# Signal Processing Self Study Notes

Fangyuan Lin

June 11, 2025

# Contents

## 0.1 Introduction

Modern Digital Signal Processing has the following difference from the classical signal processing

- Statistical: signal is random.
- Prerequisite: calculus (working knowledge), linear algebra (working knowledge), probability, signal and system, complex variables.
- Kay: Fundamentals of statistical signal processing.

In this course, we will cover the following content:

1. Linear Estimation: estimation is the center of signal processing and we will focus on linear estimation. The linear case is always the easiest case. We will discuss the fundamentals, including the statistical foundation, orthogonality and orthogonalization, typical processing methods including Wiener and Kalman filters. We

will also discussion extensions to other subjects, e.g. there is NO difference between machine learning and signal processing! We will talk about SVM, kernel, regularization, extensions of the classical processing methods.

2. Adaptive Processing: adaptive filer, LMS, RLS, related to deep learning.
3. Spectral Processing: Direct method (non-parametric), filter bulks.

Signal Processing is a young subject. Probability theory has almost 300 years of history. the material we cover is mainly developed during 1950-1980. The material is old and without research-value; however the concepts and methods are used everywhere in our daily activities. Four main concepts are linearity, orthogonality, stationarity, Gaussianness.
For the next course, we will extend to non-linear, non-orthogonal, non-stationary, and non-Gaussian.

## 0.2 Review of Probability Theory

If you know the rules of the world very well, then there is no uncertainty. Einstein: introducing unvertainty to scientific studies is a compromise to our ignorance. Stanford university did a coin flipping experiment. They made a flipping machine that eliminated all sources of uncertainty, and the result is the same for all the 10000 trials.

- Sample Space: the collection of all possible outcomes. Then we need to assign a measure of uncertainty to each outcome. This assignment is prior. e,g, we model coin flipping as 50-50, not 52-48.
- Note that Probabiity theory and statistics are two completely different subjects.
- Probabiiity theory takes us from model to decision. And statistics takes us from data to a model.
- There has been a method that takes us directly from data to decision - big data (data is too big for us to model).

### 0.2.1 Sample Space

A sample space problem - Bertrand Paradox

Suppose we have a circle and we randomly sample a chord.

- One way to do this is to sample two points on the circle.
  Suppose we construct a inscribed equilateral triangle inside the circle with one of the sampled points being the vertex. The probability that the randomly generated chord is longer than the chord generated by the triangle is $\frac{1}{3}$.
- Another way to characterize a chord is to specify a midpoint of the chord. Now generating a chord is to sample a point over all the points inside the circle. Now we have a different sample space. For the random chord to be longer than the triangle-chord, it must be located in the inscribed circle of the triangle. The area of the small circle is $\frac{1}{4}$ of the big circle.

- This difference is due to the difference in the sample space.
- Note that the random radius sampling method gives a probability of $\frac{1}{2}$.
- Take-away: the probability depends on the way randomness is defined. There is no unique way to pick a random chord.
- The true probability then depends on the experiment setup. The problem is that there is no unique way to define the action "randomly choosing a chord."

### 0.2.2 Random Variables

- The name of random variable is very misleading. What a random variable does is actually quantize the sample point in the sample space.

$$X : \Omega \to \mathbb{R}$$

#### Common Discrete Distributions

Some common discrete distributions

- Bernoulli:
$$X \sim (p, 1-p)$$
- Binomials: we shoot $n$ times and want the probability that $k$ shots are successful.
$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$$
If you go to wallstreet, you must know binomial distribution. Successful shot is exactly the increase in stock price.
- Poisson distribution:
$$P(X=k) = \frac{\lambda^k}{k!}\exp(-\lambda)$$
The Poisson distribution is actually (almost) a special case of binomial distribution.
We let the successful event to be rare in the sense that
$$p \to 0, \quad n \to \infty, \quad np = \lambda$$
Let's do some computation:
$$\begin{aligned}\binom{n}{k} p^k (1-p)^{n-k} &= \frac{n(n-1)\cdots(n-k+1)}{k!} p^k (1-p)^{n-k} \\ &= \binom{n}{k} p^k (1-p)^{n-k} \\ &= \frac{n(n-1)\cdots(n-k+1)}{k!} (\frac{\lambda}{n})^k (1-\frac{\lambda}{n})^{n-k} \\ &= \xrightarrow{n\to\infty} \frac{\lambda^k}{k!}\exp(-\lambda)\end{aligned}$$

### Common Continuous Distributions

Some common continuous distributions

- Uniform:

$$f(x) = \frac{1}{b-a} I_{[a,b]}(x)$$

- Exponential:

$$X \sim \exp(\lambda)$$

$$f(x) = \lambda \exp(-\lambda x) I([0, \infty))(x)$$

  The exponential function has an important property - memoryless property:

$$P(X > x + y | X > x) = P(X > y)$$

  The exponential distribution has been widely used in reliability theory, usually used to model the lifetime of a lightbulb.
  The lighttime of most electronics follows the bathtub-model, where the they are very likely to fail during the first few days. After a while, the lifetime follows an exponential distribution (very reliable).

$$f(X > x + y)/f(X > x) = f(X > y)$$

- Gaussian:

$$X \sim N(\mu, \sigma^2)$$

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

  This reminds of us of the Chinese doctrine of the mean.

## 0.2.3 Expectation

- Definition:

$$\mathbb{E}(X) = \begin{cases} \sum_k x_k P(X = k) \\ \int_{-\infty}^{\infty} x f_x(x) \, \mathrm{d}x \end{cases}$$

  We are trying to describe a random variable using a single number. We inevitably lose a lot of information, but it's easy. A lot of times, expectation is the only information we have.

- Linearity: The most important property of expectation is linearity. We call it magic because we are not imposing any assumption on the random variables.

  A Matching Problem

  There are $N$ people and $N$ hats. They randomly pick a hat. What's the expected number of people with the correct hat?

  - First of all, the underlying distribution $P(X = k)$ is complicated, because the fact whether I pick the correct hat affects other people.

- However, the expected value is easy to find. Everybody knows the definition of the expected value. If we already know the distribution, why the hell would we need the expected value!? We only want the expected value when we don't have the distribution.

-

$$\mathbb{E}(X) = \mathbb{E}(X_1 + \cdots + X_n)$$

where

$$X_k = \begin{cases} 1 & \text{kth matched} \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbb{E}(X_k) = P(X_k = 1) = \frac{1}{n}$$

- Let's consider the sample space,

$$\Omega = \{X_1 \cdots X_n\}$$

where are $|\Omega| = n!$ possible permutations. Among all the permutations, if we exclude one of the

- Therefore, the final answer is

$$n \cdot \frac{1}{n} = 1$$

- Variance: we can then measure the dispersion of the distribution using the following definition of variance:

$$Var(X) = \mathbb{E}(X - \mathbb{E}X)^2$$

- Convexity: let $g(x)$ be convex, i.e. for $\alpha \in (0, 1)$,

$$g(\alpha x + (1-\alpha)y) \leq \alpha g(x) + (1-\alpha)g(x)$$

The function value at a weighted sum is less then the weighted sum of the function values. This brings us to an important inequality - Jenson's inequality:

$$g(\mathbb{E}(X)) \leq \mathbb{E}(g(X))$$

Note that $\mathbb{E}X$ is the weighted sum here.

*Proof.* Here is an important property of convex function $g$: For any $a \in \mathbb{R}$, there exists a slope $L_a \in \mathbb{R}$ such that

$$g(x) \geq g(a) + L_a(x - a)$$

Therefore,

$$\mathbb{E}(g(X)) \geq g(a) + L_a(\mathbb{E}X - a)$$

for any $a$. Let $a = \mathbb{E}X$, then we see that

$$\mathbb{E}(g(X)) \geq g(\mathbb{E}X)$$

□

- Going back to variance,

$$Var(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2 \leq 0$$

by convexity of $g(x) = x^2$.

## 0.3 Estimation

Suppose we want to estimate a random variable $X$ using a constant $a$. We want to quantify how good the estimation is, so we must have a sense of distance (metric).

Mean Square Distance

$$d_{MS}(X, Y) = (\mathbb{E}(X - Y)^2)^{\frac{1}{2}}$$

Now we want to find

$$a^* = argmin_a(\mathbb{E}(X - a)^2)^{\frac{1}{2}} = argmin_a(\mathbb{E}(X - a)^2)$$

since square root is monotonically increasing.
To solve this optimization problem, we then different this expectation:

$$\frac{\mathrm{d}}{\mathrm{d}a}\mathbb{E}(X - a)^2 = 0$$

For us engineering students, let's just swap the order of differentiation and expectation. Of course there are mathematical conditions, but we don't check them. We then get

$$-2\mathbb{E}(X - a) = 0$$

Lastly,

$$a = \mathbb{E}(X)$$

Note that $\mathbb{E}(X - \mathbb{E}(X))^2$ is the variance.
The purpose of signal processing is to do the following optimization problem:

$$\min_g \mathbb{E}(X - g(Y))^2$$

This is a functional optimization problem. We can think of $Y$ as the data we collected and $g(Y)$ is our processed signal.

## 0.4 Conditional Expectation

We briefly review the most important facts about conditional expectation.

1. First of all, conditional expectation is a random variable.
2. Second of all, it retains the most important property of expectation - linearity.
3. Towering property:

$$\mathbb{E}(\mathbb{E}(X|Y)) = \mathbb{E}X$$

*Proof.*

$$\begin{aligned}\mathbb{E}(X|Y) &= \int_{-\infty}^{\infty} x f_{x|Y}(x|Y)\,\mathrm{d}x \\ \mathbb{E}(\mathbb{E}(X|Y)) &= \int_{-\infty}^{\infty} (\int_{-\infty}^{\infty} x f_{x|Y}(x|Y)\,\mathrm{d}x) f_y\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x = \mathbb{E}X\end{aligned}$$

□

4. Take-out property:

$$\mathbb{E}(Xh(Y)|Y) = h(Y)\mathbb{E}(X|Y)$$

We are computing the expectation of $X$ and we think of $Y$ as deterministic inside the conditional expectation so it's natural to take $h(Y)$ out of the conditional expectation.

5. Now what if we want to compute the expectation of a random number of random variables: Suppose all the $X_i$ are identically distributed and $N$ is independent of $X$

$$\mathbb{E}(\sum_{i=1}^{N} X_i) \neq N\mathbb{E}(X_1)$$

They are certainly not equal because expectation gives a number but $N$ is random. Our intuitive guess would be

$$\mathbb{E}(\sum_{i=1}^{N} X_i) = \mathbb{E}(N)\mathbb{E}(X_1)$$

This is indeed correct. We can take advantage of conditional expectation here - by taking care of each of the components inside the expectation one by one.

$$\mathbb{E}(\sum_{k=1}^{N} X_k) = \mathbb{E}(\mathbb{E}(\sum_{k=1}^{N})X_k|N)$$

Now that we have a new tool aka conditional expectation, let's go back to the functional optimization problem:

$$\min_g \mathbb{E}(X - g(Y))^2$$

The problem is hard because we had 2 random variables.

$$\begin{aligned}\mathbb{E}(X - g(Y^2))^2 &= \mathbb{E}_Y(\mathbb{E}_X(X - g(Y)|Y)) \\ &\geq \mathbb{E}_Y(\mathbb{E}_X(X - E(X|Y)|Y)) \\ &= \mathbb{E}(X - \mathbb{E}(X|Y))^2\end{aligned}$$

Therefore, the best estimation is $\mathbb{E}(X|Y)$. (Note that you might be confused how we just got $\mathbb{E}(X|Y)$ - it's based on intuition and in research, most progress is made possible by intuition. Many steps are not rigorous when we first write our draft. We first have insights, then prove.) This is an extremely important fact so let's make everything clear.

$$\begin{aligned}\mathbb{E}(X-g(Y))^2 &= \mathbb{E}(X-\mathbb{E}(X|Y)+\mathbb{E}(X|Y)-g(Y))^2\\ &= \mathbb{E}(X-\mathbb{E}(X|Y))^2+\mathbb{E}(\mathbb{E}(X|Y)-g(Y))^2+2\mathbb{E}((X-\mathbb{E}(X|Y))(\mathbb{E}(X|Y)-g(Y)))\end{aligned}$$

If the cross-term is 0, then we can conclude that $\mathbb{E}(X|Y)$ is the best estimator:

$$\begin{aligned}&\mathbb{E}_Y(\mathbb{E}_X((X-\mathbb{E}(X|Y))(\mathbb{E}(X|Y)-g(Y))|Y))\\ &=\mathbb{E}_Y((\mathbb{E}(X|Y)-g(Y))\mathbb{E}(X-\mathbb{E}(X|Y)|Y)\end{aligned}$$

Now

$$\mathbb{E}(X-\mathbb{E}(X|Y)|Y)=\mathbb{E}(X|Y)-\mathbb{E}(X|Y)\mathbb{E}(1|Y)=0$$

Theoretically, conditional expectation has solved our problem, but in practice, it can be hard to compute.

## 0.5 Basics of Estimation

The goal of estimation is to obtain a model (distribution) from the sample (data). There are two approaches to estimation - parametric and non-parametric (clustering classification in machine learning):

- Parametric: the sample follows a known distribution parametrized by an unknown parameter $\theta$:

  $$\{X_i\}\sim f(X,\theta)$$

  Our aim to find an estimator for $\theta$

  $$\hat{\theta}(X_1,\dots,X_n).$$

  It is called an estimator or simply statistics (doing statistics is making a statistics). *In machine learning, this is called feature and we wanna do feature extraction.* These are synonyms. We are using a frequentist approach for now we consider $\theta$ as a determined but unknown constant. (The other approach is Bayesian approach, which we will talk about later. Actually the Bayesian approach is even more popular now, because of the emergence of machine learning.)
  Our goal is to minimize:

  $$\min_{\hat{\theta}}\mathbb{E}(\hat{\theta}(X_1,\dots,X_n)-\theta)^2$$

  This implies that

  $$\hat{\theta}_{opt}=\mathbb{E}(\theta|X_1,\dots,X_n)=\theta$$

  This formula is useless because $\theta$ is what we are trying to estimate. Let's now look at $\mathbb{E}(\hat{\theta}-\theta)^2$.

  $$\begin{aligned}\mathbb{E}(\hat{\theta}-\theta)^2&=\mathbb{E}(\hat{\theta}-\mathbb{E}(\hat{\theta})+\mathbb{E}(\hat{\theta})-\theta)^2\\ &=\mathbb{E}(\hat{\theta}-\mathbb{E}(\hat{\theta}))^2+\mathbb{E}(\mathbb{E}(\hat{\theta})-\theta)^2+2\mathbb{E}((\hat{\theta}-\mathbb{E}(\hat{\theta}))(\mathbb{E}(\hat{\theta})-\theta))\end{aligned}$$

  The cross term is zero.

### Bias-Variance Trade-Off

$\mathbb{E}(\hat{\theta}-\mathbb{E}(\hat{\theta}))^2$ is the variance. $\mathbb{E}(\mathbb{E}(\hat{\theta})-\theta)^2$ is called the bias. **If we have to consider a trade-off between variance and bias, we don't mind if the bias is big, we want the variance to be small.**

#### A watch example

Imagine you have watch. It's fine if the watch is consistently faster than the standard time by 6 hours. However, if the watch is fast today, slow tomorrow, i.e. behaves randomly, then you can throw it away.

- If $\mathbb{E}(\hat{\theta})=\theta$, then we have the unbiasedness property.

Suppose we have data $X_1,\dots,X_n$, we usually use the sample mean and sample variance

$$\begin{cases}\bar{X}=\frac{1}{n}\sum_{k=1}^{n}X_k\\ \bar{S}=\frac{1}{n-1}\sum_{k=1}^{n}(X_k-X)^2\end{cases}$$

Let's analyze the reasons behind this choice. Here is an example:

### Estimating a constant

Suppose there is an unknown constant $A$ which we are trying to determine. We can observe

$$X=A+N$$

where $N$ is a noise random variable that follows a norman distribution with mean 0 and variance $\sigma^2$.

- First of all, if we only have one data point, then we can only pray to gob that the noise is small. We can let the estimator be

  $$\hat{\theta}_1(X_1,\dots,X_n)=X_1.$$

  and it's actually unbiased.

- If we have two data points $X_1,X_2$, then we can perform

  $$X_1,X_2\to\frac{X_1+X_2}{2}$$

  $$\hat{\theta}_2(X_1,\dots,X_n)=\frac{1}{2}(X_1+X_2)$$

  The goal of processing is to reduce the variance of the samples. We claim that $\hat{\theta}_2$ is a better estimator than $\hat{\theta}_1$.
  First of all,

  $$\mathbb{E}\hat{\theta}_1=\mathbb{E}\hat{\theta}_2=\theta$$

  In terms of expectation, $\hat{\theta}_1$ and $\hat{\theta}_2$ are both *Unbiased.*

$$\begin{aligned}
Var(\hat{\theta}_n) &= \mathbb{E}(\hat{\theta}_n - \mathbb{E}(\hat{\theta}_n))^2 \\
&= \mathbb{E}(\frac{1}{n}\sum_{k=1}^{n} X_k - A)^2 \\
&= \frac{1}{n^2}\mathbb{E}(\sum_{k=1}^{n}(X_k - A))^2 + \frac{1}{n^2}\sum_{i\neq j}\mathbb{E}(X_i - A)(X_j - A) \\
&= \frac{1}{n^2}\mathbb{E}(\sum_{k=1}^{n}(X_k - A))^2 \quad \text{with an uncorrelation assumption!} \\
&= \frac{1}{n}\sigma^2 \to 0 \quad \text{as } n \to \infty
\end{aligned}$$

This is called the consistency property, i.e.

$$\mathbb{E}(\hat{\theta}(X_1, \ldots, X_n) - \theta)^2 \overset{n\to\infty}{\to} 0$$

### 0.5.1 Why n-1 in the sample variance - unbiasedness

By the way, one might be confused that when estimating the variance, we used $n-1$ in $\bar{S} = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - X)^2$ for estimating the variance. If $\bar{X}$ is replaced by the true mean $A$, then we should divide by $n$. But now:

$$\begin{aligned}
\mathbb{E}((n-1)\bar{S}) &= \mathbb{E}(\sum_{k=1}^{n}(X_k-\bar{X})^2) \\
&= \sum_{k=1}^{n}\mathbb{E}(X_k-\bar{X})^2 \\
&= \sum_{k=1}^{n}\mathbb{E}(X_k^2-2X_k\bar{X}+\bar{X}^2) \\
&= \sum_{k=1}^{n}\mathbb{E}(X_k^2)-2(\sum_{k=1}^{n}X_k)\bar{X}+\sum_{k=1}^{n}(\bar{X})^2 \\
&= \mathbb{E}(\sum_{k=1}^{n}X_k^2-2n(\bar{X})^2+n(\bar{X})^2) \\
&= \mathbb{E}(\sum_{k=1}^{n}X_k^2-n(\bar{X})^2) \\
&= \mathbb{E}(\sum_{k=1}^{n}X_k^2-n(\frac{1}{n}\sum_{k=1}^{n}X_k)^2) \\
&= \mathbb{E}(\sum_{k=1}^{n}X_k^2-\frac{1}{n}(\sum_{k=1}^{n}X_k^2+\sum_{i\neq j}X_iX_j)) \\
&= \frac{n-1}{n}\sum_{k=1}^{n}\mathbb{E}(X_k^2)-\frac{1}{n}\sum_{i\neq j}\mathbb{E}(X_iX_j) \\
&= (n-1)\mathbb{E}(X_k^2)-(n-1)(\mathbb{E}X_1)^2=(n-1)Var(X_1)
\end{aligned}$$

In the last step, we assumed the $X_i$ are i.i.d.. Making this assumption is a fundamental key to our progress.
Therefore, the purpose of the "$n-1$" is to get *unbiased* estimator.
There is always a reason behind weird things - it's just that most of times, we don't have the capability to find out, or we don't need to find out. If your girlfriend decides to be with someone else, there must be a reason that you probably are not able to know.
Next, we discuss variance. We can apply the idea of conditioning to variance as well since it's also an expectation.

$$Var(X|Y)=\mathbb{E}((X-\mathbb{E}(X|Y))^2|Y)$$

One might think $Var(X)=Var(Var(X|Y))$. Of course this is wrong, since the dimension of variance is already second-order but the righthand side is a fourth order quantity. In fact,

$$Var(X)=Var(\mathbb{E}(X|Y))+\mathbb{E}(Var(X|Y))$$

*Proof.*

$$\begin{aligned}
Var(X) &= \mathbb{E}(X-\mathbb{E}X)^2\\
&= \mathbb{E}(X+\mathbb{E}(X|Y)-\mathbb{E}(X|Y)-\mathbb{E}X)^2\\
&= \mathbb{E}(X-\mathbb{E}(X|Y))^2-\mathbb{E}(\mathbb{E}(X|Y)-\mathbb{E}X)^2\\
&+ 2\mathbb{E}((X-\mathbb{E}(X|Y))(\mathbb{E}(X|Y)-X)\\
&= \mathbb{E}(X-\mathbb{E}(X|Y))^2-\mathbb{E}(\mathbb{E}(X|Y)-\mathbb{E}X)^2\\
&= \mathbb{E}(\mathbb{E}((X-\mathbb{E}(X|Y))^2|Y))+\mathbb{E}(\mathbb{E}(X|Y)-\mathbb{E}(\mathbb{E}(X|Y))^2\\
&= \mathbb{E}(Var(X|Y))+Var(E(X|Y))
\end{aligned}$$

□

## 0.6 Minimum Mean Square Error Estimation

Below we introduce a notion of distance to measure the error of an estimator.

Mean Squared Error MSE

$$d(\hat{\theta},\theta) = (\mathbb{E}(\hat{\theta}-\theta)^2)^{\frac{1}{2}}$$

When we are minimizing the MSE, we can ignore the square root.

In the setting of

$$\hat{\theta}(X_1,\ldots,X_n)\to\theta,$$

the MMSE estimator is of course $\theta$, but we are looking for $\theta$ and we don't have it. Therefore, we must construct an estimator out of the data we have.

- If $\hat{\theta}(X)\sim f(X)$ is independent of $\theta$, then it's very bad and we call it *ancillary*. Suppose

  $$X_1, X_2 \overset{i.i.d.}{\sim} N(\theta,1)$$

  If we let

  $$\hat{\theta}(X_1,X)2) := X_1 - X_2 \sim N(0,2),$$

  then $\hat{\theta}$ has nothing to do with $\theta$.
- We now would like our statistics to be uniformly optiomal as well as unbiased.

## 0.7 Minimum Variance Unbiased Estimator MVUE

Suppose $\hat{\theta}$ is unbiased, then

$$MSE(\hat{\theta}) = \mathbb{E}(\hat{\theta}-\theta)^2 = \mathbb{E}(\hat{\theta}-\mathbb{E}(\hat{\theta}))^2 = Var(\hat{\theta})$$

Under the unbiasedness assumption, then MSE becomes MVUE.

- We also want to have sufficiency for our estimator (Ronald A. Fisher), i.e. we want the statistics to contain all the information about $\theta$.

Sufficient Statistics

Mathematically, Fisher meant for our statistics $S$ conditioned on $t$,

$$f(X, \theta | S = t)$$

is independent of $\theta$. If this condition is true, then $S$ is called a sufficient statistics.

## 0.8 Sufficient Statistics

Example of sufficient statistics

Suppose we have $X_1, \dots, X_n \overset{i.i.d.}{\sim} Bernoulli(p)$. Then

$$\begin{aligned} f(x_1, \dots, x_n) &= P(X_1 = x_1, \dots, X_n = x_n) \\ &= \prod_{k=1}^{n} P(X_k = x_k) \\ &= p^{\sum_{k=1}^{n} x_k}(1-p)^{n-\sum_{i=1}^{n} x_k} \end{aligned}$$

Suppose our statistics is the sum of all the $x_i$. Then

$$\begin{aligned} & P(X_1 = x_1, \dots, X_n = x_n | \sum_{k=1}^{n} x_k = t) \\ &= P(X_1 = x_1, \dots, X_n = x_n, \sum_{k=1}^{n} x_k = t) / P(\sum_{k=1}^{n} x_k = t) \\ &= \frac{p^t(1-p)^{n-t}}{\binom{n}{t} p^t (1-p)^{n-t}} \\ &= \frac{1}{\binom{n}{t}} \end{aligned}$$

Now we see that our statistics $S$ is sufficient, because the distribution doesn't depend on the parameter $p$ anymore after the conditioning. The information of the number of success contains the information of the success rate $p$.

Now we present a way to characterize a sufficient statistics:

Neyman Factorization

$S$ is a sufficient statistics if and only if

$$f(x;\theta) = h(x)g(\theta, S(x))$$

In the example above,

$$S(x) = \sum_{k=1}^{n} x_k$$

$$f(x,\theta) = p^S(x)(1-p)^{n-S(x)}$$

where $g(\theta, S(x)) = p^S(1-p)^{n-S}$.
Here is another example,

Poisson Sufficient Statistic Example

Let $X_k$ be i.i.d. Poisson distributions of rate $\lambda$. Recall that

$$P(X_k = x_k) = \frac{\lambda^{x_k}}{x_k!} \exp -]\lambda.$$

Then

$$f(x_1,\dots,x_n,\lambda) = \prod_{k=1}^{n} \frac{\lambda^{x_k}}{(x_k)!} \exp -\lambda = \lambda^{\sum x_k} \exp(-n\lambda)$$

It seems that the sum of the $x_i$ is an interesting quantity - we will see this guy again later.

Gaussian Sufficient Statistic Example

Suppose $X_1,\dots,X_n$ are i.i.d. Gaussian random variables with unknown mean $\theta = \mu$, and known variance $\sigma^2$.

$$\begin{aligned} f(X_1,\dots,X_n,\mu) &= (\frac{1}{\sqrt{2\pi\sigma}}) \exp(-\frac{1}{2\sigma^2}\sum_{k=1}^{n}(X_k-\mu)^2) \\ &= (\frac{1}{\sqrt{2\pi\sigma}}) \exp(-\frac{1}{2\sigma^2}\sum X_k^2 - 2(\sum X_k\mu) + n\mu^2) \\ &= (\frac{1}{\sqrt{2\pi\sigma}}) \exp(-\frac{1}{2\sigma^2}\sum X_k^2) \exp(-2(\sum X_k\mu) + n\mu^2) \end{aligned}$$

One sufficient statistics is then $\sum X_k$.
A natural extension: if both $\theta$ and $\sigma^2$ are known, the sufficient statistics would be $\sum X_k$ as well as $\sum X_k^2$. (Fangyuan: to see this, I think we can think of the parameter in the factorization as a tuple.)

### Estimation using one, two, three... samples

Now let's look at the mean squared distance.

$$\begin{aligned} d(\hat{\theta}_1, \theta) &= (\mathbb{E}(\theta_1 - \theta)^2)^{\frac{1}{2}} \\ &= (\mathbb{E}(X_1 - \theta)^2)^{\frac{1}{2}} = 1 \end{aligned}$$

because $X_1 - \theta$ is Gaussian distribution with mean 0 and variance 1.
Now

$$\begin{aligned} (\mathbb{E}(\hat{\theta}_2 - \theta)^2)^{\frac{1}{2}} &= (\mathbb{E}(\frac{1}{2}(X_1 - \theta) + \frac{1}{2}(X_2 - \theta))^2)^{1/2} \\ &= (\frac{1}{4}\mathbb{E}((X_1 - \theta) + (X_2 - \theta))^2)^{1/2} \\ &= \frac{1}{2}(\mathbb{E}(X_1 - \theta)^2 + \mathbb{E}(X_2 - \theta)^2 + 2\mathbb{E}(X_1 - \theta)(X_2 - \theta))^{\frac{1}{2}} \\ &= \frac{1}{\sqrt{2}} < 1 \end{aligned}$$

This shows the important to have independent samples because otherwise, we cannot have

$$\mathbb{E}(X_1 - \theta)(X_2 - \theta) = \mathbb{E}(X_1 - \theta)\mathbb{E}(X_2 - \theta)$$

Unsurprisingly, in general,

$$\hat{\theta}_n = \frac{1}{n}\sum_{k=1}^{n} X_k, \quad (\mathbb{E}(\hat{\theta}_n - \theta)^2)^{\frac{1}{n}} = \frac{1}{\sqrt{n}}$$

To estimate the variance, we let

$$\hat{\sigma^2} = \frac{1}{n}\sum_{k=1}^{n}(X_k - \bar{X})^2$$

Now we show why we have a $\frac{1}{n-1}$ instead of a $1\frac{1}{n}$.

$$\begin{aligned} \sum_{k=1}^{n}(X_k - \bar{X})^2 &= \sum_{k=1}^{n} X_k^2 - 2\sum X_k\bar{X} + n(\bar{X})^2 \\ &= \sum_{k=1}^{n} X_k^2 - 2n(\bar{X})^2 + n(\bar{X})^2 \\ &= \sum_{k} X_k^2 - n(\bar{X})^2 \\ \mathbb{E}(\sum(X_k - \bar{X})^2) &= \mathbb{E}(\sum X_k^2) - n\mathbb{E}(\bar{X})^2 \end{aligned}$$

Note that we don't know $\theta$ so we used $\bar{X}$.

- The second approach is non-parametric approach - clustering classification.

## 0.9 Rao-Blackwell Procedure

Construction of a New Estimator Using Sufficient Statistic

Suppose $\hat{\theta}(X), \mathbb{E}(\hat{\theta}) = \theta$ is an unbiased estimator. For all sufficient statistic $S$, we can construct a new estimator defined as

$$\hat{\theta}' = \mathbb{E}(\hat{\theta}|S) = g(S)$$

with

$$Var(\hat{\theta}') \leq Var(\hat{\theta})$$

*Proof.*

$$\begin{aligned} Var(\hat{\theta}) &= Var(\mathbb{E}(\hat{\theta}|S)) + \mathbb{E}(Var(\hat{\theta}|S)) \\ Var(\hat{\theta}) &= Var(\hat{\theta}') + \mathbb{E}(Var(\hat{\theta}|S)) \\ MSE(\hat{\theta}) &= MSE(\hat{\theta}') + \mathbb{E}(Var(\hat{\theta}|S)) \end{aligned}$$

□

We used the law of total variance in the proof. Note that $\hat{\theta}'$ is unbiased by the towering property of conditional expectation.
I was smart enough to see that sufficiency of $S$ was never used. The importance of sufficiency will be demonstrated later.
Update: The sufficiency of $S$ is needed because we want to make sure that the newly constructed statistic $\hat{\theta}'$ does not depend on $\theta$ because otherwise, it would be cheating.

An Example

Suppose

$$X_1, \ldots, X_n \sim N(\mu, \sigma^2)$$

and $\sigma^2$ is known. Let

$$S = \sum_{k=1}^{n} X_k$$

be the sufficient statistic mentioned previously. Suppose we take the naive estimator

$$\hat{\theta} = X_1$$

and let's see how much Rao-Blackwell can improve it.

$$\hat{\theta}'(X) = \mathbb{E}(X_1|\sum_{k=1}^{n} X_k)$$

By symmetry,

$$\mathbb{E}(X_i|\sum X_k) = \mathbb{E}(X_j|\sum X_k)$$

Therefore,

$$\hat{\theta}' = \frac{1}{n}\mathbb{E}(\sum X_k | X_k) = \frac{1}{n}\sum_k X_k$$

Now, we can look at the reduction in MSE:

$$MSE(\hat{\theta}) = MSE(X_1) = \sigma^2$$

becomes

$$MSE(\hat{\theta}') = MSE(\frac{1}{n}\sum_k X_k) = \frac{\sigma^2}{n}$$

We will see that we cannot make a better improvement than this, so Rao-Blackwell is indeed useful.

Next, we introduce a new concept called completeness to show that there is a limit for which we can reduce the MSE.

## 0.10 Complete Statistics

Complete Statistic

Let $X \sim f(x, \theta)$. If $T = T(X)$ is a complete statistic, then for any $\theta$ and function $h$,

$$\mathbb{E}_\theta(h(T)) = 0 \implies h(T) = 0 \quad a.s.$$

Lehrmann–Scheffe theorem: Sufficiency and Completeness gives Minimum Variance Unbiased Estimator

If $T$ is sufficient and complete, then

$$\mathbb{E}(h(T)) = \theta \implies h(T) \text{ is MVUE}$$

*Proof.* Let $\hat{\theta}$ be any unbiased estimator, let's perform Rao-Blackwell on $\hat{\theta}$, then

$$\hat{\theta}' = \mathbb{E}(\hat{\theta}|T) = \hat{\theta}'(T)$$
$$\implies MSE(\hat{\theta}) \geq MSE(\hat{\theta}'(T))$$

Now note that $\hat{\theta}(T)$ and $h(T)$ both depend on the complete statistic $T$ and they are both unbiased.

$$\mathbb{E}(h(T) - \hat{\theta}'(T)) = \theta - \theta = 0$$

Since the inside of the expectation is just another function of $T$, by completeness of $T$,

$$h(T) = \hat{\theta}'(T) \quad a.s.$$

Therefore, $\hat{\theta}(T)$ always have larger variance than $h(T)$. □

From the above proof, we see that intuitively, completeness means that we only get one possible improvement with Blackwellization.

Let's summarize what we have done so far. We introduced a property called sufficiency and we can improve sufficient statistic through Rao-Blackwellization. The best statistic would be the MVUE. If we have a sufficient and complete statistic, MVUE is easier to find by Lehrmann-Scheffe theorem.

## 0.11 Cramer-Rao Lower Bound CRLB

Like information theory, we want to have bound that tells us the fundamental limits for which we can do certain things.

Cramer-Rao Lower Bound

For any unbiased statistic $\hat{\theta}, \quad \mathbb{E}(\hat{\theta}) = \theta$, then the MSE of the estimator has a lower bound that only depends on the model itself, i.e.

$$\exists A, \quad Var(\hat{\theta}) \geq A(\theta).$$

*Proof.* We know that $\hat{\theta}$ is unbiased, so

$$\theta = \int_{-\infty}^{\infty} \hat{\theta}(x) f(x,\theta)\,\mathrm{d}x$$

We engineering people just ignore the condition for which integration and differentiation can be swapped.

$$\begin{aligned}\mathbb{E}(\hat{\theta}) = \theta &= \int_{-\infty}^{\infty} \hat{\theta}(x) f(x,\theta)\,\mathrm{d}x \\ 1 &= \frac{\mathrm{d}}{\mathrm{d}\theta}\int_{-\infty}^{+\infty} \hat{\theta}(x) f(x,\theta)\,\mathrm{d}x \\ &= \int_{-\infty}^{+\infty} \hat{\theta}(x) \frac{\mathrm{d}}{\mathrm{d}\theta} f(x,\theta)\,\mathrm{d}x\end{aligned}$$

The first time Leibniz worked on complex number, he felt guilty and thought god wouldn't forgive him for taking the square root of negative one; however, we must keep going to be enlightened.
Note also that $f$ is a density function, so

$$\int_{-\infty}^{\infty} f(x,\theta)\,\mathrm{d}x = 1.$$

Then

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\int_{-\infty}^{+\infty} f(x,\theta)\,\mathrm{d}x = \int_{-\infty}^{+\infty} \frac{\mathrm{d}}{\mathrm{d}\theta} f(x,\theta)\,\mathrm{d}x = 0$$

Next, note that

$$\int_{-\infty}^{+\infty} \theta \frac{\mathrm{d}}{\mathrm{d}\theta} f(x,\theta)\,\mathrm{d}x = 0$$

Then we see that

$$\int_{-\infty}^{+\infty} (\hat{\theta} - \theta) \frac{\mathrm{d}}{\mathrm{d}\theta} f(x,\theta)\,\mathrm{d}x = 1$$

We are using the derivative of the density function, but we need the density itself. Fisher applied the following trick.

$$\int_{-\infty}^{+\infty}(\hat{\theta}-\theta)\frac{\mathrm{d}}{\mathrm{d}\theta}f(x,\theta)\,\mathrm{d}x=\int_{-\infty}^{+\infty}(\hat{\theta}-\theta)\frac{\mathrm{d}}{\mathrm{d}\theta}(\log f(x,\theta))f(x,\theta)\,\mathrm{d}x=1$$

Next, we apply the Cauchy-Schwarz inequality:

$$|\int f(x)g(x)\,\mathrm{d}x|^2\leq\int f^2(x)\,\mathrm{d}x\int g^2(x)\,\mathrm{d}x$$

We now split $f$ into $\sqrt{f}\sqrt{f}$. Then we have

$$\begin{aligned}1^2&=\left(\int_{-\infty}^{\infty}(\hat{\theta}-\theta)\sqrt{f(x,\theta)}\frac{\mathrm{d}}{\mathrm{d}\theta}(\log f(x,\theta))\sqrt{f(x,\theta)}\,\mathrm{d}x\right)^2\\&\leq\int_{-\infty}^{\infty}(\hat{\theta}-\theta)^2f(x,\theta)\,\mathrm{d}x\int_{-\infty}^{\infty}\frac{\mathrm{d}}{\mathrm{d}\theta}(\log f(x,\theta))^2f(x,\theta)\,\mathrm{d}x\\&=\mathbb{E}(\hat{\theta}-\theta)^2\mathbb{E}(\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(x,\theta))^2\end{aligned}$$

Therefore

$$Var(\hat{\theta})=\mathbb{E}(\hat{\theta}-\theta)^2\geq 1/I(\theta)$$

where

$$I(\theta)=\mathbb{E}(\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(x,\theta))^2$$

I is called the Fisher Information. □

Proposition

$$I(\theta)=\mathbb{E}(\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(x,\theta))^2=-\mathbb{E}(\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log f(x,\theta))$$

*Proof.* Apply the Fisher trick to $1=\int_{-\infty}^{\infty}f(x,\theta)\,\mathrm{d}x$.

$$\begin{aligned}0&=\int_{-\infty}^{\infty}\frac{\mathrm{d}}{\mathrm{d}\theta}f(x,\theta)\,\mathrm{d}x\\&=\int_{-\infty}^{\infty}(\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(x,\theta))f(x,\theta)\,\mathrm{d}x\\\implies 0&=\int_{-\infty}^{\infty}\frac{\mathrm{d}}{\mathrm{d}\theta}\left((\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(x,\theta))f(x,\theta)\right)\,\mathrm{d}x\\&=\int_{-\infty}^{\infty}(\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log f(x,\theta))f(x,\theta)\,\mathrm{d}x+\int_{-\infty}^{\infty}(\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(x,\theta))\frac{\mathrm{d}}{\mathrm{d}\theta}f(x,\theta)\,\mathrm{d}x\\&=\mathbb{E}(\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log f(x,\theta))+\int_{-\infty}^{\infty}(\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(x,\theta))\frac{\mathrm{d}}{\mathrm{d}\theta}f(x,\theta)\,\mathrm{d}x\end{aligned}$$

For the second term, apply the Fisher's trick again to get □

Note that the second derivative corresponds to the concept of "curvature." The bigger the second derivative, the bigger the curvature (small radius of curvature).

Big curvature means an easier estimation: a change in $\theta$ can be more sensitively reflected in our data.

### Normal Distribution Fisher Information

Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$, $\sigma^2$, find $\mu$ CRLB.

1. First of all, we need to get the model right:

$$f(x,\theta) = (\frac{1}{\sqrt{2\pi}\sigma})^n \exp(-\frac{1}{2\sigma^2}\sum_{k=1}^{n}(x_k-\mu)^2)$$

Then we take the log:

$$\log f(x,\theta) = -n\log(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2}\sum_{k=1}^{n}(x_j-\mu)^2$$

2. Step Two: We now take the derivative with respect to the parameter of interest:

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(x,\theta) = \frac{1}{\sigma^2}\sum_{k=1}^{n}(x_k-\mu)$$

3. Sometimes, it's more convenient to take the second derivative, so let's take a look:

$$\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log f(x,\theta) = -\frac{n}{\sigma^2}$$

4. Step Three: we compute the Fisher information.

$$I(\theta) = -\mathbb{E}(\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log f(x,\theta)) = \frac{n}{\sigma^2}$$

The Cramer-Rao lower bound is then

$$CRLB = \frac{1}{I(\theta)} = \frac{\sigma^2}{n}$$

We saw that Blackwellization directly took us to the MVUE.

### Poisson Cramer-Rao Bound

Suppose $X_1, \ldots, X_n$ are i.i.d. Poisson random variables with mean $\lambda$.

- Step one: the model is

$$f(x,\theta) = \prod_{k=1}^{n}\frac{\lambda^{x_k}}{(x_k)!}\exp(-\lambda) = \lambda^{\sum_{k=1}^{n}x_k}\exp(-n\lambda)(\prod_{k=1}^{n}(x_k)!)^{-1}$$

Then we take the log:

$$\log f(x,\theta) = (\sum_{k=1}^{n} x_k)\log\lambda - n\lambda - \sum_{k=1}^{n}\log(x_k)!$$

- We then take the derivative:

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\log f(x,\theta) = \frac{1}{\lambda}(\sum_{k=1}^{n} x_k) - n$$

$$\frac{\mathrm{d}^2}{\mathrm{d}\lambda^2}\log f(x,\theta) = -\frac{1}{\lambda^2}(\sum_{k=1}^{n} x_k)$$

- Now we need to take expectation to find the Fisher information. Remark: It's good to write about Cramer-Rao lower bound in articles because if we don't want to grind over different algorithms, then we can just derive some bound.

$$I(\theta) = -\mathbb{E}(-\frac{1}{\lambda^2}(\sum_{k=1}^{n} X_k)) = \frac{n}{\lambda}$$

Therefore, the CRLB is $\frac{\lambda}{n}$. The variance of Poisson variable is $\lambda$. Therefore, in the Poisson case, the sample average also achieves the CRLB.

Note that to achieve equality in Cauchy-Schwarz inequality, we need to have linear dependence, i.e.

$$f(x) = k\cdot g(x).$$

In the case of Cramer-Rao, equality means

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(x,\theta) = k(\theta)(\hat{\theta}-\theta)$$

Now let's see what kind of statistics achieve the CRLB. Let's do integration to see what happens:

$$\begin{aligned}\log f(x,\theta) &= \int k(\theta)\hat{\theta}(x) - \int k(\theta)\cdot\theta + h(X)\\ &= A(\theta)\hat{\theta}(X) + B(\theta) + h(X)\\ \implies f(x,\theta) &= \exp(A(\theta)\hat{\theta}(X))\exp(B(\theta))\exp(h(X))\end{aligned}$$

This is called exponential family.

### Exponential Family

A density function belongs to the exponential family if

$$f(x,\theta)=\exp(\sum_{k=1}^{m}T_k(\theta)S_k(X))g(\theta)h(X)$$

If $X\sim f(x,\theta)$ is in the exponential family, then the group of statistics $(S_1(X),\ldots,S_m(X))$ must be sufficient and complete.

### Bernoulli is Exponential Family

$$\begin{aligned}Bern(p):f(x,\theta)&=p^{\sum_{k=1}^{n}x_k}(1-p)^{n-\sum_{k=1}^{n}x_k}\\&=\exp((\sum_{k=1}^{n}x_k)\log p-(\sum_{k=1}^{n}x_k)\log(1-p))(1-p)^n\\&=\exp((\sum_{k=1}^{n}x_k)\log(\frac{p}{1-p}))(1-p)^n\end{aligned}$$

This is exponential family where

$$S_1(X)=\sum_{k=1}^{n}X_k,\quad T_1(p)=\log\frac{p}{1-p},\quad g(p)=(1-p)^n$$

### Poisson is Exponential Family

$$\begin{aligned}Pois(\lambda):f(x,\theta)&=\lambda^{\sum_{k=1}^{n}x_k}\exp(-n\lambda)(\prod_{k=1}^{n}(x_k)!)^{-1}\\&=\exp((\sum_{k=1}^{n}x_k)\log\lambda)\exp(-\lambda n)(\prod_{k=1}^{n}(x_k)!)^{-1}\end{aligned}$$

### Gaussian is Exponential Group

$$f(x,\theta)=\exp(-\frac{1}{2\sigma^2}\sum_{k=1}^{n}x_k^2+\frac{\mu}{\sigma^2}\sum_{k=1}^{n}x_k-\frac{n\mu^2}{2\sigma^2})\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n$$

In the above expression, we have $T_2,S_2,T_1,S_1,g(\mu,\sigma^2)$ respectively.

## 0.12 More on Cramer-Rao Lower Bound

Here's a quick review. Let $X \sim (X_1, \dots X_n)^T \sim P(X, \theta)$. Let $\hat{\theta}(X)$ be our estimator, a function of the sample data. Let $\hat{\theta}$ be unbiased, i.e.

$$\mathbb{E}(\hat{\theta}) = \theta.$$

Therefore,

$$Var(\hat{\theta}) = MSE(\hat{\theta}) = \mathbb{E}(\hat{\theta}(X) - \theta)^2 \geq \frac{1}{I(\theta)}$$

Multi-Dimensional Fisher Information

The Fisher information is

$$I(\theta)_{ij} = \mathbb{E}(\frac{d}{d\theta_i} \log P(X, \theta) \frac{d}{d\theta_j} \log P(X, \theta)) = -\mathbb{E}(\frac{d^2}{d\theta_i d\theta_j} \log P(X, \theta))$$

Now

1. Sometimes, the quantity of our interest is funciton/transformation of the model parameter. In that case, if
$$E(\hat{\theta}) = g(\theta),$$
how do we need to modify the Cramer-Rao bound?
$$g(\theta) = \int_{-\infty}^{\infty} \hat{\theta}(x) p(x, \theta)\,\mathrm{d}x.$$
We apply Fisher's trick:
$$\begin{aligned} g'(\theta) &= \int_{-\infty}^{\infty} \hat{\theta}(x)(\frac{\mathrm{d}}{\mathrm{d}\theta} \log P(x, \theta)) p(x, \theta)\,\mathrm{d}x \\ 0 &= \int_{-\infty}^{\infty} \frac{\mathrm{d}}{\mathrm{d}\theta} p(x, \theta)\,\mathrm{d}x = \int_{-\infty}^{\infty} (\frac{\mathrm{d}}{\mathrm{d}\theta} \log p(x, \theta)) p(x, \theta)\,\mathrm{d}x \\ 0 &= \int_{-\infty}^{\infty} g(\theta)(\frac{\mathrm{d}}{\mathrm{d}\theta} \log p(x, \theta)) p(x, \theta)\,\mathrm{d}x \\ g'(\theta) &= (\hat{\theta}(x) - g(\theta))(\frac{\mathrm{d}}{\mathrm{d}\theta} \log p(x, \theta)) p(x, \theta)\,\mathrm{d}x \end{aligned}$$
The only difference is that the left hand side is $g'(\theta)$ now instead of 1.
$$\mathbb{E}(\hat{\theta}(X) - g(\theta))^2 \geq |g'(\theta)|^2 / I(\theta)$$

2. What if now we consider *Multi-Dimensional* CRLB:
$$\theta \in \mathbb{R}^m, \quad \theta = (\theta_1, \dots, \theta_m)^T$$
Note that usually $m < n$, the number of sample data.

Covariance Matrix

$$Cov(\hat{\theta}) = \mathbb{E}(\hat{\theta} - \mathbb{E}(\hat{\theta}))(\hat{\theta} - \mathbb{E}(\hat{\theta}))^T$$

The diagnal entries are variances.

Multivariate Fisher Information

$$I(\theta) = \mathbb{E}((\nabla_\theta \log p(x,\theta))(\nabla_\theta \log p(x,\theta))^T) = -\mathbb{E}(H_\theta(\log p(x,\theta)))$$

where the gradient $\nabla_\theta$ is

$$\nabla_\theta = (\frac{\mathrm{d}}{d\mathrm{d}\theta_1}, \dots, \frac{\mathrm{d}}{\mathrm{d}\theta_n})$$

and the Hessian $H_\theta$ is

$$(H_\theta(f))_{ij} = \frac{\mathrm{d}^2 f}{\mathrm{d}\theta_i\,\mathrm{d}\theta_j}$$

Multivariate Version of Cramer-Rao

$$Cov(\hat{\theta}) \geq I^{-1}(\theta).$$

We define the ordering on the set of matrices in the following sense:

$$A \geq B \iff A - B \text{ is p.d.}$$

where

$$A \in \mathbb{R}^{n\times m} \text{ is p.d.} \iff \forall \alpha \in \mathbb{R}^n, \alpha^T A \alpha \geq 0$$

Here are some quick reminders about facts about positive definite matrices:

- The eigenvalues of positive definite matrices are all positive.

This is a generalization of the old Cramer-Rao lower bound.

*Proof.* The essence of the scalar version of Cramer-Rao is Cauchy-Schwarz inequality. The proof of the multivariate version is actually different.
First of all,

$$\hat{\theta} : \mathbb{R}^n \to \mathbb{R}^m$$

since $\hat{\theta}$ takes $n$ sample data and outputs an estimate of the model parameter which is $m$-dimensional.

$$g : \mathbb{R}^m \to \mathbb{R}^m$$

Let

$$\alpha_1 = (\hat{\theta}(X) - g(\theta)) \in \mathbb{R}^m$$

Let

$$\alpha_2 = \nabla_\theta \log P(x,\theta) \in \mathbb{R}^m$$

Let

$$\alpha = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} \in \mathbb{R}^{2m}$$

$\mathbb{E}(\alpha\alpha^T)$ is definitely positive definite. Let $y \in \mathbb{R}^{2m}$.

$$\begin{aligned} y^T(\mathbb{E}(\alpha\alpha^T))y &= \mathbb{E}(y^T\alpha\alpha^T y) \quad \text{by linearity} \\ &= \mathbb{E}((\alpha^T y)^2) \geq 0 \end{aligned}$$

Now note that

$$\mathbb{E}(\alpha\alpha^T) = \mathbb{E}\begin{bmatrix} \alpha_1\alpha_1^T & \alpha_1\alpha_2^T \\ \alpha_2\alpha_1^T & \alpha_2\alpha_2^T \end{bmatrix}$$

$$\mathbb{E}(\alpha_1\alpha_1^T) = \mathbb{E}((\hat{\theta}(X) - g(\theta))(\hat{\theta}(X) - g(\theta))^T) = Cov(\hat{\theta})$$

$$\mathbb{E}(\alpha_2\alpha_2^T) = \mathbb{E}(\nabla_\theta \log P(X,\theta))(\nabla_\theta \log P(X,\theta))^T = I(\theta)$$

$$\begin{aligned} \mathbb{E}(\alpha_1\alpha_2^T) &= \mathbb{E}\left((\hat{\theta}(X) - g(\theta))(\nabla_\theta \log P(X,\theta))^T\right) \\ &= \int_{\mathbb{R}^n} (\hat{\theta}(X) - g(\theta))(\nabla_\theta \log P(X,\theta)^T)dX \end{aligned}$$

Let's take a look at the $ij$-th entry of the above matrix.

$$\begin{aligned} \mathbb{E}(\alpha_1\alpha_2^T)_{ij} &= \mathbb{E}\left((\hat{\theta}_i(X) - g_i(\theta))(\frac{d}{d\theta_j}\log P(X,\theta))\right) \\ &= \int_{\mathbb{R}^n}\left((\hat{\theta}_i(X) - g_i(\theta))(\frac{d}{d\theta_j}\log P(X,\theta))\right)P(X,\theta)dX \end{aligned}$$

Note that

$$0 = \int_{\mathbb{R}^n}\frac{d}{d\theta_j}P(X,\theta)dx$$

Then

$$0 = \int_{\mathbb{R}^n}(\frac{d}{d\theta_j}\log P(X,\theta))P(X,\theta)dX = \int_{\mathbb{R}^n} g(\theta)(\frac{d}{d\theta_j}\log P(X,\theta))P(X,\theta)dX$$

$$g_i(\theta) = \int_{\mathbb{R}^n}\hat{\theta}_i(X)P(X,\theta)dX$$

$$\frac{dg_i}{d\theta_j} = \int_{\mathbb{R}^n}\hat{\theta}_i(X)(\frac{d}{d\theta_j}\log P(X,\theta))P(X,\theta)dX$$

Therefore,

$$\mathbb{E}(\alpha_1\alpha_2^T) = \frac{dg_i}{d\theta_j}$$

This is called the Jacobian matrix of $g$, i.e.

$$\mathbb{E}(\alpha_1\alpha_2^T) = J_\theta g \implies \mathbb{E}(\alpha_2\alpha_1^T) = (J_\theta g)^T$$

Note that usually Hessian is symmetric unless there's problem with smoothness, but Jacobian is not symmetric because $\frac{dg_i}{d\theta_j} \neq \frac{dg_j}{d\theta_i}$. Now

$$\mathbb{E}(\alpha\alpha^T) = \begin{bmatrix} Cov(\hat{\theta}) & J_\theta g \\ (J_\theta g)^T & I(\theta) \end{bmatrix} \geq 0$$

Let's do a matrix transformation to get rid of the non-diagonal entries. This is a technique called "holing."

$$\begin{bmatrix} I & -(J_\theta g)I^{-1}(\theta) \\ 0 & I \end{bmatrix} \begin{bmatrix} Cov(\hat{\theta}) & J_\theta g \\ (J_\theta g)^T & I(\theta) \end{bmatrix} = \begin{bmatrix} Cov(\hat{\theta}) - (J_\theta g)I^{-1}(\theta)(J_\theta g)^T & 0 \\ (J_\theta g)^T & I(\theta) \end{bmatrix}$$

$$\begin{bmatrix} I & -(J_\theta g)I^{-1}(\theta) \\ 0 & I \end{bmatrix} \begin{bmatrix} Cov(\hat{\theta}) & J_\theta g \\ (J_\theta g)^T & I(\theta) \end{bmatrix} \begin{bmatrix} I & 0 \\ -I^{-1}(\theta)(J_\theta g)^T & I \end{bmatrix}$$

$$= \begin{bmatrix} Cov(\hat{\theta}) - (J_\theta g)I^{-1}(\theta)(J_\theta g)^T & 0 \\ 0 & I(\theta) \end{bmatrix}$$

If $A$ is positive definite, then $B^TAB$ is also positive definite. Therefore,

$$Cov(\hat{\theta}) - (J_\theta g)I^{-1}(\theta)(J_\theta g)^T$$

is positive definite. (Random remark: Luogeng Hua is so comfortable with matrices that he uses them as integers, John von Neumann said that. Be like Luogeng Hua.) Therefore

$$Cov(\hat{\theta}) \geq (J_\theta g)I^{-1}(\theta)(J_\theta g)^T$$

Note that for the case of unbiased statistics, $J_\theta \theta$ is just the identity matrix. □

Next, we will discuss specific estimation technique. As a bound, Cramer-Rao gives us more insights about our estimation problems but Cramer-Rao does not solve our problems.

## 0.13 Linear Estimation

Recall that if we want to estimate $Y$ using a function $g$ of the data. Then we know that

$$g_{opt}(X) = \mathbb{E}(Y|X)$$

However, conditional expectation is usually difficult to obtain.
Among all the estimation techniques, linear estimation is the easiest. We consider the estimator as

$$g(X) = \alpha X.$$

Let $n$ be the number of sample data (i.e. $X$ is $n$-dimensional) and $Y$ be $m$-dimensional. Consider the mean square metric as the measure of estimator performance.

(a) If $m = 1, n = 1$, then we are considering

$$\min_\alpha \mathbb{E}(Y - \alpha X)^2, \quad h(\alpha) = \mathbb{E}(Y - \alpha X)^2$$

This is an optimization problem so we differentiate $h$ with respect to $\alpha$.

$$\frac{d}{d\alpha} h(\alpha) = -2\mathbb{E}(X(Y - \alpha X))$$

$$\alpha = (\mathbb{E}X^2)^{-1}\mathbb{E}(XY)$$

We now interpret the above terms from a geometry perspective.

- First of all, the cross moment operator

  $$\mathbb{E}(XY) = \langle X, Y \rangle$$

  is an inner product operator $H \times H \to \mathbb{R}$.
- Recall that an inner product is symmetric, positive definite ($\langle X, X \rangle \geq 0$ with equality if and only if $X = 0$) and bilinear. The cross moment operator is apparently an inner product.
- Now that we have a notion of inner product, then we can view random variables geometrically:

  $$\cos(X, Y) = \frac{\langle X, Y \rangle}{\|X\| \cdot \|Y\|} = \frac{\langle X, Y \rangle}{|\langle X, X \rangle \langle Y, Y \rangle|^{\frac{1}{2}}}$$

  To minimize the distance between $Y$ and $X$, we want the line between $Y$ and $X$ to be perpendicular to $X$. To find the length of $X$, we simply notice that it should be equal to

  $$\|Y\| \cos(X, Y)$$

  Therefore, the optimal linear estimator under the MSE metric is

  $$\frac{\langle X, Y \rangle}{\|X\|} \cdot \frac{X}{\|X\|} = \frac{\langle X, Y \rangle}{\|X\|^2} X = \frac{\mathbb{E}[XY]}{\mathbb{E}[X^2]} X$$

(b) Suppose $m = 1, n > 1$, i.e. we have a collection of sample data $X = (X_1, \ldots, X_n)^T$ to estimate $Y$. To make a linear estimator, choose a collection of linear coefficients

$$\theta = (\theta_1, \ldots, \theta_n)^T, \quad \theta^T X = \sum_{k=1}^{n} \theta_k X_k$$

Now consider

$$h(\theta) = \mathbb{E}(Y - \theta^T X)^2 \implies \frac{dh}{d\theta_k} = -2\mathbb{E}\left[(Y - \sum_{i=1}^{n} \theta_i X_i) X_k\right] = 0, \quad k = 1, \ldots, n$$

- Orthogonality is a very important concept here: we what the *residue* $Y - X$ to be perpendicular to $X$. In the above equations, we see that to solve the optimization problem, $X_k$ must be orthogonal to the residue

  $$Y - \sum_{i=1}^{n} \theta_i X_i$$

-

  $$\begin{aligned} h(\theta) &= \mathbb{E}(Y - \theta^T X)^2 = \mathbb{E}Y^2 - 2\mathbb{E}(Y\theta^T X) + \theta^T \mathbb{E}(XX^t)\theta \\ &= \mathbb{E}Y^2 - 2\theta^T \mathbb{E}(YX) + \theta^T \mathbb{E}(XX^T)\theta \end{aligned}$$

  We now take the gradient with respect to $\theta$. Here is a quick review of something useful high-dimensional derivatives.

- $$\nabla_\theta \theta^T \alpha = \alpha = \nabla_\theta(\alpha^T\theta)$$

- $$\nabla_\theta(\theta^T A\theta) = (A + A^T)\theta$$

$$\nabla_\theta h(\theta) = -2\mathbb{E}(XY) + 2\mathbb{E}(XX^T)\theta = 0 \implies \theta = (\mathbb{E}(XX^T))^{-1}\mathbb{E}(XY)$$

(c) In our final case, assume $m > 1, n > 1$. Then we want to minimize:

$$\min_\theta \sum_{k=1}^m \mathbb{E}(Y_k - \theta^T X_k)^2$$

Alternatively, we can weight the errors:

$$\min_\theta \sum_{k=1}^m \lambda_k \mathbb{E}(Y_k - \theta^T X_k)^2$$

Another common practice is "Minimax:"

$$\min_\theta \max_k \mathbb{E}(Y_k - \theta^T X_k)^2$$

Now the data vector

$$X = \begin{pmatrix} X_1^T \\ \vdots \\ X_m^T \end{pmatrix} \in \mathbb{R}^{m\times n}$$

Now

$$\sum_{k=1}^m \mathbb{E}(Y_k - \theta^T X_k)^2 = \mathbb{E}(Y - X\theta)^T(Y - X\theta)$$

Our target function is then

$$h(\theta) = \mathbb{E}(Y - X\theta)^T(Y - X\theta) = \mathbb{E}[Y^TY] - \mathbb{E}[\theta^T X^T Y] - \mathbb{E}[Y^T X\theta] + \theta^T\mathbb{E}(X^TX)\theta$$

Let's compute the gradient:

$$\nabla_\theta h(\theta) = 0 - 2\mathbb{E}[X^TY] + 2\mathbb{E}[X^TX]\theta = 0$$

Then

$$\theta = (\mathbb{E}[X^TX])^{-1}\mathbb{E}(X^TY)$$

## 0.14 Continuous-Time Signal Linear Estimation

Consider a stochastic signal as a continuous-time stochastic process. Then a linear system $h$ would take $X(t)$ as the input and output

$$\hat{Y}(t) = \int_{-\infty}^{\infty} h(t-\tau)X(t)d\tau$$

The goal is to minimize:

$$\min_h \mathbb{E}\left(Y(t) - \int_{-\infty}^{\infty} h(t-\tau)X(\tau)d\tau\right)^2$$

- Although we are not necessarily able to differentiate the target function, but we know that the optimal residue is orthogonal to $X$ at any time. Note that we also assume stationarity of the system. (i.e. correlation only depends on difference in time.)
-

$$\begin{aligned} &\mathbb{E}\left(\left(Y(t) - \int_{-\infty}^{\infty} h_{opt}(t-\tau)X(\tau)d\tau\right)X(S)\right) = 0 \quad \forall S \\ \implies &\mathbb{E}[Y(t)X(S)] - \int_{-\infty}^{\infty} h_{opt}(t-\tau)\mathbb{E}(X(\tau)X(s))d\tau = 0 \quad \forall S \\ \implies &R_{YX}(t-S) = \int_{-\infty}^{\infty} h_{opt}(t-\tau)R_X(\tau - S)d\tau \\ &R_{YX}(\tau) = (h_{opt} * R_X)(\tau) \implies S_{YX}(\omega) = H_{opt}(\omega)\cdot S_X(\omega) \\ &H_{opt}(\omega) = (S_X(\omega))^{-1}S_{XY}(\omega) \end{aligned}$$

To summarize,

$$\theta_{opt} = (\mathbb{E}(X^TX))^{-1}\mathbb{E}(X^TY)$$

The first term is used for normalizing. The second term is correlation, which corresponds to the "angle" or "alignment."

Recall the concept of projection matrix:

$$Proj_A X = A(A^TA)^{-1}A^TX$$

This result employs the same orthogonality idea.
Remark: Don't doubt your mathematical ability. This is not math; this is just symbols. Von Neumann said: "We never truly understood something. We only get used to things." If you pretend to understand something long enough, you will start to really understand it. We never truly understood love - we just get more and more used to each other :)

## 0.15 Wiener Filtering

What we have been covering in this course has mainly been statistical modeling. The next topic focuses on signal processing.
Suppose that we have a collection of sample data $X_1, X_2, \ldots, X_n$. Our purpose is to use the data to estimate a parameter $\theta \in \mathbb{R}^m$ with an estimator $\hat{\hat{\theta}}(X_1, \ldots, X_m)$.

Filtering

Filtering:

- $\theta_k, k = 1, 2, \ldots, n$ changes over time.