possible values taken by the random variable) can be interpreted as providing a relative likelihood

that the value of the random variable would equal that sample. In other words, while the absolute likelihood for a continuous random variable to take on any particular value is 0 (since there is an infinite set of possible values to begin with), the value of the PDF at two different samples can be used to infer, in any particular draw of the random variable, how much more likely it is that the random variable would equal one sample compared to the other sample.

In a more precise sense, the PDF is used to specify the probability of the random variable falling within a particular range of values, as opposed to taking on any one specific value. This probability is given by the integral of this variable's PDF over that range, that is, it is given by the area under the density function but above the horizontal axis and between the lowest and greatest values of the range. The probability density function is non-negative everywhere, and its integral over the entire space is equal to 1.

For a continuous random variable x with a PDF P(x):

1. $P(x) \geq 0$
2. $\int_{-\infty}^{\infty} P(x)\,dx \ = 1$
3. More generally, for a set A, $P(x \in A) = \int_A P(x)\,dx$

2.1.3 Most commonly used PDF: Gaussian

A RV X with the following PDF is called a Gaussian RV:

$$p_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-m)^2/2\sigma^2} \tag{2.1}$$

where:

- m = Mean of RV X
- σ^2 = Variance of RV X
- σ = S.D of RV X

Notation: When X has a Gaussian PDF, we say

$$X \sim N(m, \sigma^2) \tag{2.2}$$

Gaussians are common because of CLT

2.1.4 Central Limit Theorem (CLT)

The sum of N independent RVs has a PDF that tends to be Gaussian (normally distributed) as $N \longrightarrow \infty$, even if the original variables themselves are not normally distributed

2.1.5 Joint PDF of RVs X and Y - $p_{XY}(x, y)$ [15]

Joint probability is the statistical measure where the likelihood of two events occurring together and at the same point in time are calculated. Because of this, it can only be applied to situations where more than one observation can be made at the same time. A bivariate distribution is a joint

distribution with two variables of interest. This can also be applied to numerous events or RVs being measured at one time (multivariate distribution).

For example, in the case of a bivariate distribution of X and Y, the probability that X lies in interval [a,b] and Y lies in interval [c,d] is given by:

$$Pr\{(a < X < b)\ and\ (c < Y < d)\} = \int_a^b \int_c^d p_{XY}(x, y)\ dxdy \tag{2.3}$$

This can be represented by the volume element shown in the figure below (2.2)

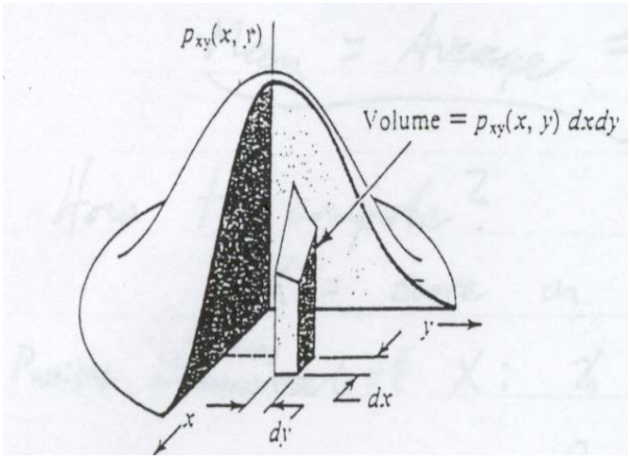


Figure 2.2: Graph from B. P. Lathi's book: Modern Digital & Analog Communication Systems

The probability of two events, X and B, both occurring simultaneously is expressed as:

$$P(X, Y) \tag{2.4}$$

If X, Y, Z are independent random variables, then

$$P(X, Y, Z) = P(X)P(Y)P(Z) \tag{2.5}$$

Notation:

- $P(X \cap Y) = P(X \text{ and } Y) = P(X, Y)$
- $P(X \cup Y) = P(X \text{ or } Y)$

2.1.6 Conditional PDFs of Two RVs

When we have two RVs, we often ask: What is the PDF of Y if X is constrained to take on a specific value? In other words, what is the PDF of Y conditioned on the fact that X is constrained to take on a specific value.

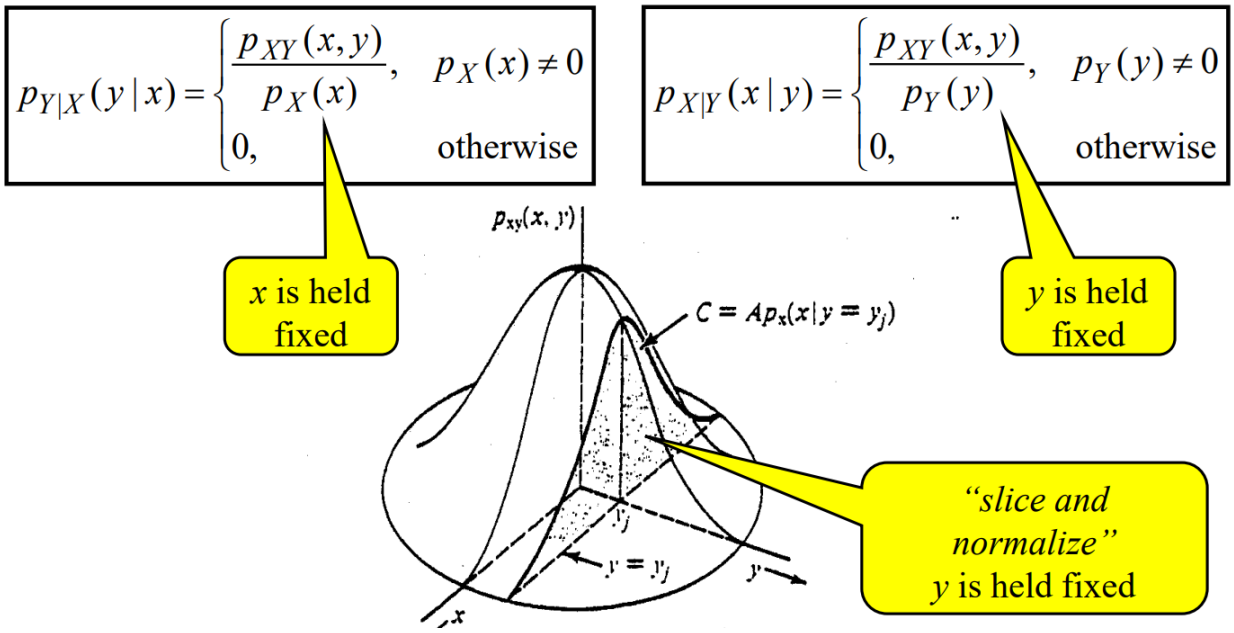


Figure 2.3: The graph shows the conditional PDF $p_{X|Y}(x|y = y_j)$

2.1.7 Conditional Probability [16]

Dependent events:

Events can be dependent, meaning they can be affected by previous events.
Example: Marbles in a bag
5 marbles are in a bag. 2 blue and 3 red. The probability of picking a blue marble is 2 in 5. But after taking one out, the chances change.
If in the first draw, we got a **red** marble, then the chance of picking a blue marble next is **2 in 4**.
If in the first draw, we got a **blue** marble, then the chance of picking a blue marble next is **1 in 4**.

Event A $\longrightarrow$ get a Blue marble first.

$$P(A) = 2/5$$

Event B $\longrightarrow$ get a Blue marble second, but we now have 2 choices:

- If we got a Blue Marble first the chance is now 1/4
- If we got a Red Marble first the chance is now 2/4

Probability of getting event B given event A $\longrightarrow P(B|A) = 1/4$

In summary, the conditional probability of an event B is the probability that the event will occur given the knowledge that an event A has already occurred. This probability is written $P(B|A)$, notation for the probability of B given A. In the case where events A and B are independent (where event A has no effect on the probability of event B), the conditional probability of event B given event A is simply the probability of event B, that is P(B).

Now, the probability of getting 2 blue marbles is: $P(A) \times P(B|A) = \frac{1}{10}$

$$P(A \text{ and } B) = P(A) \times P(B|A) \tag{2.6}$$

Probability of event A and event B equals the probability of event A times the probability of event B given event A.

Rearranging (2.6), we get:

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)} \tag{2.7}$$

Example: 70% of your friends like Chocolate $P(\text{Chocolate})$, and 35% like Chocolate AND like Strawberry $P(\text{Chocolate and Strawberry})$. What percent of those who like Chocolate also like Strawberry $P(\text{Strawberry}|\text{Chocolate})$?

$$P(\text{Strawberry}|\text{Chocolate}) = \frac{P(\text{Chocolate and Strawberry})}{P(\text{Chocolate})} = \frac{0.35}{0.7} = 50\% \tag{2.8}$$

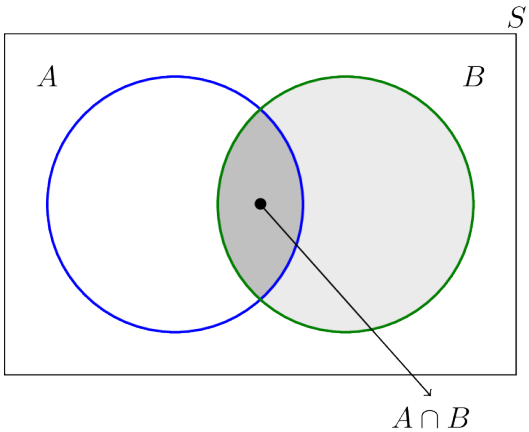
Note!

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)} \tag{2.9}$$

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} \tag{2.10}$$

Hence, $P(A|B) = P(B|A)$ only when P(A)=P(B). In general, $P(A|B) \neq P(B|A)$. The ordering of the conditioning is important. Additionally, $P(A \text{ and } B) = P(B \text{ and } A)$

Equation (2.10) can be represented by the following Venn diagram:



$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Independent events:

Events can be independent, meaning each event is not affected by other events.
Example: Coin Toss
Each coin toss is perfectly isolated. The outcome of the previous toss does not affect the present outcome. The chance of heads is still 50%

Two events A and B are independent iff $P(A, B) = P(A)P(B)$ (2.11)

For independent events, equation (2.7), becomes:

$$P(B|A) = \frac{P(A, B)}{P(A)} = \frac{P(A)P(B)}{P(A)} = P(B)$$

$$\therefore P(A|B) = P(A) \tag{2.12}$$
$$P(B|A) = P(B) \tag{2.13}$$

The definition of independence can be extended to the case of three or more events.

Three events A, B, and C are independent if **all** of the following conditions hold:

- P(A,B) = P(A)P(B)
- P(A,C) = P(A)P(C)
- P(B,C) = P(B)P(C)
- P(A,B,C) = P(A)P(B)P(C)

Warning! One common mistake is to confuse independence with being disjoint. These are different concepts. When two events A and B are disjoint, it means that if one of them occurs, the other one cannot occur, i.e, $A \cap B = \emptyset$. Therefore, event A usually gives a lot of information about event B, which means they are not independent.
For disjoint events:

$$P(A, B) = P(A \cap B) = 0 \neq P(A)P(B)$$

We see that disjoint events are not independent

Conditional Independence

Remember from (2.11) and (2.12), that A and B are independent if $P(A, B) = P(A)P(B)$, or equivalently, $P(A|B) = P(A)$. We can extend this concept to conditionally independent events. In particular,

Two events A and B are **conditionally independent** given an event C if:

$$P(A, B|C) = P(A|C)P(B|C) \tag{2.14}$$
$$P(A|B, C) = P(A|C) \tag{2.15}$$

The two above equations, are equivalent statements of conditional independence

(2.14) can be easily derived from (2.11). From (2.11), we know that if A and B are independent events, then

$$P(A, B) = P(A)P(B)$$

By conditioning on C, we obtain (2.14)

$$P(A, B|C) = P(A|C)P(B|C)$$

(2.15) can be derived in the following way:

$$P(A|B) = \frac{P(A, B)}{P(B)} \qquad \text{(From Eq(2.7))}$$

By conditioning on C, we obtain

$$P(A|B, C) = \frac{P(A, B|C)}{P(B|C)}$$

If A and B are conditionally independent given C, we know that:

$$P(A, B|C) = P(A|C)P(B|C) \qquad \text{(From Eq(2.14))}$$

$$\therefore P(A|B, C) = \frac{P(A, B|C)}{P(B|C)} = \frac{P(A|C)P(B|C)}{P(B|C)} = P(A|C)$$

Therefore, if A and B are conditionally independent given C, then $P(A|B, C) = P(A|C)$

2.1.8 Characterizing RVs, using Mean & Variance

PDFs completely describe RVs, however sometimes, they give us more information than is required. We can define a few characteristics of PDFs to help us compare different PDFs without actually comparing their functions over all space.

1. Mean of an RV (describes the centroid of the PDF)
2. Variance of an RV (Describes the spread of the PDF)
3. Correlation of RVs (Describes the "tilt" of joint PDFs)

Example: Rectangular distribution (Uniform PDF): We would like to measure the phase ϕ of a phase shifter. We first make a rudimentary measurement of ϕ and we measure it to be $\pi \pm 0.1\pi$. Here, $\pi = \phi$ and we define $\Delta/2$ to be 0.1π. We assume that a measurement of 0.90π, 0.91π, 0.92π ...1.09π, are equally likely. Statistically, this is the **Rectangular distribution**.

A rectangular/uniform distribution is a continuous probability distribution with a probability density function shaped like a rectangle. It is the simplest form of continuous probability distributions, due to its shape; it is also known as a rectangular distribution. It is defined in the following way:

$$p(\phi) = \begin{cases} 0, & \text{for } \phi < a \\ \frac{1}{(b-a)}, & \text{for } a \leq \phi \leq b \\ 0, & \text{for } \phi > b \end{cases} \tag{2.16}$$

where $p(\phi)$ is the uniform probability distribution function, a is the lower limit (in this example 0.90π), b is the upper limit (in this example 1.10π), and ϕ is the phase variable.

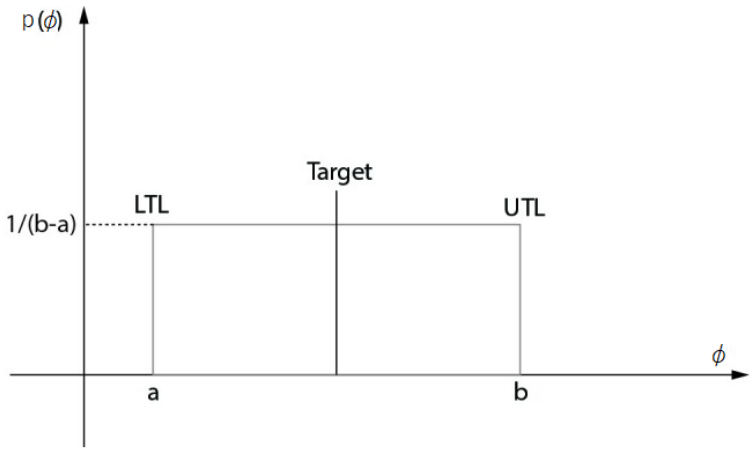


Figure 2.4: Uniform rectangular distribution. In contrast to other distributions, the probability density function of uniform distribution is constant between the upper (UTL) and lower target limits (LTL). Here $a = \phi - \Delta/2$ and $b = \phi + \Delta/2$

To intuitively understand why $p(\phi) = 1/\Delta$, let us assume that that the spectrum of measurable ϕ values is discrete, and that we can only measure n of them.

Since the probability distribution of measuring ϕ with an uncertainty of Δ is uniform, the probability of measuring a singular outcome (among the n possible ones), is 1 over n, or $P(\phi_{particular}) = \frac{1}{N}$.

Now, if the measurable ϕ spectrum was continuous, the probability of measuring a singular outcome (now among $b - a = (\phi + \Delta/2) - (\phi - \Delta/2) = \Delta$ possible ones), is $P(\phi_{particular}) = \frac{1}{\Delta}$.

This is the same as saying the following: The probability of measuring $\phi_{particular}$ in the interval $-\Delta/2 \leq \phi \leq \Delta/2$ has to be 1. Since the probability distribution is flat, the area of the rectangle ($\Delta \times height$) should be equal to 1. That is why the height of the rectangle should be $\frac{1}{b-a} = \frac{1}{\Delta} = P(\phi_{particular})$

We now seek to derive the mean, variance and standard deviation of a rectangular distribution. We start by assuming discreteness and then modify the equations to account for the continuity of phase measurements

Mean, variance and standard deviation assuming discreteness

As shown in (2.16), the probability density function is constant between a and b. Similar to all other probability distributions, the area under the curve of the uniform distribution is normalized and set to 1.

To intuitively understand, let us assume that the spectrum of measurable ϕ values is discrete, and that we have n of them. The mean, variance and standard deviation can be calculated in the

following way:

$$< \phi > = \frac{\sum_{i=1}^{n} \phi_i}{n} \tag{2.17}$$

$$(\delta\phi)^2 = \frac{\sum_{i=1}^{n} (\phi_i - < \phi >)^2}{n} \tag{2.18}$$

$$(\delta\phi) = \sqrt{\frac{\sum_{i=1}^{n} (\phi_i - < \phi >)^2}{n}} \tag{2.19}$$

Our intuitive understanding of the mean and the variance for the random variable ϕ is the following:

- The mean (a.k.a average a.k.a expected value), symbolically represented as: $< \phi >= \mu = E(\phi))$ is the sum of each term divided by the total number of terms.

- The variance (a.k.a $(\delta\phi)^2 = \sigma^2 = var(\phi) = E[(\phi - \mu)^2]$ is the sum of the squared distances of each term in the distribution from the mean (μ), divided by the number of terms in the distribution. This is equivalent to saying that the variance is the mean of square of the difference of each term from the mean: $E[(\phi - \mu)^2]$

- The standard deviation (a.k.a $\delta\phi = \sigma = SD = \Delta =$ uncertainty) is the square root of the variance and represents the spread of the terms.

Modifying expressions to account for continuity

In the case of continuous distributions, the sums become integrals and the bounds change appropriately. However, we need to get rid of the denominator since it is a function of n and in a continuous distribution, we don't have n. Let us introduce the probability function $p(\phi)$ such that:

$$< \phi >= \frac{\sum_{i=1}^{n} \phi_i}{n} = \sum_{i=1}^{n} \frac{\phi_i}{n} = \sum_{i=1}^{n} \phi_i p(\phi_i) \tag{2.20}$$

$p(\phi_i)$ as we saw earlier, is the probability of measuring the particular outcome ϕ_i among n possible ones. This way, we got rid of the n in the denominator. As of now, the discrete formula tells us to take a weighted sum of the values ϕ_i, where the weights are the probabilities $p(\phi_i)$. Let $E(\phi)$ be the expectation value or the average of ϕ. Hence, let $E[\phi]=\mu = < \phi >$

In the continuous case, (2.20) now becomes:

$$< \phi >= \sum_{i=1}^{n} \phi_i p(\phi_i) = \int_a^b \phi p(\phi) d\phi \tag{2.21}$$

Similarly, for the variance:

$$(\delta\phi)^2 = \frac{\sum_{i=1}^{n} (\phi_i - < \phi >)^2}{n} = \sum_{i=1}^{n} \frac{(\phi_i - < \phi >)^2}{n}$$

$$= \sum_{i=1}^{n} (\phi_i - < \phi >)^2 p(\phi_i) = \int_a^b (\phi - < \phi >)^2 p(\phi) d\phi = E[(\phi - \mu)^2$$

Note that the denominators are constants and can be pulled in and out of the summation. Also note that $\int_a^b (\phi - <\phi>)^2 p(\phi) d\phi = \int_a^b (\phi - \mu)^2 p(\phi) d\phi$, is just the average, or the expectation value of $(\phi - \mu)^2$. Hence, it makes sense that $(\delta\phi)^2 = E[(\phi - \mu)^2]$.

Now, with the following manipulations,

$$
\begin{aligned}
(\delta\phi)^2 &= E[(\phi-\mu)^2] \\
&= E[\phi^2 + \mu^2 - 2\phi\mu] \\
&= E[\phi^2] + E[\mu^2] - E[2\phi\mu] && \text{(Distributivity can be inferred from integral expression)} \\
&= E[\phi^2] + \mu^2 - 2\mu E[\phi] && \text{(μ is constant, we can take it out from the expected value)} \\
&= E[\phi^2] + \mu^2 - 2\mu^2 && (E(\phi) = \mu) \\
&= E[\phi^2] - \mu^2 && (2.22)
\end{aligned}
$$

we arrive at this formula for variance:

$$(\delta\phi)^2 = E[\phi^2] - <\phi>^2 = \left(\int_a^b \phi^2 p(\phi) d\phi \right) - <\phi>^2 \tag{2.23}$$

The variance can also be remembered as "mean of squares"-"square of means".
(MOSSOM)

$$(\delta\phi)^2 = <\phi^2> - <\phi>^2$$

	Discrete Random Variable	Continuous Random Variable
Mean (Expected Value)	$\mu = E(X) = \sum_{i=1}^{n} x f(x)$	$\mu = E(X) = \int_{-\infty}^{\infty} x f(x) dx$
Variance	$\sigma^2 = V(X) = \sum_{i=1}^{n} (x-\mu)^2 f(x)$	$\sigma^2 = V(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx$
Standard Deviation (Standard Error)	$\sigma = \sqrt{\sigma^2}$	$\sigma = \sqrt{\sigma^2}$

Figure 2.5: Discrete and continuous random variable formulas

Now that we are familiar with different expressions for the variance, mean and standard deviation, let's take a look at the what values they take for the bounds a and b for a rectangular probability distribution.

$$1 = \int_a^b p(\phi)d\phi = p(\phi)\phi\Big|_a^b = p(\phi)(b-a) \implies \boxed{p(\phi) = \frac{1}{b-a}} \tag{2.24}$$

$$< \phi > = \int_a^b \phi p(\phi)d\phi = p(\phi)\int_a^b \phi d\phi = \frac{1}{b-a}\frac{\phi^2}{2}\Big|_a^b = \frac{1}{b-a}\frac{b^2-a^2}{2} = \frac{a+b}{2}$$

$$\implies \boxed{< \phi >= \frac{a+b}{2}} \tag{2.25}$$

$$(\delta\phi)^2 = \int_a^b (\phi - \mu^2) = \int_a^b \phi^2 p(\phi)d\phi - \mu^2$$

$$= \frac{1}{b-a}\int_a^b \phi^2 d\phi - \left(\frac{a+b}{2}\right)^2 = \frac{1}{b-a}\frac{\phi^3}{3}\Big|_a^b - (\frac{a+b}{2})^2 = \frac{1}{b-a}\left[\frac{b^3-a^3}{3} - \frac{(a+b)^2}{4}\right]$$

$$= \frac{(b-a)^2}{12} \implies \boxed{(\delta\phi)^2 = \frac{(b-a)^2}{12}} \tag{2.26}$$

(2.24) comes from normalization and $p(\phi)$ can be removed from the integrals since it is constant.

For a measurement of ϕ with a total uncertainty of Δ, and with the rectangular distribution(2.1.8), the prior probability, the mean, variance and the standard deviation take the following values:

$$p(\phi) = \frac{1}{(\Delta/2) - (-\Delta/2)} = \frac{1}{\Delta} \tag{2.27}$$

$$< \phi > = \frac{a+b}{2} = \frac{(\phi - \Delta/2) + (\phi + \Delta/2)}{2} = \phi_{target} \tag{2.28}$$

$$(\delta\phi)^2 = \frac{(b-a)^2}{12} = \frac{((\phi + \Delta/2) - (\phi - \Delta/2))^2}{12} = \frac{\Delta^2}{12} \tag{2.29}$$

$$(\delta\phi) = \frac{b-a}{2\sqrt{3}} = \frac{(\phi + \Delta/2) - (\phi - \Delta/2)}{2\sqrt{3}} = \frac{\Delta}{2\sqrt{3}} \tag{2.30}$$

2.2 Introduction to Classical Estimation - FI Approach

2.2.1 Estimation Problem

Assume that we have the N-point data set (N samples of measurement outcomes of x):

$$x[n] = x[0], x[1], x[2], ..., x[N-1] \tag{2.31}$$

which depend on an unknown but deterministic parameter θ. Our job is to find estimator functions that map the data into estimates:

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

Since the data in the data set are inherently RVs, we can describe x[n] by its PDF $p(x[0], x[1], ..., x[N-1]; \theta)$. The PDF is parameterized by the unknown parameter θ, in other words, we have a class of PDFs where each one is different due to a different value of θ

For example, we have a PDF $p(x[0]; \theta_1)$, which is different from $p(x[0]; \theta_2)$, and so on.

2.2.2 Classical vs Bayesian Estimation Approaches

We will be discussing two estimation approaches:

1. Classical Estimation $\longrightarrow$ When the unknown parameter θ that we are trying to estimate is a deterministic quantity. There is no way to gather a priori information about θ.
2. Bayesian Estimation $\longrightarrow$ When the unknown parameter θ that we are trying to estimate is viewed as a realization of the random variable θ. There exists a way to gather some prior knowledge about θ. Since θ is assumed to be a random variable, we can assign a PDF to it. This prior PDF influences our choice of estimator.

2.2.3 Assessing Estimator Performance

Example: Consider the following DC Noise measurements:

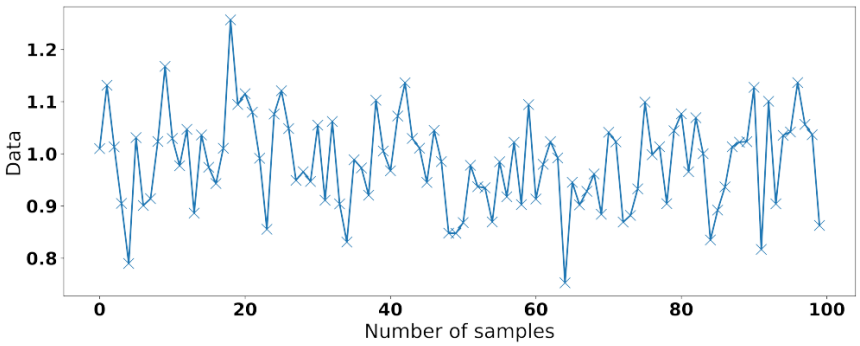


Figure 2.6: Include caption

$$x[n] = \theta + w[n], \ n = 0, ..., N-1, \tag{2.32}$$

Where,

- $x[n]$ is the measured data (which is an RV)
- θ is the constant DC signal, and is deterministic but unknown
- $w[n]$ is a zero mean noise process (which is an RV)

We would like to estimate the constant DC signal. The following are potential estimators for θ:

- $\hat{\theta}_1 = x[0]$
- $\hat{\theta}_2 = \frac{1}{N}\sum_{n=0}^{N-1} x[n]$
- $\hat{\theta}_3 = \frac{a}{N}\sum_{n=0}^{N-1} x[n]$, for some constant a

We would like to determine which estimator is an optimal estimator. In order to do that, we could compare their PDFs, but from (2.1.8), we saw that we could compare their mean (and check for bias) and their variances instead.

Unbiased estimator: An unbiased estimator on average yields the true parameter value:

$$E(\hat{\theta}) = \theta \text{ or } bias(\theta) = E(\hat{\theta}) - \theta = 0 \tag{2.33}$$

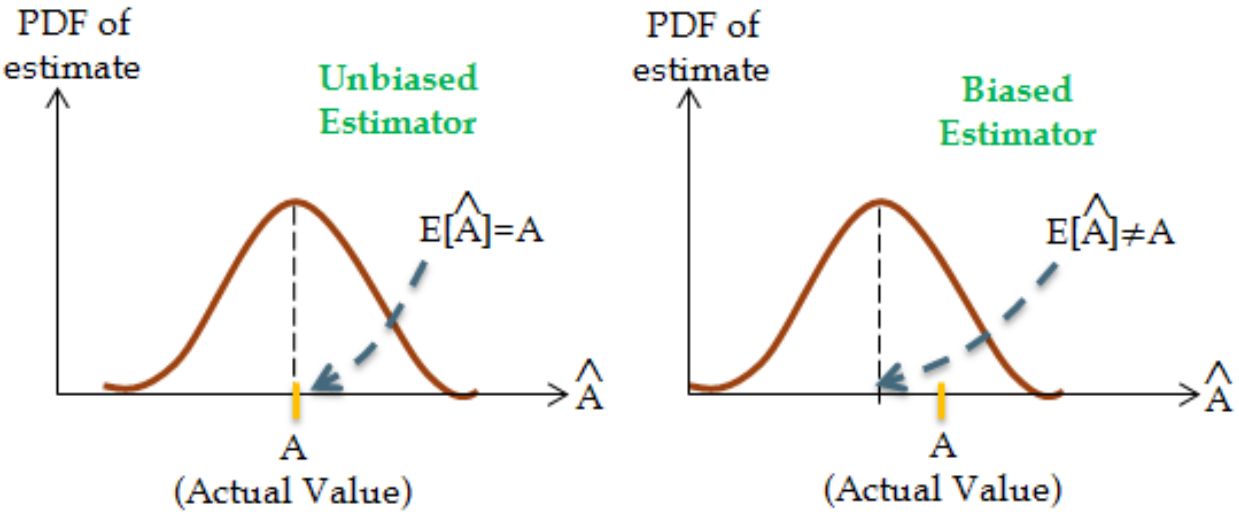


Figure 2.7: The difference between an unbiased estimator (left) and an biased estimator (right)

We compare the bias of our potential estimators of θ:

- $E(\hat{\theta_1}) = E(x[0]) = \theta$
- $E(\hat{\theta_2}) = E(\frac{1}{N}\sum_{n=0}^{N-1} x[n]) = \frac{1}{N}\sum_{n=0}^{N-1} E(x[n]) = \theta$
- $E(\hat{\theta_3}) = E(\frac{a}{N}\sum_{n=0}^{N-1} x[n]) = \frac{a}{N}\sum_{n=0}^{N-1} E(x[n]) = a\theta$

We see that $\hat{\theta_1}, \hat{\theta_2}$ are unbiased estimators while $\hat{\theta_3}$ is biased. Note that an unbiased estimator does not mean that it is optimal. We then compare the variances of our potential estimators of θ:

- $var(\hat{\theta_1}) = \sigma^2$

- $var(\hat{\theta_2}) = var(\frac{1}{N}\sum_{n=0}^{N-1} x[n]) = \frac{1}{N}\sum_{n=0}^{N-1} var(x[n]) = \frac{\sigma^2}{N}$
- $var(\hat{\theta_3}) = var(\frac{a}{N}\sum_{n=0}^{N-1} x[n]) = \frac{a^2}{N}\sum_{n=0}^{N-1} var(x[n]) = \frac{a^2\sigma^2}{N}$

We make the following observations:

1. As $N \to \infty$, $var(\hat{\theta_2} \to 0)$, $var(\hat{\theta_3} \to 0)$
2. $var(\hat{\theta_3})$ is a function of a constant a
3. $var(\hat{\theta_2}) < var(\hat{\theta_1})$ and $\hat{\theta_2}$ is unbiased

Can we then say that $\hat{\theta_2}$ is optimal? Not yet. There are several criteria for optimality. The most logical one is that the estimator should minimize the MSE.

Optimality Criterion

$$
\begin{aligned}
mse(\hat{\theta}) &= E[(\hat{\theta} - \theta)^2] \\
&= E\left[(\hat{\theta} - E[\hat{\theta}] + E[\hat{\theta}] - \theta)^2\right] && \text{Adding and subtracting } E[\hat{\theta}] \\
&= E\left[(\hat{\theta} - E[\hat{\theta}] + b(\theta))^2\right] && \text{bias, } b(\theta) = E[\hat{\theta}] - \theta \\
&= E\left[(\hat{\theta} - E[\hat{\theta}])^2 + 2b(\theta)(\hat{\theta} - E[\hat{\theta}]) + b^2(\theta)\right] \\
&= E\left[(\hat{\theta} - E[\hat{\theta}])^2\right] + 2b(\theta)(E\left[\hat{\theta} - E[\hat{\theta}]\right]) + E\left[b^2(\theta)\right] && \text{The expected value is distributive} \\
&= var(\hat{\theta}) + b^2(\theta) && \text{The second term is 0}
\end{aligned}
$$

We see that the MSE usually depends on θ and that is a problem, unless the estimator is unbiased.

2.2.4 Minimum Variance Unbiased Estimator (MVUE)

We saw earlier that

$$mse(\hat{\theta}) = var(\hat{\theta}) + b^2(\theta) \tag{2.34}$$

We constrain the bias to 0 and find the estimator that minimizes the variance. Out of all unbiased estimates, this amounts to finding and selecting the one with the lowest variance. Usually this is the sample mean. We call this estimator, the **Minimum Variance Unbiased Estimator (MVUE)**.
Can we then infer that the MVUE is the same as an unbiased MMSE.
Sometimes there is no MVUE, this can happen in 2 ways:

1. There may be no unbiased estimators
2. None of the unbiased estimators has a uniformly minimum variance over all θ.

Example: Assume that there are only 3 unbiased estimators for a given problem, and we have 2 possible cases:

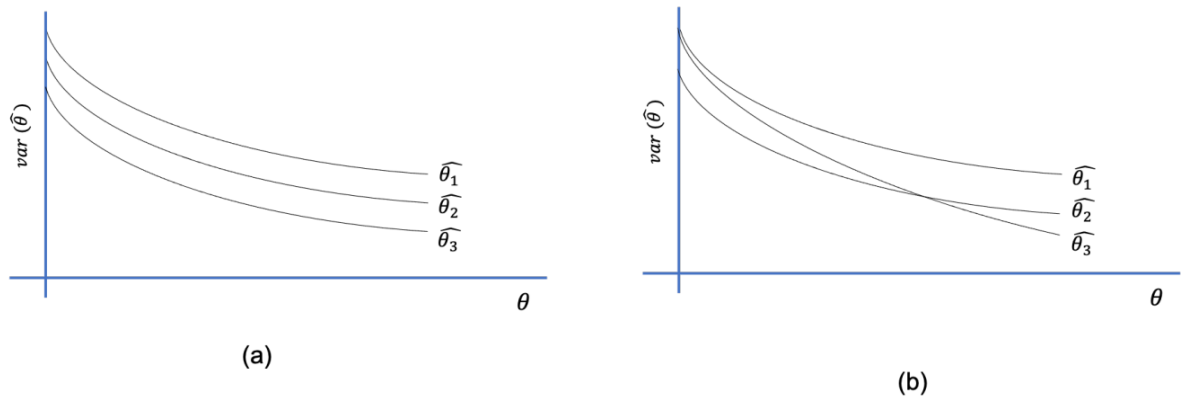


Figure 2.8: a) $\hat{\theta_3}$ is the MVUE b)MVUE does not exist

2.2.5 Cramer Rao Lower Bound (CRB or CRLB)

CRB is a lower bound on the variance of any unbiased estimator. If an estimator attains the bound for all values of the unknown parameter, then CRB allows us to assert that this estimator is the MVUE. At worst, it provides a benchmark against which we can compare the performance of any unbiased estimator. It also gives us an impossibility condition of finding an unbiased estimator whose variance is less than the bound.

If $\hat{\theta}$ is an unbiased estimator of θ, then

$$\sigma^2_{\hat{\theta}}(\theta) \geq CRB_{\hat{\theta}}(\theta) \tag{2.35}$$

Likelihood and Probability

Not clear Recall that RVs x are samples from a random process that depends on θ. The PDF that describes this dependence is given by: $L(x;\theta)$, where L is a function parametrized by θ with the data sample x fixed. Broadly speaking, when a PDF is viewed as a function of the unknown parameter, it is called the likelihood function

- The probability of obtaining outcome θ (given outcome x) $\longrightarrow P(\theta;x)$
- The likelihood of obtaining outcome θ (given outcome x) $\longrightarrow L(x;\theta)$

Clearly, $L(x;\theta) = P(\theta;x)$, and $L(\theta;x) = P(x;\theta)$

Fisher Information

Let's define the Log-Likelihood function as:

$$ln\ [P(x;\theta)] \tag{2.36}$$

Intuitively, the "sharpness" of the likelihood function determines how accurately we can estimate the unknown parameter. Looking at (2.9), we can see that the red line is much "sharper" than the

blue one. Clearly, the likelihood of measuring x and finding that x=3 is much greater on the red curve.

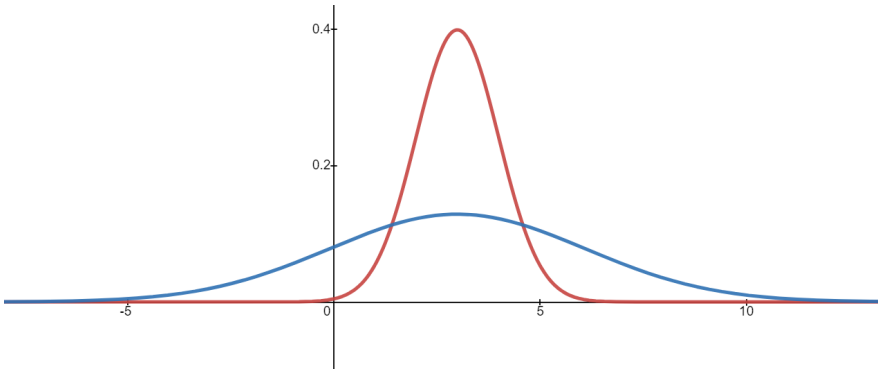


Figure 2.9: Comparing two normal likelihood functions

Mathematically, we can measure the sharpness of the likelihood function by measuring the curvature of the log-likelihood function. Since log of 0 is undefined, taking the log-likelihood of a PDF effectively removes the part of the PDF that goes to zero at the edges, it only keeps the general bell shape of a normal distribution. Therefore the "sharpness" is given by the following expression:

$$\frac{-\partial^2 \, ln \, p(x;\theta)}{\partial \theta^2} \tag{2.37}$$

But this is for a particular data sample x. If we want a general sharpness, we average over all data samples to give us the average curvature:

$$FI = I(\theta) = -E\left[\frac{\partial^2 \, ln \, p(x;\theta)}{\partial \theta^2}\right] \tag{2.38}$$

(2.38) is also called the Fisher Information $I(\theta)$. Clearly FI has the following properties:

1. $I(\theta) \geq 0$
2. $I(\theta)$ is additive for independent observations

Cramer-Rao Lower Bound

Assume that the PDF $p(x;\theta)$ satisfies the "regularity" condition:

$$E\left[\frac{\partial \, ln \, p(x;\theta)}{\partial \theta}\right] = 0 \;\; \forall \, \theta$$

Then the variance of any unbiased estimator $\hat{\theta}$ must satisfy:

$$var(\hat{\theta}) \geq \frac{1}{-E\left[\frac{\partial^2 \, ln \, p(x;\theta)}{\partial \theta^2}\right]} \tag{2.39}$$

where the derivative is evaluated at the true value of θ and the expectation is taken w.r.t $p(x|\theta)$. Furthermore, an unbiased estimator may be found that the bound for all θ iff

$$\frac{\partial \, ln \, p(x;\theta)}{\partial \theta} = I(\theta)(g(x) - \theta)$$

for some functions g and I. That estimator, which is the MVUE, is then $\hat{\theta} = g(x)$ and the minimum variance is $\frac{1}{I(\theta)}$

2.2.6 Maximum Likelihood Estimator (MLE)

Often, MVUEs do not exist or cannot be found. However, difficulties in saturating the CRB are only present in the finite-N regime. In the asymptotic limit of multiple shots/ large sample size (large number of photons). A particular estimator called the Maximum Likelihood (ML) estimator saturates the CRB.

Definition of MLE

The MLE for a scalar parameter is defined to be the value of θ that maximizes $p(x;\theta)$ for a specific value of x; i.e, the value that maximizes the likelihood function. The maximization is performed over the allowable range for θ.

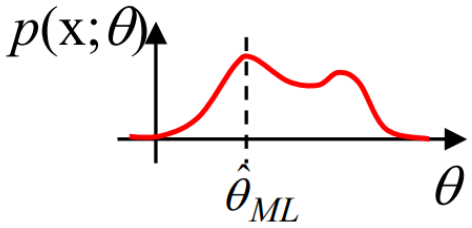


Figure 2.10: $\hat{\theta_{ML}}$ maximizes the likelihood function

Because $ln(z)$ is a monotonically increasing function, $\hat{\theta_{ML}}$ maximizes the log-likelihood function $ln[p(x;\theta)]$.

Example: For the DC noise measurement in (2.32), let's assume that the w[n] term represents a white gaussian noise (WGN). WGN is chosen to be a zero mean noise process.

$$x[n] = \theta + w[n]$$

The likelihood function $p(x;\theta)$ is the PDF of x parametrized by θ:

$$p(x;\theta) = \frac{1}{(2\pi\theta)^{N/2}} e^{\frac{-1}{2\theta}\sum_{n=0}^{N-1}(x[n]-\theta)^2}$$

We then find the estimator that maximizes the log-likelihood function. We take $\frac{\partial ln(p(x;\theta))}{\partial\theta}$, set it to 0, and solve for $\hat{\theta}_{ML}$.

$$\hat{\theta}_{ML} = -\frac{1}{2} + \sqrt{\frac{1}{N}\sum_{n=0}^{N-1} x^2[n] + \frac{1}{4}}$$

It can be shown that this esimator is biased: $E(\hat{\theta}) \neq \theta$. But it is asymptotically unbiased. By using the law of large numbers, the sample mean $\rightarrow$ true mean:

$$\sum_{n=0}^{N-1} x^2[n] \underset{as\ N\to\infty}{\longrightarrow} E[x^2[n]]$$

where $E[x^2[n]] = \theta^2 + \theta$, hence $E[\theta] = \theta$ and therefore the $CRB = \frac{1}{I(\theta)}$

$$CRB = \frac{A^2}{N(A+1/2)}$$

The asymptotic properties are captured in the following theorem:

Theorem: Asymptotic properties of the MLE
If the PDF $p(x;\theta)$ of the data x satisfies some "regularity" conditions, then the MLE of the unknown parameter θ is asymptotically distributed according to (Notation defined here (2.1.3)):

$$\hat{\theta} \overset{a}{\sim} N(\theta, I^{-1}(\theta)) \tag{2.40}$$

This theorem only states what happens asymptotically. When N is small, there is no guarantee how the MLE behaves. To answer how large N must be to achieve asymptotic properties, we use the Monte Carlo Simulation. We can then explore via plots how the Bias and variance vary with, θ and N.

2.3 Introduction to Bayesian Estimation - Bayesian Approach

In the **classical (FI) approach**, we assumed that the parameter θ was deterministic but unknown. This had a few ramifications

- Variance of the estimate could depend on θ.
- In Monte Carlo simulations:
  - M runs done at the same θ
  - must do M runs at each θ of interest
  - averaging done over data
  - averaging over θ values

In the **Bayesian approach**, we assume θ is random with prior PDF $p(\theta)$. In other word, θ is an RV and the statistics of θ is known. This has a few ramifications:

- Variance of the estimate cannot depend on θ ($E(\cdot)$ *w.r.t* $p(x, \theta)$ which is a joint PDF)
- In Monte Carlo simulations:
  - each run done at a **randomly chosen** θ
  - averaging done over data **and** over θ value.

Classical vs Bayesian MSE

There are several different optimization criteria within the Bayesian framework. The most widely used is the MMSE of the BMSE. The MMSE is an estimator that minimizes the BMSE

$$Bmse(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] \tag{2.41}$$

However, both x and θ are random, and the statistic of $\hat{\theta}$ depend on the statistics of both x and θ. Note the difference between the two approaches:

$$Bmse(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] = \int\int [\theta - \hat{\theta}(x)]^2 \, p(x, \theta) \, dxd\theta \tag{2.42}$$

$$mse(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] = \int\int [\theta - \hat{\theta}(x)]^2 \, p(x; \theta) \, dxd\theta. \tag{2.43}$$

We know from (2.9) or (2.10) that,

$$p(x, \theta) = p(x \text{ and } \theta) = p(\theta|x)p(x) \tag{2.44}$$

In the Bayesian case, the BMSE cannot depend on θ. $E[\cdot]$ is taken w.r.t **joint PDF** of x and θ, and the PDF $p(\theta|x)$ is a **conditional PDF**, hence the "|" separator. Whereas, in the classical case, the MSE can depend on θ, and the PDF is of x **parameterized by** θ, hence the separator ";".

MMSE

Now that we found our BMSE, we want to find the estimator function that minimizes the BMSE (2.42):

$$Bmse(\hat{\theta}) = \int \int [\theta - \hat{\theta}]^2 \, p(x, \theta) \, dx d\theta.$$

Hence we have:

$$Bmse(\hat{\theta}) = \int \Big[\int [\theta - \hat{\theta}]^2 \, p(\theta|x) \, d\theta \Big] p(x) \, dx \tag{2.45}$$

Since $p(x) \geq 0$, we minimize the term in the bracket for **each** x value. So, we fix x, take its partial derivative w.r.t $\hat{\theta}$ and set it to 0.

$$\begin{aligned}
&= \frac{\partial}{\partial \hat{\theta}} \int [\theta - \hat{\theta}]^2 \, p(\theta|x) \, d\theta \\
&= \int \frac{\partial}{\partial \hat{\theta}} [\theta - \hat{\theta}]^2 \, p(\theta|x) \, d\theta \\
&= \int 2[\theta - \hat{\theta}] \, p(\theta|x) \, d\theta \\
&= 2 \int \theta p(\theta|x) \, d\theta - 2\hat{\theta} \int p(\theta|x) \, d\theta = 0
\end{aligned}$$

Now, $\int p(\theta|x) \, d\theta = 1$ over the entire interval. Therefore:

$$\hat{\theta} = \int \theta p(\theta|x) \, d\theta = E[\theta|x] \tag{2.46}$$

What we then see here is that:

Bayesian MMSE = Mean of the posterior PDF.
$$\hat{\theta} = E[\theta|x] \tag{2.47}$$

Given the prior PDF, we can use Bayes' Theorem to give us the posterior PDF. We discuss Bayes' Theorem for multiple measurement strategies in the next sections. For each of the measurement strategies, we derive the posterior PDF. Once we have the posterior PDF, we find its mean which is also the Bayesian MMSE (2.47), and we are set with the "best" estimator.

2.3.1 Bayes' Theorem [12]

From (2.9) and (2.10), we see that:

$$P(A) \times P(B|A) = P(A \text{ and } B) = P(A|B)P(B) \tag{2.48}$$

Provided that $P(B) \neq 0$, dividing by P(B) leads us to Bayes Theorem:

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} \tag{2.49}$$

where A and B are events/outcomes and P(B) $\neq 0$.

- $P(A|B)$ is a conditional probability: the probability of event A occurring given that B is true. It is also called the **posterior probability** of A given B
- $P(B|A)$ is also a conditional probability of event B occurring given that A is true. It can also be interpreted as the **likelihood** of A given a fixed B because $P(B|A) = L(A|B)$
- $P(A)$ is the probability of observing A without any given conditions; it is known as the marginal probability or **prior probability**
- $P(B)$ is called the **evidence**.
- A and B must be different events

Bayes' Theorem is a relationship between the likelihood and probability. $P(A|B)$ can be interpreted in two ways:

- The probability of obtaining outcome A (given outcome B) $P(A|B)$
- The likelihood of obtaining outcome B (given outcome A) $L(B|A)$

It may be helpful to think of Bayes' theorem using the diagram below:

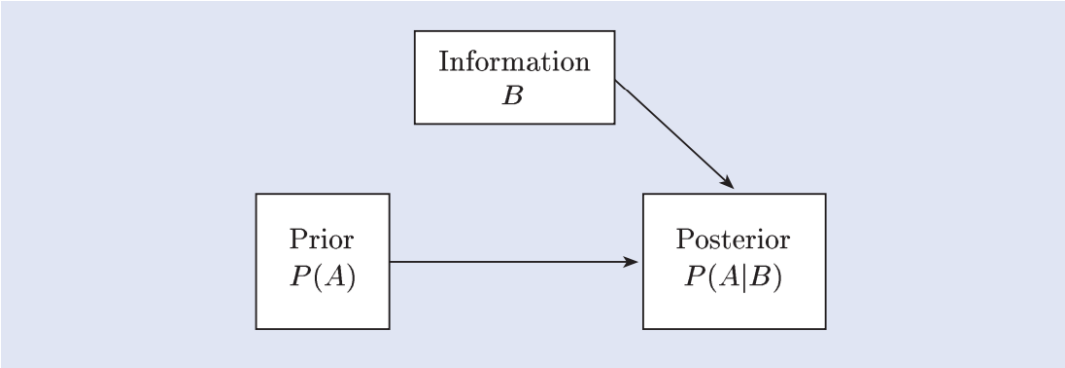


Figure 2.11: Bayes' theorem in practice [13]

Bayes' Theorem for continuous probability distributions

If the probability were a density function, then Bayes theorem would take the following form:

$$P(A|B) = \frac{P(B|A) \times P(A)}{\int_{-\infty}^{\infty} P(B|A)P(A)\,dA} \tag{2.50}$$

This can be derived in the following way. Rearranging Bayes theorem (2.49), and integrating both sides from $-\infty$ to $+\infty$, we have:

$$\int_{-\infty}^{\infty} P(A|B)P(B)\,dA \;= \int_{-\infty}^{\infty} P(B|A)P(A)\,dA \tag{2.51}$$

Now, $P(A|B)$ is a p.d.f on A (2.1.2), and P(B) is a constant w.r.t A. Therefore $\int_{-\infty}^{\infty} P(A|B)\,dA \;= 1$. We then have:

$$\int_{-\infty}^{\infty} P(A|B)P(B)\,dA \;= P(B)\int_{-\infty}^{\infty} P(A|B)\,dA \;= P(B)$$

Equation (2.51) becomes:

$$P(B) = \int_{-\infty}^{\infty} P(B|A)P(A)\,dA \tag{2.52}$$

Bayes' Theorem for multiple measurements

Bayes' theorem can be extended to multiple measurements $(m_1, m_2, m_3, m_4, ...m_M)$, where M is the total number of measurements. An evidence is gathered every time a measurement is performed. It is important to note that each measurement m is independent as well as conditionally independent. One does not imply the other, we are just going to assume that they are both. For a sequence of 2 independent measurements m_1, m_2, equation (2.49) becomes:

$$P(A|m_1, m_2) = \frac{P(m_1, m_2|A)P(A)}{P(m_1, m_2)}$$

Assuming conditional independence (2.14),

$$P(m_1, m_2|A) = P(m_1|A)P(m_2|A)$$

Therefore, Bayes' Theorem for a sequence of 2 independent measurements m_1, m_2 is:

$$P(A|m_1, m_2) = \frac{P(m_1|A)P(m_2|A)P(A)}{P(m_1, m_2)} \tag{2.53}$$

where, from (2.52)

$$P(m_1, m_2) = \int_{-\infty}^{\infty} P(m_1|A)P(m_2|A)P(A)\,dA \tag{2.54}$$

The above equations can be extended to multiple independent measurements. We continue to assume that they are conditionally independent too

$$P(A|m_1, m_2, m_3, ..., m_M) = \frac{P(A)P(m_1|A)P(m_2|A)P(m_3|A)...P(m_M|A)}{P(m_1, m_2, m_2...m_M)} \tag{2.55}$$

where

$$P(m_1, m_2, m_2...m_M) = \int_{-\infty}^{\infty} P(A)P(m_1|A)P(m_2|A)P(m_3|A)...P(m_M|A)\, dA \tag{2.56}$$

Therefore, Bayes' Theorem for a sequence of M independent measurements $m_1, m_2, m_2...m_M$ is:

$$\therefore P(A|m_1, m_2, m_3, ..., m_M) = \frac{P(A)P(m_1|A)P(m_2|A)P(m_3|A)...P(m_M|A)}{\int_{-\infty}^{\infty} P(A)P(m_1|A)P(m_2|A)P(m_3|A)...P(m_M|A)\, dA} \tag{2.57}$$

Bayes' Sequential updating theorem [13]

Suppose that after observing information B, the posterior probability $P(A|B)$ has been calculated but now some additional information C is available. Bayes' theorem can be used to further revise the probability estimate. In this case, $P(A|B)$ becomes the prior probability, as this is the estimate before observing C. The posterior probability is $P(A|C, B)$. This can be calculated using Bayes' theorem:

$$P(A|C, B) = \frac{P(A|B) \times P(C|A, B)}{P(C|B)} \tag{2.58}$$

Notice that all the probabilities in (2.58) are conditional on B, since information B is now part of the prior information. It may be helpful to think of sequential updating in a Bayesian analysis using the diagram below:

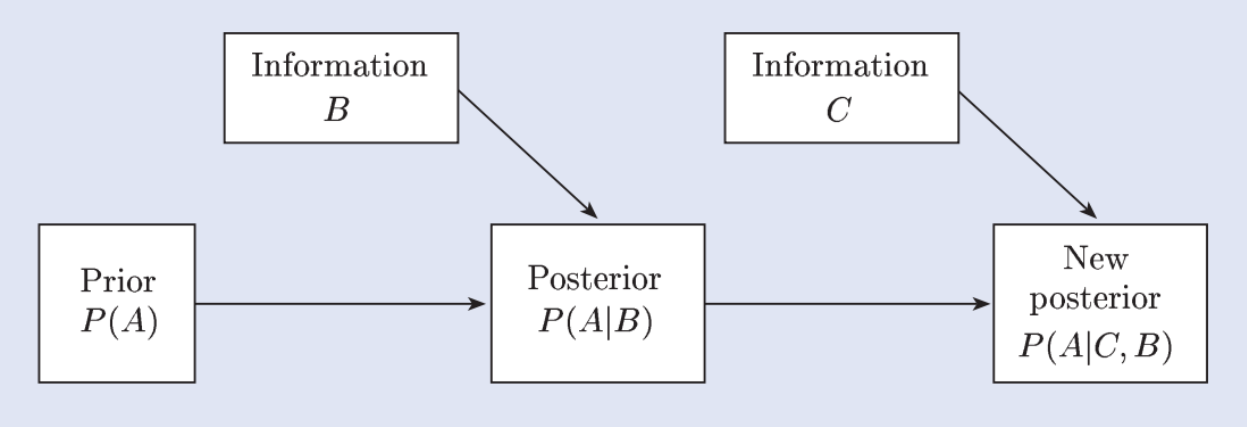


Figure 2.12: Sequential updating [13]

The following is an intuitive derivation to (2.58):

Multiplying both sides of (2.58) by $P(C|B)P(B)$, we get:

$$P(A|C,B)P(C|B)P(B) = P(A|B)P(C|A,B)P(B)$$

Now from (2.7), we know that $P(A|C,B) = \frac{P(A,C,B)}{P(C,B)} = \frac{P(A,B,C)}{P(B)P(C|B)}$

$$\therefore LHS = P(A|C,B)P(C|B)P(B) = P(A,B,C)$$

again, from (2.7), $P(C|A,B) = \frac{P(C,A,B)}{P(A,B)} = \frac{P(A,B,C)}{P(B)P(A|B)}$

$$RHS = P(A|B)P(C|A,B)P(B) = P(A,B,C)$$

Bayes' Sequential updating theorem for multiple updates

2.3.2 General Insights on Bayesian and Classical estimation

Before taking any data, the best estimate of θ of the two approaches are the following

- Classical: No best guess exists
- Bayesian: Prior PDF exists

Prior knowledge leads to a more accurate estimator. To see why this is, assume that the PDF of an RV is a Gaussian.

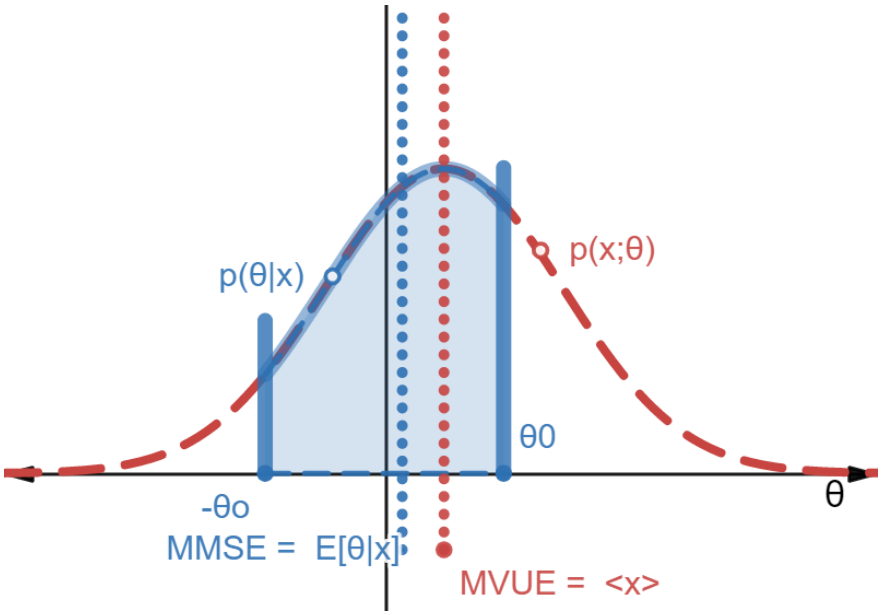


Figure 2.13: Improvement of estimator via prior knowledge

(2.13) If θ could take on any value in the interval $-\infty < \theta < \infty$, then the sample mean (the mean of the red PDF over all space) $< x >$ is the MVUE as we discussed in (2.2.4). However, if we had an "idea" about the parameter θ prior to the measurement x that θ belonged in the interval $-\theta_0 \leq \theta \leq \theta_0$, then we could truncate the PDF at those two limits, since any value outside of this interval would be highly unlikely/impossible(?). This adjusted PDF is shown by the blue line. The mean of this truncated PDF is the MMSE. Note that these two PDFs represent our observed knowledge of a given data sample x (posterior PDF), with (blue) and without (red) the adjustment of a prior estimate.

Intuitively, the effect of observing data will be to concentrate the PDF of θ as shown in figure (2.14). This is because knowledge of the data should reduce our uncertainty about θ. We can clearly see that MMSE $= E[\theta|x] \neq < x >$. This is due to the truncation of the blue posterior PDF, unless θ_0 is so large that there is effectively no truncation. Having θ_0 to be really large would cause (2.14) a) to be extremely flat and long, effectively being the same as having no prior knowledge. In this case, (when θ_0 is reasonable small to make an observable effect) the estimator (MMSE) will be "biased" towards 0 as opposed to being equal to the sample mean $< x >$ because the prior knowledge $p(\theta)$, in the absence of the data x, would force the MMSE to 0: $\hat{\theta} = E(\theta) = 0$

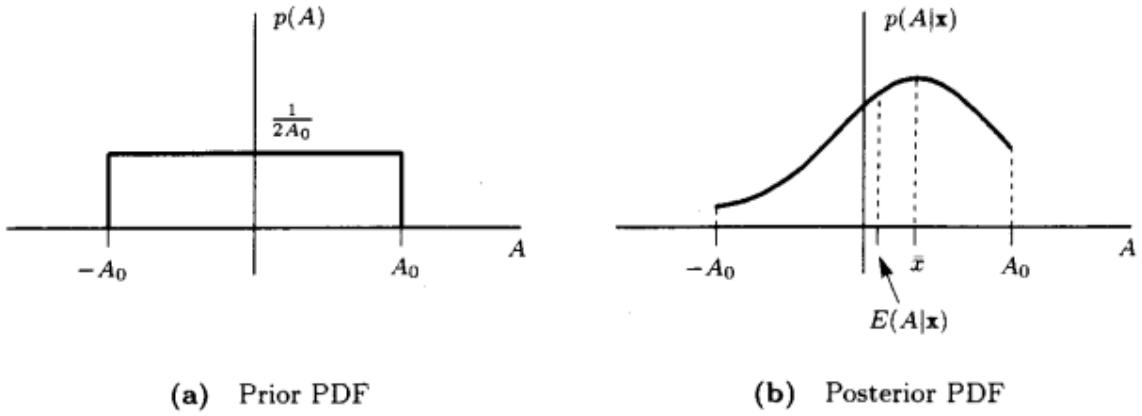


Figure 2.14: Comparison of prior and posterior PDFs

When the number of measurements N, becomes large, the prior data knowledge become very important, and the posterior PDF gets narrower and approaches $< x >= \overline{x}$. The MMSE relies less and less on the prior knowledge and more on the data. The data "swamps out" the prior knowledge. We see that for large number of measurements N, the MMSE becomes "unbiased". In this case, the MMSE becomes an unbiased estimator that minimizes the MSE and therefore attains the CRB on that limit.

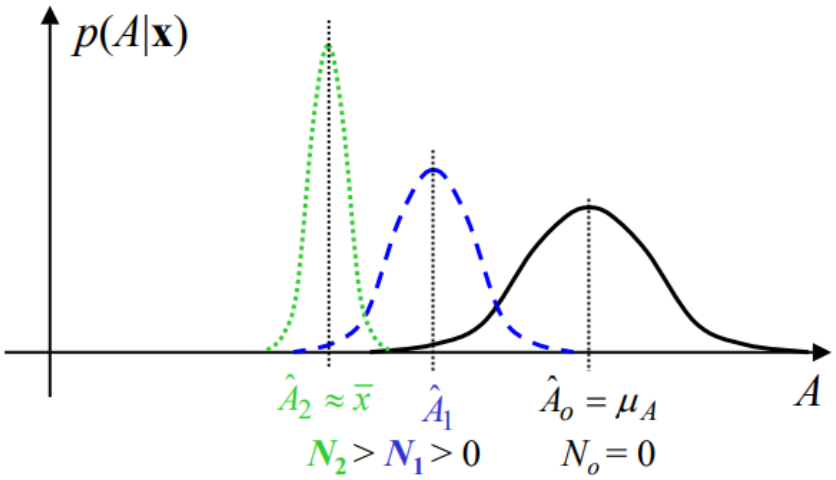


Figure 2.15: Effect of increasing data record on posterior PDF

[20] The MMSE will in general depend on the prior knowledge as well as the data. If the prior knowledge is weak relative to the data, then the estimator will ignore the prior knowledge. Otherwise, the estimator will be "biased" towards the prior mean. As expected, the use of prior information always improves the estimation accuracy.

The choice of a prior PDF is critical in Bayesian estimation. The wrong choice will result in a poor estimator, similar to the problems of a classical estimator designed with an incorrect data model. Much of the controversy surrounding the use of Bayesian estimator stems from the inability

in practice to be able to justify the prior PDF. Suffice it to say that unless the prior PDF can be based on the physical constraints of the problem, then classical estimation is more appropriate.

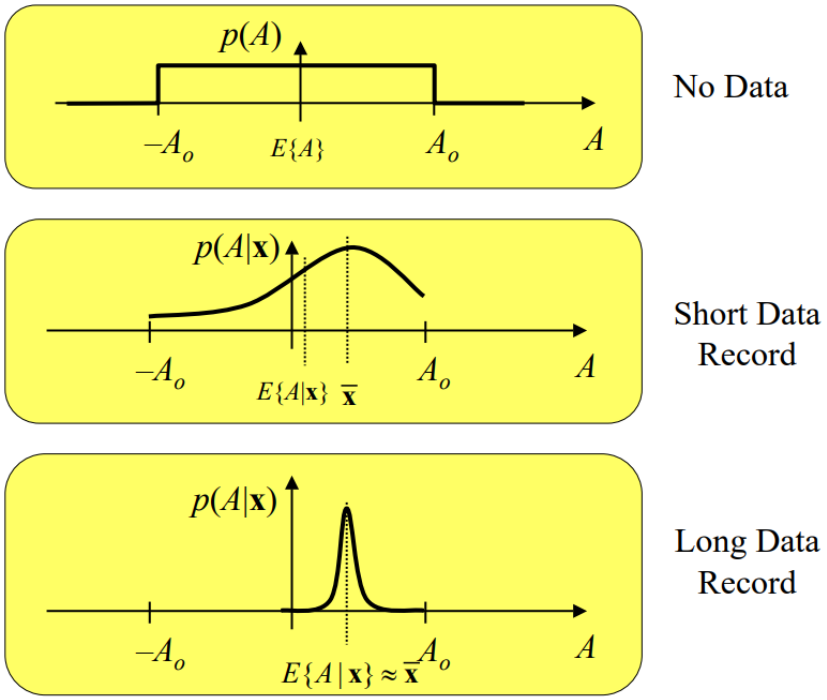


Figure 2.16: How the Bayesian approach balances a priori and a posteriori info

Bibliography

[1] Nielsen, M. A., Chuang, I. L. (2000). Quantum Computation and Quantum Information. Cambridge University Press.

[2] Gerry, C., & Knight, P. (2004). Introductory Quantum Optics. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511791239

[3] Zettili, Nouredine. 2009. Quantum mechanics: concepts and applications. Chichester, U.K.: Wiley.

[4] Chapter Four - Quantum Limits in Optical Interferometry https://reader.elsevier.com/reader/sd/pii/S0079663815000049?token=0054584D677DE24F22914D70AD5F1E81DFB19DC9E645A0C853899DD57C0643089209FAEB5A57747B93E009F61C originRegion=us-east-1&originCreation=20210721151446

[5] Statistical distributions commonly used in measurement uncertainty in laboratory medicine - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6999182/

[6] Taylor J.R. Introduction to error analysis 2ed - https://www.niser.ac.in/sps/sites/default/files/basic_page/John%20R.%20Taylor%20-%20An%20Introduction%20to%20Error%20Analysis_%20The%20Study%20of%20Uncertainties%20in%20Physical%20Measurements-University%20Science%20Books%20(1997).pdf

[7] https://en.wikipedia.org/wiki/Estimator

[8] feedforward.pdf (in Docs)

[9] http://farside.ph.utexas.edu/teaching/qmech/Quantum/node31.html

[10] An Introduction to Bayesian analysis (lecture notes)- Agustín Blasco https://www.mastergr.upv.es/Asignaturas/Apuntes/08.%20Cuantitativa%203/LECTURE%20NOTES.pdf

[11] StatQuest: Probability vs Likelihood https://www.youtube.com/watch?v=pYxNSUDSFH4

[12] https://en.wikipedia.org/wiki/Bayes%27_theorem

[13] Bayesian Statistics- https://www.open.edu/openlearn/ocw/pluginfile.php/1061828/mod_resource/content/3/Bayesian%20statistics%20PDF.pdf

[14] https://en.wikipedia.org/wiki/Probability_density_function

[15] https://eng.libretexts.org/Bookshelves/Industrial_and_Systems_Engineering/Book%3A_Chemical_Process_Dynamics_and_Controls_(Woolf)/13%3A_Statistics_and_Probability_Background/13.04%3A_Bayes_Rule%2C_conditional_probability%2C_independence

[16] https://www.probabilitycourse.com/chapter1/1_4_0_conditional_probability.php

[17] https://en.wikipedia.org/wiki/Second_quantization

[18] https://journals.aps.org/pra/abstract/10.1103/PhysRevA.87.012340

[19] metrologyv2LK2.pdf (check notes)

[20] Kay, S. M. (1993). Fundamentals of statistical signal processing: Estimation theory (S. E. Alan V. Oppenheim, Ed.). Prentice Hall: Upper Saddle River.

[21] http://www.ws.binghamton.edu/fowler/fowler%20personal%20page/ee522.htm https://cas.tudelft.nl/Education/courses/et4386/index.php

[22] Giovannetti, V., Lloyd, S. & Maccone, L. Advances in quantum metrology. Nature Photon 5, 222–229 (2011). https://doi.org/10.1038/nphoton.2011.35 https://arxiv.org/pdf/1102.2318.pdf