

EE 278: Probability and Statistical Inference

Lecture Notes, Spring 2026

Abhiram Gorle, Instructor: Tsachy Weissman

September 21, 2026

Contents

1	Lecture 1: Course Overview	2
2	Lecture 2: Random Variables and Transformations	11
3	Lecture 3: Expectation, Variance, and MMSE Estimation	19
4	Lecture 4: Inner-Product Geometry, Linear Estimation, and Jointly Gaussian Random Variables	32
5	Lecture 5: Random Vectors, Gaussian Random Vectors	44
6	Lecture 6: Covariance Matrices and Linear MMSE Estimation	54
7	Lecture 7: Gaussian Conditioning and Innovations	60
8	Lecture 8: Kalman Filtering	65
9	Lecture 9: Kalman Filtering II	71
10	Lecture 10: Random Processes I	77
11	Lecture 11: Random Processes II	93
12	Lecture 12: Stationarity and Power Spectral Density	101
13	Lecture 13: Linear Systems with Random Process Inputs	108
14	Lecture 14: Cross-Correlation, Cross-Spectral Density, and White Noise	114
15	Lecture 15: Linear Estimation, Wiener Filtering, and Spectral Factorization	120
16	Lecture 16: Spectral Factorization and Causal Wiener Filtering	125
17	Lecture 17: Causal Wiener Filtering, Lookahead, and Linear Prediction	131
18	Lecture 18: Course Wrap-Up and Prediction in Practice	136

Lecture 1: Course Overview

For all administrative details, please refer to the [course website](#).

About the course

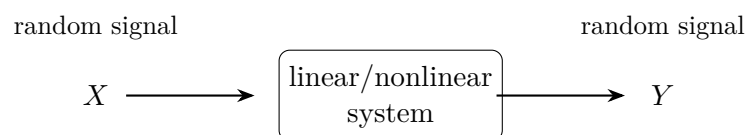
Historically, this course has been framed as a course on “*Introduction to Statistical Signal Processing*.”

In this offering, we will instead take a broader and more modern viewpoint and focus on an intersection of topics in signal processing and statistical inference, presenting it as a course on “*Probabilistic Signals and Systems*”

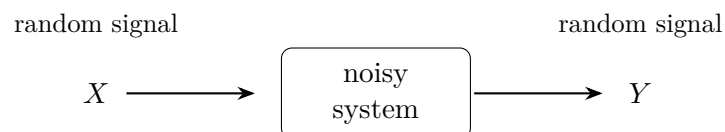
The emphasis shifts from purely signal processing techniques to a more unified probabilistic perspective on signals, systems, and inference. Signals are inherently random, and systems operate on these random objects. Understanding their probabilistic structure is central to modern engineering, data science, and machine learning.

Elements of the course

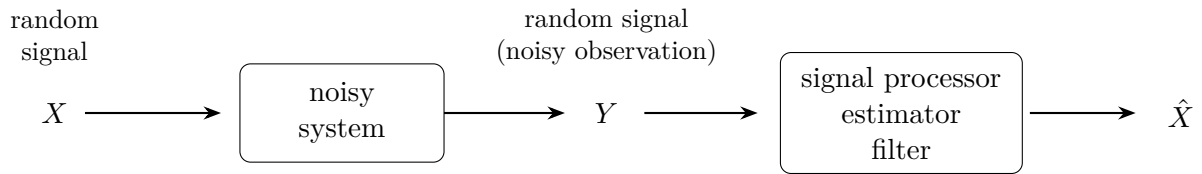
- A central theme of the course is the study of **stochastic processes**, which serve as mathematical models for **random signals**. In particular, the course will emphasize:
 - **Formalization:** rigorous characterization and basic properties of random processes
 - **Canonical examples:** Gaussian processes, Poisson processes, and other standard models
 - **Characterization:** joint distributions, finite-dimensional distributions, autocorrelation functions, spectral densities, and related descriptors
- Another central theme of the course is understanding how **random signals** behave when they pass through systems. Two important settings of interest include:
 - **Deterministic systems:** a random input signal is transformed by a linear or nonlinear system, and we study the probabilistic relationship between the input and output.



- **Noisy systems:** the system itself introduces additional randomness, so the output depends both on the input signal and on the noise injected by the system.



- Another central theme of the course is the **processing of noisy signals**. The basic problem is to recover an underlying signal X from a noisy observation Y produced by a system. This leads naturally to questions pertaining to **estimation**, **filtering**, and **signal processing** more broadly.



Typical examples of underlying signals include digital communication signals, audio, images, and biomedical signals such as EEG.

- Another important theme of the course is to quantify what it means to **recover well**. Depending on the problem, this may involve criteria for **estimation** or for **detection**. Once such performance measures (estimation error, detection performance, etc.) are made precise, they guide the design of **optimal algorithms** for signal processing, including classical methods such as **Wiener filtering** and **Kalman filtering** (which will be introduced in the course).

Big Picture

A key modeling viewpoint in this course is that the underlying **probabilistic model is known**. That is, we will generally assume known statistical descriptions for the **signal of interest**, the **noise**, and the **observation mechanism**. This corresponds to a **Bayesian formulation** of the inference problem. The emphasis of the course is therefore on how to perform estimation, detection, and related inference tasks when the model is available. By contrast, in later courses on **statistical inference**, **information theory**, and especially **machine learning**, one often studies how to **learn the model itself** from data.

A prototypical inference problem

A basic inference problem of interest throughout the course can be represented by the pipeline

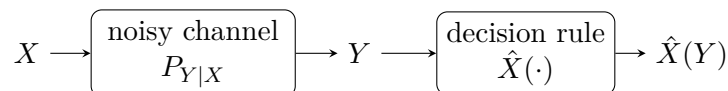
$$X \longrightarrow Y \longrightarrow \hat{X},$$

where:

- X denotes the underlying signal, message, state, or hypothesis of interest,
- Y denotes the observation available to us,
- $\hat{X} = \hat{X}(Y)$ denotes the estimate or decision formed from the observation.

In many applications, X is not directly observable. Instead, we only have access to an observation Y that is related to X through some probabilistic mechanism. The goal is then to infer X from Y . Depending on the context, this same mathematical structure appears under different names, including *detection*, *hypothesis testing*, *decoding*, *classification*, and *decision making*. Although the terminology varies across fields, the underlying problem is essentially the same: use the observation Y to decide which value of X most plausibly generated it.

A convenient block-diagram representation is the following:



Here the block labeled $P_{Y|X}$ represents the probabilistic law relating the hidden variable X to the

observation Y . In communication this may literally be a noisy channel; in statistics it may be an observation model; in machine learning it may be viewed as a data-generation mechanism.

Discrete formulation and probability of error

More formally, suppose X is a discrete random variable taking values in a finite alphabet

$$X \in \mathcal{X} = \{x_1, x_2, \dots, x_m\},$$

with prior distribution P_X , and suppose the observation Y takes values in

$$Y \in \mathcal{Y} = \{y_1, y_2, \dots, y_\ell\}.$$

A detector (or decision rule) is any mapping

$$\hat{X} : \mathcal{Y} \rightarrow \mathcal{X}.$$

A natural performance criterion is the *probability of error*

$$P_e = \mathbb{P}(\hat{X} \neq X).$$

Equivalently,

$$P_e = 1 - \mathbb{P}(\hat{X} = X).$$

Thus minimizing P_e is the same as maximizing the probability of correct decision.

Probability of correct decision

Since Y is discrete, we can decompose according to the possible observed values of Y and further condition on its possible values:

$$\mathbb{P}(\hat{X} = X) = \sum_{i=1}^{\ell} \mathbb{P}(X = \hat{X}(Y), Y = y_i).$$

Therefore,

$$P_e = 1 - \sum_{i=1}^{\ell} \mathbb{P}(X = \hat{X}(Y), Y = y_i).$$

Now, we can write

$$\mathbb{P}(X = \hat{X}(Y), Y = y_i) = \mathbb{P}(X = \hat{X}(Y) \mid Y = y_i) \mathbb{P}(Y = y_i),$$

and substituting this gives

$$P_e = 1 - \sum_{i=1}^{\ell} \mathbb{P}(X = \hat{X}(Y) \mid Y = y_i) \mathbb{P}(Y = y_i).$$

But on the event $\{Y = y_i\}$, the detector output is simply the fixed value $\hat{X}(y_i)$. Hence

$$P_e = 1 - \sum_{i=1}^{\ell} \mathbb{P}(X = \hat{X}(y_i) \mid Y = y_i) \mathbb{P}(Y = y_i).$$

It is often convenient to write this using the posterior distribution $P_{X|Y}$:

$$P_e = 1 - \sum_{i=1}^{\ell} P_{X|Y}(\hat{X}(y_i) | y_i) P_Y(y_i).$$

Why is this expansion useful? The expression

$$P_e = 1 - \sum_{i=1}^{\ell} P_{X|Y}(\hat{X}(y_i) | y_i) P_Y(y_i)$$

reveals an important structural fact: the contribution of the decision rule at one observation value y_i is separated from its contribution at every other observation value. Indeed, once the observation alphabet $\{y_1, \dots, y_\ell\}$ is fixed, the rule $\hat{X}(\cdot)$ affects the above sum only through the individual choices

$$\hat{X}(y_1), \hat{X}(y_2), \dots, \hat{X}(y_\ell).$$

For a given observation y_i , the term involving that choice is

$$P_{X|Y}(\hat{X}(y_i) | y_i) P_Y(y_i).$$

Since $P_Y(y_i)$ is nonnegative and does not depend on which value of x we choose as $\hat{X}(y_i)$, the only quantity we can influence is

$$P_{X|Y}(\hat{X}(y_i) | y_i).$$

Therefore, to maximize the overall probability of correct decision, it is enough to maximize this posterior probability separately for each possible observation y_i . In other words, once $Y = y_i$ has been observed, the best decision is to declare the value of x that is most probable under the posterior distribution conditioned on that observation.

The MAP rule

The discussion above shows that minimizing P_e is equivalent to choosing, for each $y \in \mathcal{Y}$, a value $\hat{X}(y)$ that maximizes

$$P_{X|Y}(\hat{X}(y) | y).$$

Thus an optimal decision rule is

$$\hat{X}(y) \in \arg \max_{x \in \mathcal{X}} P_{X|Y}(x | y).$$

This is called the *maximum a posteriori* (MAP) rule.

Definition 1.1. The *MAP decision rule* is any rule of the form

$$\hat{X}_{\text{MAP}}(y) \in \arg \max_{x \in \mathcal{X}} P_{X|Y}(x | y),$$

that is, for each observed value y , it chooses a value of x having largest posterior probability given $Y = y$.

Interpretation: Before any observation is made, our uncertainty about X is described by the prior distribution P_X . After observing $Y = y$, that uncertainty is updated to the conditional distribution $P_{X|Y}(\cdot | y)$. The MAP rule simply says: once the data y has been seen, choose the value of X that is now most probable in light of that data.

After observing y , guess the value of X with highest posterior probability.

Why is it called *a posteriori*?

The phrase *a posteriori* means “after the observation” or “after the data has been seen.” It refers to the fact that the decision is based on the **posterior** distribution

$$P_{X|Y}(x | y),$$

rather than on the prior distribution $P_X(x)$.

The distinction is the following:

- the **prior** $P_X(x)$ describes our probabilistic belief about X before observing Y ;
- the **posterior** $P_{X|Y}(x | y)$ describes our updated probabilistic belief about X after observing $Y = y$.

Thus the MAP rule is called *maximum a posteriori* because it selects the value of x that maximizes the posterior probability.

Posterior distribution via Bayes' rule

The posterior distribution can be expressed in terms of the prior and the observation model through Bayes' rule:

$$P_{X|Y}(x | y) = \frac{P_X(x) P_{Y|X}(y | x)}{\sum_{\tilde{x} \in \mathcal{X}} P_X(\tilde{x}) P_{Y|X}(y | \tilde{x})}.$$

This identity is fundamental: it shows that the posterior combines two ingredients,

- the **prior** term $P_X(x)$, which captures how plausible x was before seeing the data;
- the **likelihood** term $P_{Y|X}(y | x)$, which captures how compatible the observed value y is with the hypothesis $X = x$.

Since the denominator does not depend on x , maximizing the posterior over x is equivalent to maximizing the numerator. Hence the MAP rule may also be written as

$$\hat{X}_{\text{MAP}}(y) \in \arg \max_{x \in \mathcal{X}} P_X(x) P_{Y|X}(y | x).$$

Interpretation of the two factors: Under the MAP rule, a candidate value x is favored when two considerations work in its favor:

- it was already relatively plausible *a priori*, as reflected by $P_X(x)$;
- it explains the observation y well, as reflected by the likelihood $P_{Y|X}(y | x)$.

Thus MAP balances prior belief and data fit in a natural Bayesian manner.

MAP versus Maximum Likelihood (ML)

It is useful to compare MAP with the *maximum likelihood* (ML) rule. If the **prior distribution** P_X is **uniform on $\mathcal{X}$** , then $P_X(x)$ is the same for every x , and so it does not affect the maximization.

In that case,

$$\hat{X}_{\text{MAP}}(y) \in \arg \max_{x \in \mathcal{X}} P_{Y|X}(y | x),$$

which is precisely the maximum likelihood (ML) rule.

Thus:

- **MAP** uses both the prior $P_X(x)$ and the likelihood $P_{Y|X}(y | x)$;
- **ML** uses only the likelihood and ignores the prior.

Takeaway

Many problems in detection, decoding, classification, or hypothesis testing fit into the same probabilistic template presented earlier

$$X \rightarrow Y \rightarrow \hat{X}.$$

When the goal is to minimize probability of error, the optimal decision is obtained by choosing, for each observation y , the value of X with largest posterior probability $P_{X|Y}(x | y)$. This is the MAP rule.

Example: binary signaling over an additive Gaussian noise channel

We now illustrate the MAP rule on a classic example from detection theory: binary signaling over an additive Gaussian noise channel.

Assume the observation is

$$Y = X + Z,$$

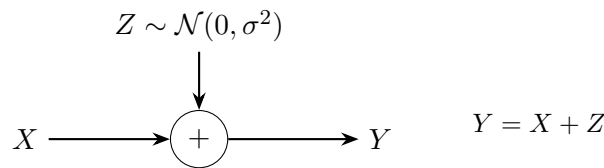
where

$$Z \sim \mathcal{N}(0, \sigma^2),$$

and X and Z are independent. The input X is binary and takes the two values

$$X \in \{-\sqrt{P}, +\sqrt{P}\}, \mathbb{P}(X = +\sqrt{P}) = \mathbb{P}(X = -\sqrt{P}) = \frac{1}{2}.$$

Thus the transmitted signal is equally likely to be either of the two antipodal points $\pm\sqrt{P}$, and from the noisy observation $Y = y$ we would like to decide which one was sent.



Applying the MAP rule: As we saw above, the MAP rule chooses the value of $x \in \{-\sqrt{P}, +\sqrt{P}\}$ that maximizes the posterior probability $P_{X|Y}(x | y)$. Since the two priors are equal,

$$P_X(+\sqrt{P}) = P_X(-\sqrt{P}) = \frac{1}{2},$$

the prior does not favor either hypothesis, and so the MAP rule reduces to the maximum-likelihood rule:

$$\hat{X}_{\text{MAP}}(y) \in \arg \max_{x \in \{-\sqrt{P}, +\sqrt{P}\}} f_{Y|X}(y | x).$$

Thus we decide $+\sqrt{P}$ when

$$f_{Y|X}(y | +\sqrt{P}) > f_{Y|X}(y | -\sqrt{P}),$$

and decide $-\sqrt{P}$ when the reverse inequality holds.

Conditional distribution of Y : Conditioned on $X = x$, we have

$$Y = x + Z.$$

Since $Z \sim \mathcal{N}(0, \sigma^2)$, it follows that

$$Y | \{X = x\} \sim \mathcal{N}(x, \sigma^2),$$

and therefore

$$f_{Y|X}(y | x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-x)^2}{2\sigma^2}\right).$$

In particular,

$$f_{Y|X}(y | +\sqrt{P}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\sqrt{P})^2}{2\sigma^2}\right),$$

and

$$f_{Y|X}(y | -\sqrt{P}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y+\sqrt{P})^2}{2\sigma^2}\right).$$

Deriving the decision rule: The MAP detector chooses $+\sqrt{P}$ whenever

$$f_{Y|X}(y | +\sqrt{P}) > f_{Y|X}(y | -\sqrt{P}).$$

Substituting the Gaussian densities yields

$$\exp\left(-\frac{(y-\sqrt{P})^2}{2\sigma^2}\right) > \exp\left(-\frac{(y+\sqrt{P})^2}{2\sigma^2}\right).$$

Since the exponential function is strictly increasing, we may compare the exponents:

$$-\frac{(y-\sqrt{P})^2}{2\sigma^2} > -\frac{(y+\sqrt{P})^2}{2\sigma^2}.$$

Multiplying by $-2\sigma^2$ reverses the inequality, giving

$$(y-\sqrt{P})^2 < (y+\sqrt{P})^2.$$

Expanding both sides,

$$y^2 - 2y\sqrt{P} + P < y^2 + 2y\sqrt{P} + P,$$

so

$$-2y\sqrt{P} < 2y\sqrt{P}.$$

Since $\sqrt{P} > 0$, this simplifies to

$$y > 0.$$

We conclude that the MAP rule is

$$\hat{X}_{\text{MAP}}(y) = \begin{cases} +\sqrt{P}, & y > 0, \\ -\sqrt{P}, & y < 0. \end{cases}$$

At the tie point $y = 0$, both choices are equally good, and either may be selected.

Interpretation: This example has a very simple geometric interpretation. The two possible transmitted values are the points

$$-\sqrt{P} \text{ and } +\sqrt{P}$$

on the real line, and the observation y is a noisy version of whichever point was transmitted. The detector declares the signal point that is more likely to have produced the observed value. Because the hypotheses are symmetric and equally likely, the decision boundary lies exactly at the midpoint between the two signal points, namely at 0.

Thus the detector reduces to the intuitive rule:

decide $+\sqrt{P}$ if the observation is positive, and decide $-\sqrt{P}$ if the observation is negative.

Probability of error: Having derived the detector, it is natural to probe how often it makes an error. In this example, the rule is especially simple: we decide $+\sqrt{P}$ when $Y > 0$ and $-\sqrt{P}$ when $Y < 0$. Thus an error occurs exactly when the noise is large enough to push the observation across the decision boundary at 0.

If $X = +\sqrt{P}$, then

$$Y = +\sqrt{P} + Z,$$

and an error occurs when the detector decides $-\sqrt{P}$, that is, when $Y < 0$. This is equivalent to

$$+\sqrt{P} + Z < 0 \iff Z < -\sqrt{P}.$$

Similarly, if $X = -\sqrt{P}$, then

$$Y = -\sqrt{P} + Z,$$

and an error occurs when $Y > 0$, which is equivalent to

$$-\sqrt{P} + Z > 0 \iff Z > \sqrt{P}.$$

Therefore,

$$P_e = \frac{1}{2} \mathbb{P}(Y < 0 \mid X = +\sqrt{P}) + \frac{1}{2} \mathbb{P}(Y > 0 \mid X = -\sqrt{P}).$$

By symmetry of the setup here, these two conditional error probabilities are equal, and so

$$P_e = \mathbb{P}(Z > \sqrt{P}).$$

Since $Z \sim \mathcal{N}(0, \sigma^2)$, this Gaussian tail probability is

$$P_e = Q\left(\frac{\sqrt{P}}{\sigma}\right),$$

where

$$Q(t) = \frac{1}{\sqrt{2\pi}} \int_t^\infty e^{-u^2/2} du.$$

Equivalently, in terms of the signal-to-noise ratio P/σ^2 ,

$$P_e = Q\left(\sqrt{\frac{P}{\sigma^2}}\right).$$

Thus, after obtaining the MAP rule explicitly, we can also quantify its performance in closed form for this simple binary Gaussian model.

Lecture 2: Random Variables and Transformations

Random Variables

We now begin with a formal review of *random variables*, which provide a numerical way to describe outcomes of a random experiment.

A random variable X is a function from the sample space Ω to the real line:

$$X : \Omega \rightarrow \mathbb{R}.$$

Rather than working directly with outcomes $\omega \in \Omega$, we often study the induced probabilistic behavior of X .

Cumulative distribution function (CDF): For any random variable X , its *cumulative distribution function* is defined by

$$F_X(x) \triangleq \mathbb{P}(X \leq x), \quad x \in \mathbb{R}.$$

The CDF is one of the most fundamental ways to describe the distribution of a random variable. It is defined for *every* random variable, whether discrete, continuous, or mixed.

Some basic properties of F_X are:

1. $0 \leq F_X(x) \leq 1$,
2. $F_X(x)$ is nondecreasing in x ,
3. $\lim_{x \rightarrow -\infty} F_X(x) = 0$, $\lim_{x \rightarrow \infty} F_X(x) = 1$, and F_X is right-continuous.

Discrete random variables and the PMF: If X is discrete, then its distribution can be described by its *probability mass function* (PMF),

$$p_X(x) \triangleq \mathbb{P}(X = x).$$

In this case, the CDF can be written as

$$F_X(x) = \sum_{t \leq x} p_X(t).$$

Thus, for a discrete random variable, the PMF gives the probability assigned to each point, while the CDF gives the accumulated probability up to x .

Absolutely continuous random variables and the PDF: A random variable X is *absolutely continuous* if there exists a nonnegative function f_X such that

$$F_X(x) = \int_{-\infty}^x f_X(u) du.$$

The function f_X is called the probability density function of X .

Equivalently, F_X is absolutely continuous, its derivative exists almost everywhere, and

$$f_X(x) = F'_X(x)$$

for almost every x .

For an absolutely continuous random variable, point probabilities vanish:

$$\mathbb{P}(X = x) = 0 \text{ for all } x,$$

and interval probabilities are obtained by integrating the density:

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(u) du.$$

Function of a random variable:

A very common problem is to understand the distribution of a transformed random variable

$$Y = g(X),$$

where g is some deterministic function.

The starting point is always the CDF of Y :

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(g(X) \leq y).$$

This identity is fundamental: to find the distribution of Y , we rewrite the event $\{g(X) \leq y\}$ in terms of X , and then evaluate its probability using the distribution of X .

Monotone transformation rule: If g is strictly increasing and invertible, then

$$F_Y(y) = \mathbb{P}(g(X) \leq y) = \mathbb{P}(X \leq g^{-1}(y)) = F_X(g^{-1}(y)).$$

If g is strictly decreasing and invertible, then

$$F_Y(y) = \mathbb{P}(g(X) \leq y) = \mathbb{P}(X \geq g^{-1}(y)).$$

When densities exist and g is one-to-one and differentiable, this leads to the change-of-variables formula

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right|.$$

Example 1: Affine transformation. Let $Y = aX + b$, $a > 0$. Then

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(aX + b \leq y).$$

Since $a > 0$, dividing by a preserves the inequality, so

$$F_Y(y) = \mathbb{P}\left(X \leq \frac{y-b}{a}\right) = F_X\left(\frac{y-b}{a}\right).$$

If X has a density f_X , then differentiating gives

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{1}{a} f_X\left(\frac{y-b}{a}\right).$$

Thus, an affine transformation shifts and rescales the distribution as shown above.

Example 2: The probability integral transform. Suppose F_X is strictly increasing and continuous, so that the inverse function F_X^{-1} exists. Define $Y = F_X(X)$, we now determine the distribution of Y .

First, since $0 \leq F_X(x) \leq 1$ for every x , it follows immediately that $0 \leq Y \leq 1$. Hence,

$$F_Y(y) = \mathbb{P}(Y \leq y)$$

must satisfy

$$F_Y(y) = 0 \quad \text{for } y < 0,$$

and

$$F_Y(y) = 1 \quad \text{for } y > 1.$$

It remains to evaluate $F_Y(y)$ for $0 \leq y \leq 1$. For such y ,

$$F_Y(y) = \mathbb{P}(F_X(X) \leq y).$$

Since F_X is strictly increasing, we may apply F_X^{-1} to both sides of the inequality:

$$F_X(X) \leq y \iff X \leq F_X^{-1}(y).$$

Therefore,

$$F_Y(y) = \mathbb{P}(X \leq F_X^{-1}(y)) = F_X(F_X^{-1}(y)) = y.$$

Combining all three cases, we obtain

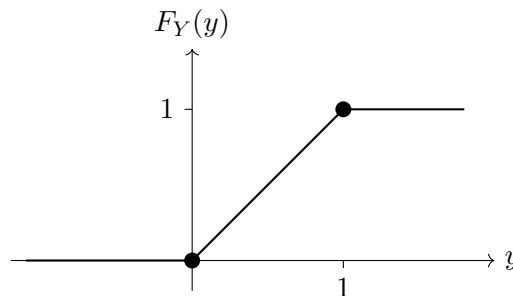
$$F_Y(y) = \begin{cases} 0, & y < 0, \\ y, & 0 \leq y \leq 1, \\ 1, & y > 1. \end{cases}$$

This is precisely the CDF of a uniform random variable on $[0, 1]$. Hence,

$$Y \sim \text{Unif}[0, 1].$$

Remark. This result is known as the *probability integral transform*. It is extremely useful in both theory and simulation: applying a CDF to a random variable produces a uniform random variable on $[0, 1]$, provided the CDF is continuous and strictly increasing.

Graphical intuition. The CDF is flat at 0 to the left of the origin, rises linearly with slope 1 on $[0, 1]$, and stays flat at 1 thereafter. Figure below shows the CDF of $Y \sim \text{Unif}[0, 1]$.



Converse: The above result showed that if X has continuous, strictly increasing CDF F_X , then

$$F_X(X) \sim \text{Unif}[0, 1].$$

Conversely, if

$$U \sim \text{Unif}[0, 1],$$

then applying the inverse CDF recovers a random variable with distribution F_X . More precisely,

$$F_X^{-1}(U) \sim F_X,$$

or equivalently,

$$F_X^{-1}(U) \stackrel{d}{=} X,$$

where $\stackrel{d}{=}$ denotes equality in distribution, i.e., both random variables have the same CDF.

Indeed, for any x ,

$$\mathbb{P}(F_X^{-1}(U) \leq x) = \mathbb{P}(U \leq F_X(x)) = F_X(x),$$

since $U \sim \text{Unif}[0, 1]$. Therefore the CDF of $F_X^{-1}(U)$ is exactly F_X .

Simulation via the inverse CDF: The above converse gives a very useful fundamental sampling principle: to generate a random variable with CDF F_X , it suffices to first generate

$$U \sim \text{Unif}[0, 1]$$

and then set

$$X = F_X^{-1}(U).$$

More generally, even when F_X is not strictly increasing, one may define the generalized inverse (or quantile function) by

$$F_X^{-1}(u) \triangleq \inf\{x \in \mathbb{R} : F_X(x) \geq u\}, \quad 0 < u < 1,$$

and then $F_X^{-1}(U)$ still has distribution F_X . This is the basis of the *inverse transform method*: to sample from a target distribution, first generate a uniform random variable and then pass it through the target quantile function.

Notation:

1. It is often convenient to use the following shorthand:

$$Z \sim F \iff \mathbb{P}(Z \leq z) = F(z),$$

that is, Z has CDF F .

2. Also,

$$Z \stackrel{d}{=} V \iff F_Z = F_V,$$

meaning that Z and V are equal in distribution.

Gaussian random variables: A random variable Y is said to be Gaussian with mean μ and variance σ^2 , also written as $Y \sim \mathcal{N}(\mu, \sigma^2)$ if its PDF is

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right), \quad -\infty < y < \infty.$$

In particular, when $\mu = 0$ and $\sigma^2 = 1$, we obtain the *standard normal* random variable

$$X \sim \mathcal{N}(0, 1),$$

whose density is

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad -\infty < x < \infty.$$

Standard normal CDF and the Q-function: For a standard normal random variable $X \sim \mathcal{N}(0, 1)$, its CDF is denoted by

$$\Phi(x) \triangleq F_X(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du.$$

The corresponding upper-tail probability is denoted by the Q -function:

$$Q(x) \triangleq 1 - \Phi(x) = \mathbb{P}(X > x).$$

Thus, $Q(x)$ represents the area under the standard normal density to the right of x .

Affine transformation of a Gaussian random variable: Let

$$X \sim \mathcal{N}(0, 1), \quad Z = aX + b, \quad a > 0.$$

Using the change-of-variables formula,

$$f_Z(z) = \frac{1}{a} f_X\left(\frac{z-b}{a}\right).$$

Substituting the standard normal density,

$$f_Z(z) = \frac{1}{\sqrt{2\pi}a} \exp\left(-\frac{1}{2}\left(\frac{z-b}{a}\right)^2\right) = \frac{1}{\sqrt{2\pi}a^2} \exp\left(-\frac{(z-b)^2}{2a^2}\right).$$

This is exactly the density of a Gaussian random variable with mean b and variance a^2 . Hence,

$$Z \sim \mathcal{N}(b, a^2).$$

Remark (affine transformations). Let X be a random variable and let $Z = aX + b$, where $a, b \in \mathbb{R}$ are deterministic constants and $a \neq 0$.

- **General density transformation:** If X has density f_X , then Z has density

$$f_Z(z) = \frac{1}{|a|} f_X\left(\frac{z-b}{a}\right).$$

Thus affine transformations shift and rescale distributions as shown above.

- **Closure of common families under affine maps:** Certain distribution families are preserved under affine transformations.

In particular, for the Gaussian family:

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

then

$$aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2).$$

Thus the Gaussian family is closed under affine transformations.

Likewise, for the Uniform family:

$$X \sim \text{Unif}[c, d],$$

then $aX + b$ is uniform on the image of the interval $[c, d]$ under the map $x \mapsto ax + b$. If $a > 0$, then

$$aX + b \sim \text{Unif}[ac + b, ad + b],$$

while if $a < 0$, the same interval appears with the endpoints reversed, i.e.

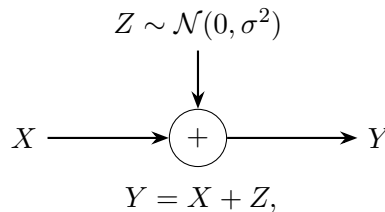
$$aX + b \sim \text{Unif}[ad + b, ac + b].$$

So this closure property is not unique to Gaussians; it also appears in other location-scale families.

Example revisited: binary signaling over an additive Gaussian noise channel

We now return to the basic detection example from the first lecture, in light of our discussion of Gaussian random variables and affine transformations. A more detailed derivation of the MAP rule and its probability-of-error analysis already appears in the Lecture 1 notes, so here we only summarize the key steps.

Consider the additive Gaussian noise model



where

$$X = \begin{cases} +\sqrt{P}, & \text{with probability } \frac{1}{2}, \\ -\sqrt{P}, & \text{with probability } \frac{1}{2}, \end{cases} \quad Z \sim \mathcal{N}(0, \sigma^2),$$

and X and Z are independent.

Since the two possible values of X are equally likely, the MAP rule reduces to the ML rule:

$$\hat{X}(y) \in \arg \max_{x \in \{-\sqrt{P}, \sqrt{P}\}} f_{Y|X}(y|x).$$

Now, conditioned on $X = x$,

$$Y = x + Z.$$

Since $Z \sim \mathcal{N}(0, \sigma^2)$, it follows from the affine transformation property of Gaussian random variables that

$$Y | \{X = x\} \sim \mathcal{N}(x, \sigma^2).$$

Hence

$$f_{Y|X}(y|x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-x)^2}{2\sigma^2}\right).$$

In particular,

$$f_{Y|X}(y|\sqrt{P}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\sqrt{P})^2}{2\sigma^2}\right),$$

and

$$f_{Y|X}(y|-\sqrt{P}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y+\sqrt{P})^2}{2\sigma^2}\right).$$

Therefore,

$$f_{Y|X}(y|\sqrt{P}) > f_{Y|X}(y|-\sqrt{P}) \iff (y-\sqrt{P})^2 < (y+\sqrt{P})^2.$$

Since both sides are nonnegative, this is equivalent to

$$|y - \sqrt{P}| < |y - (-\sqrt{P})|.$$

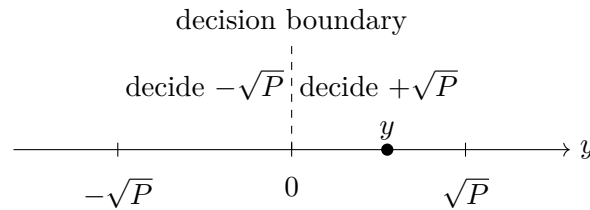
So, the observation y is assigned to the hypothesis whose signal point is *closer* to y on the real line. Thus the detector chooses the nearest of the two candidate points $\pm\sqrt{P}$.

Geometrically, the decision boundary is therefore the midpoint between $-\sqrt{P}$ and $+\sqrt{P}$, that is 0 (as we've already seen earlier). Hence

$$|y - \sqrt{P}| < |y + \sqrt{P}| \iff y > 0.$$

So the detector is

$$\hat{X}(y) = \begin{cases} \sqrt{P}, & y > 0, \\ -\sqrt{P}, & \text{otherwise.} \end{cases}$$



Probability of error: An error occurs when the noise pushes the observation across the origin. Hence

$$P_e = \mathbb{P}(\hat{X}(Y) \neq X) = \frac{1}{2} \mathbb{P}(Y \leq 0 | X = \sqrt{P}) + \frac{1}{2} \mathbb{P}(Y > 0 | X = -\sqrt{P}).$$

Using $Y = X + Z$, this becomes

$$P_e = \frac{1}{2} \mathbb{P}(Z \leq -\sqrt{P}) + \frac{1}{2} \mathbb{P}(Z > \sqrt{P}).$$

By symmetry of the Gaussian distribution,

$$P_e = \mathbb{P}(Z > \sqrt{P}).$$

Since $\frac{Z}{\sigma} \sim \mathcal{N}(0, 1)$,

$$P_e = \mathbb{P}\left(\frac{Z}{\sigma} > \frac{\sqrt{P}}{\sigma}\right) = Q\left(\frac{\sqrt{P}}{\sigma}\right) = Q\left(\sqrt{\frac{P}{\sigma^2}}\right).$$

Here, we can also define the signal-to-noise ratio (SNR) as:

$$\text{SNR} \triangleq \frac{P}{\sigma^2},$$

so equivalently

$$P_e = Q\left(\sqrt{\text{SNR}}\right).$$

Exercise

Find the MAP decoder and its probability of error P_e for the model

$$Y = X + Z, \quad Z \sim \mathcal{N}(0, \sigma^2),$$

with X and Z independent, and

$$X = \begin{cases} \sqrt{P}, & \text{with probability } q, \\ -\sqrt{P}, & \text{with probability } 1 - q. \end{cases}$$

Lecture 3: Expectation, Variance, and MMSE Estimation

Expectation

One of the most basic numerical summaries of a random variable is its *expectation* (or mean), which captures its average value. If X is a discrete random variable with PMF p_X , then

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x p_X(x),$$

provided the sum converges absolutely.

If X is a continuous random variable with PDF f_X , then

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x) dx,$$

provided the integral is finite.

Thus, expectation is computed by averaging the possible values of the random variable, weighted by their probabilities (or density).

Basic properties of expectation

Expectation is especially useful because it behaves very naturally under deterministic transformations and under combinations of random variables.

1) Expectation of a function of one random variable:

Let

$$Y = g(X).$$

Then we may compute $\mathbb{E}[Y]$ directly from the distribution of X , without first finding the distribution of Y .

If X is discrete, then

$$\mathbb{E}[Y] = \mathbb{E}[g(X)] = \sum_{x \in \mathcal{X}} g(x) p_X(x).$$

If X is continuous with density f_X , then

$$\mathbb{E}[Y] = \mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx.$$

This principle is often called the *law of the unconscious statistician* (LOTUS). The name reflects the fact that to compute $\mathbb{E}[g(X)]$, we do *not* need to first find the distribution of the transformed random variable $Y = g(X)$. Instead, we can work directly with the original distribution of X and simply average the values of $g(X)$. In this sense, we evaluate the expectation of the transformed quantity almost “unconsciously,” without explicitly deriving the law of Y . The main purpose of LOTUS is therefore computational: it provides a direct and often much simpler way to compute expectations of functions of random variables.

2) Expectation of a function of multiple random variables:

Now let

$$Z = g(X, Y).$$

If (X, Y) has joint PMF $p_{X,Y}$, then

$$\mathbb{E}[Z] = \mathbb{E}[g(X, Y)] = \sum_x \sum_y g(x, y) p_{X,Y}(x, y).$$

If (X, Y) has joint density $f_{X,Y}$, then

$$\mathbb{E}[Z] = \mathbb{E}[g(X, Y)] = \iint g(x, y) f_{X,Y}(x, y) dx dy.$$

So once the joint distribution is known, expectations of functions of several random variables are computed in exactly the same spirit.

3) Independence implies factorization of expectation:

If X and Y are independent and the relevant expectations exist, then

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y].$$

Indeed, in the continuous case independence means

$$f_{X,Y}(x, y) = f_X(x)f_Y(y),$$

so

$$\mathbb{E}[XY] = \iint xy f_{X,Y}(x, y) dx dy = \iint xy f_X(x)f_Y(y) dx dy.$$

Separating the integrals gives

$$\mathbb{E}[XY] = \left(\int x f_X(x) dx \right) \left(\int y f_Y(y) dy \right) = \mathbb{E}[X]\mathbb{E}[Y].$$

The discrete case is completely analogous.

4) Linearity of expectation:

For any random variables X, Y and any constants a, b ,

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y].$$

More generally, for random variables $X_1, \dots, X_n$ and constants $a_1, \dots, a_n$,

$$\mathbb{E}\left[\sum_{i=1}^n a_i X_i\right] = \sum_{i=1}^n a_i \mathbb{E}[X_i].$$

It is worthwhile to note that *linearity does not require independence*. This makes expectation much easier to work with than many other quantities.

Variance and standard deviation

Expectation measures the center of a random variable, but it does not describe how spread out the variable is around that center. This motivates the variance.

The *variance* of X is defined by

$$\text{Var}(X) \triangleq \mathbb{E}[(X - \mathbb{E}[X])^2].$$

Thus, variance is the expected squared deviation from the mean.

A closely related quantity is the *standard deviation*,

$$\sigma_X \triangleq \sqrt{\text{Var}(X)}.$$

Equivalently,

$$\text{Var}(X) = \sigma_X^2.$$

A useful identity: Expanding the square and using linearity of expectation,

$$\text{Var}(X) = \mathbb{E}[X^2 - 2X\mathbb{E}[X] + (\mathbb{E}[X])^2].$$

Since $\mathbb{E}[X]$ is a constant,

$$\text{Var}(X) = \mathbb{E}[X^2] - 2(\mathbb{E}[X])^2 + (\mathbb{E}[X])^2.$$

Therefore,

$$\boxed{\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2.}$$

This alternate formula is often the most convenient way to compute variance in practice.

Inequalities

In probability, we are often interested in questions such as how likely a random variable is to be unusually large, or how likely it is to deviate significantly from its mean. In other words, we seek to compute probabilities of the form

$$\mathbb{P}(X \geq t), \mathbb{P}(|X - \mathbb{E}[X]| \geq t).$$

If the full distribution of X is known explicitly, then in principle such probabilities can be computed exactly. In many important situations, however, this is not realistic. The distribution may be unknown, difficult to derive, or too complicated to manipulate directly. By contrast, quantities like the mean $\mathbb{E}[X]$ and the variance $\text{Var}(X)$ are often much easier to compute or estimate.

This leads to a natural question: *what can we still say about tail probabilities if we know only partial information about the random variable?* Probabilistic inequalities answer precisely this question. They convert coarse information, such as nonnegativity, the mean, or the variance, into general upper bounds on the probabilities of certain events.

In this sense, inequalities are useful because they:

- provide rigorous bounds even when exact distributions are unknown,

- quantify how unlikely large deviations or rare events can be,
- help show that random quantities concentrate around typical values,
- serve as basic tools in proving broader results such as laws of large numbers and limit theorems.

In this lecture, we begin with two of the most fundamental examples: Markov's inequality and Chebyshev's inequality. These are simple but powerful bounds that require only limited information about the random variable.

Markov's inequality

Markov's inequality is the most basic tail bound. It applies to nonnegative random variables and provides an upper bound on an upper-tail probability using only the mean. In particular, a small mean prevents too much probability mass from lying far out in the upper tail.

Theorem 3.1 (Markov's inequality). *Let X be a nonnegative random variable and let $t > 0$. Then*

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}.$$

Proof. Since $X \geq 0$, we may write

$$X = X \mathbf{1}_{\{X \geq t\}} + X \mathbf{1}_{\{X < t\}}.$$

Taking expectations on both sides gives

$$\mathbb{E}[X] = \mathbb{E}[X \mathbf{1}_{\{X \geq t\}}] + \mathbb{E}[X \mathbf{1}_{\{X < t\}}].$$

Both terms on the right-hand side are nonnegative, so

$$\mathbb{E}[X] \geq \mathbb{E}[X \mathbf{1}_{\{X \geq t\}}].$$

Now on the event $\{X \geq t\}$, we have $X \geq t$, and therefore

$$X \mathbf{1}_{\{X \geq t\}} \geq t \mathbf{1}_{\{X \geq t\}}.$$

Taking expectations once again, we obtain

$$\mathbb{E}[X] \geq t \mathbb{E}[\mathbf{1}_{\{X \geq t\}}] = t \mathbb{P}(X \geq t).$$

Rearranging yields $\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}$. □

Example 3.2 (A crude upper-tail bound). Suppose X denotes the height of a randomly chosen Stanford student, measured in centimeters, and suppose $\mathbb{E}[X] = 165$. Since heights are nonnegative, Markov's inequality gives

$$\mathbb{P}(X \geq 180) \leq \frac{165}{180}.$$

This bound is quite loose in practice, but it illustrates the point: one can obtain a valid tail bound using only nonnegativity and the mean.

Note

Markov's inequality is very general: it assumes only that the random variable is nonnegative and has a finite mean. The tradeoff is that the resulting bound is often conservative.

If $\mathbb{E}[X] > 0$, it is often convenient to set $t = a\mathbb{E}[X]$. Markov's inequality then gives, for every $a > 0$,

$$\mathbb{P}(X \geq a\mathbb{E}[X]) \leq \frac{1}{a}.$$

For example, if $X \geq 0$ and $\mathbb{E}[X] = 35$, then

$$\mathbb{P}(X \geq 70) \leq \frac{1}{2}.$$

So even without knowing the full distribution of X , the mean alone already places a universal upper bound on the probability of very large values.

Chebyshev's inequality

Markov's inequality applies only to nonnegative random variables, but it becomes much more useful when applied to carefully chosen transformations. In particular, since the squared deviation from the mean is always nonnegative, applying Markov's inequality to $(X - \mathbb{E}[X])^2$ yields Chebyshev's inequality.

Theorem 3.3 (Chebyshev's inequality). *Let X be a random variable with mean $\mu = \mathbb{E}[X]$ and variance $\sigma^2 = \text{Var}(X)$. Then for every $t > 0$,*

$$\mathbb{P}(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}.$$

If $\sigma > 0$, this can equivalently be written, for every $b > 0$, as

$$\mathbb{P}(|X - \mu| \geq b\sigma) \leq \frac{1}{b^2}.$$

Proof. Define the nonnegative random variable

$$Y = (X - \mu)^2.$$

Applying Markov's inequality to Y with threshold t^2 , we obtain

$$\mathbb{P}((X - \mu)^2 \geq t^2) \leq \frac{\mathbb{E}[(X - \mu)^2]}{t^2}.$$

But by definition,

$$\mathbb{E}[(X - \mu)^2] = \text{Var}(X) = \sigma^2.$$

Also,

$$\{(X - \mu)^2 \geq t^2\} = \{|X - \mu| \geq t\}.$$

Therefore,

$$\mathbb{P}(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2},$$

as claimed. □

Key takeaway

Chebyshev's inequality is a two-sided concentration bound around the mean. Unlike Markov's inequality, it uses second-moment information through the variance, and therefore gives a stronger statement about how unlikely large deviations from the mean are.

For instance, taking $t = 2\sigma$ gives

$$\mathbb{P}(|X - \mu| \geq 2\sigma) \leq \frac{1}{4},$$

and taking $t = 3\sigma$ gives

$$\mathbb{P}(|X - \mu| \geq 3\sigma) \leq \frac{1}{9}.$$

These bounds are often not tight, but they are remarkably general: they hold for *every* random variable with finite variance, regardless of the underlying distribution.

Covariance and correlation

Expectation and variance describe the mean and spread of a single random variable. When two random variables are present, we also want a way to measure how they vary *together*. This leads us to the notion of covariance.

Definition 3.4. The *covariance* between X and Y is

$$\text{Cov}(X, Y) \triangleq \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])].$$

Expanding the product and using linearity of expectation gives the useful identity

$$\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

Proposition 3.5. If X and Y are independent, then

$$\text{Cov}(X, Y) = 0.$$

Proof. If X and Y are independent, then

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y].$$

Substituting into the identity above yields

$$\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0.$$

□

Note

Independence implies zero covariance, but the converse is not necessarily true: uncorrelated random variables need not be independent.

Since covariance depends on the units of X and Y , it is often convenient to normalize it.

Definition 3.6. Assuming $\sigma_X > 0$ and $\sigma_Y > 0$, the *correlation coefficient* between X and Y is

$$\rho_{X,Y} \triangleq \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}.$$

This dimensionless quantity always satisfies

$$|\rho_{X,Y}| \leq 1.$$

Equality holds *if and only if* $Y = aX + b$ almost surely for some constants a, b with $a \neq 0$. In particular, $\rho_{X,Y} = 1$ corresponds to perfect positive linear dependence, while $\rho_{X,Y} = -1$ corresponds to perfect negative linear dependence.

Conditional expectation

Suppose now that two random variables X and Y are jointly distributed, and that we observe the value of Y . A natural question is: what is the average value of X once this information is known? The *conditional expectation of X given $Y = y$* is denoted by $\mathbb{E}[X|Y = y]$.

For example, in the discrete case,

$$\mathbb{E}[X | Y = y] = \sum_x x p_{X|Y}(x | y),$$

and in the continuous case,

$$\mathbb{E}[X | Y = y] = \int_{-\infty}^{\infty} x f_{X|Y}(x | y) dx.$$

Since the quantity $\mathbb{E}[X | Y = y]$ depends on y , it defines a function of y . If we write

$$g(y) \triangleq \mathbb{E}[X | Y = y],$$

then the *conditional expectation of X given Y* is the random variable

$$\mathbb{E}[X | Y] \triangleq g(Y).$$

Thus, $\mathbb{E}[X | Y]$ is itself a random variable: once Y is observed, it outputs the corresponding conditional mean of X .

Key takeaway

The quantity $\mathbb{E}[X | Y]$ is the best mean-based summary of X after observing Y . It is no longer a number, but a function of the observation Y .

Law of iterated expectations

A fundamental property of conditional expectation is that if we average it once more over the randomness of Y , we recover the ordinary expectation of X .

Theorem 3.7 (Law of iterated expectations (tower property)). *Let X and Y be random variables, and assume that $\mathbb{E}[|X|] < \infty$. Then*

$$\mathbb{E}[\mathbb{E}[X | Y]] = \mathbb{E}[X].$$

Proof. Let

$$g(y) = \mathbb{E}[X | Y = y].$$

Then by definition,

$$\mathbb{E}[X | Y] = g(Y).$$

Hence

$$\mathbb{E}[\mathbb{E}[X | Y]] = \mathbb{E}[g(Y)].$$

In the discrete case,

$$\mathbb{E}[g(Y)] = \sum_y g(y) p_Y(y) = \sum_y \mathbb{E}[X | Y = y] p_Y(y).$$

Substituting the definition of conditional expectation,

$$\mathbb{E}[g(Y)] = \sum_y \left(\sum_x x p_{X|Y}(x | y) \right) p_Y(y).$$

Using

$$p_{X|Y}(x | y) p_Y(y) = p_{X,Y}(x, y),$$

we obtain

$$\mathbb{E}[g(Y)] = \sum_y \sum_x x p_{X,Y}(x, y) = \sum_x x \left(\sum_y p_{X,Y}(x, y) \right).$$

But

$$\sum_y p_{X,Y}(x, y) = p_X(x),$$

so

$$\mathbb{E}[g(Y)] = \sum_x x p_X(x) = \mathbb{E}[X].$$

The continuous case is entirely analogous, with sums replaced by integrals. □

The law of iterated expectations essentially tells us that conditioning on Y and then averaging over Y does not change the overall mean. In other words, conditioning refines the information, but averaging removes that refinement and returns us to the original expectation.

Conditional variance

Just as conditioning leads to a conditional mean, it also leads to a conditional notion of spread.

Definition 3.8. The *conditional variance of X given $Y = y$* is

$$\text{Var}(X | Y = y) \triangleq \mathbb{E}[(X - \mathbb{E}[X | Y = y])^2 | Y = y].$$

Correspondingly, the *conditional variance of X given Y* is the random variable $\text{Var}(X | Y)$.

As with ordinary variance, there is a useful alternative formula:

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

Law of total variance

The ordinary variance of X can be decomposed into two parts: the average uncertainty remaining after observing Y , and the variability of the conditional mean itself.

Theorem 3.9 (Law of total variance). *Also called the law of conditional variances,*

$$\text{Var}(X) = \mathbb{E}[\text{Var}(X | Y)] + \text{Var}(\mathbb{E}[X | Y]).$$

Proof. Write

$$X - \mathbb{E}[X] = (X - \mathbb{E}[X | Y]) + (\mathbb{E}[X | Y] - \mathbb{E}[X]).$$

Squaring both sides,

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

Taking an expectation on both sides gives

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

We now show that the cross term is zero. Using conditional expectation with respect to Y ,

$$\mathbb{E}[(X - \mathbb{E}[X | Y])(\mathbb{E}[X | Y] - \mathbb{E}[X])] = \mathbb{E}[\mathbb{E}[(X - \mathbb{E}[X | Y])(\mathbb{E}[X | Y] - \mathbb{E}[X]) | Y]].$$

Since $\mathbb{E}[X | Y] - \mathbb{E}[X]$ is a function of Y , it can be pulled outside the inner conditional expectation:

$$= \mathbb{E}[(\mathbb{E}[X | Y] - \mathbb{E}[X])\mathbb{E}[X - \mathbb{E}[X | Y] | Y]].$$

But

$$\mathbb{E}[X - \mathbb{E}[X | Y] | Y] = \mathbb{E}[X | Y] - \mathbb{E}[X | Y] = 0,$$

so the cross term vanishes.

Therefore,

$$\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X | Y])^2] + \mathbb{E}[(\mathbb{E}[X | Y] - \mathbb{E}[X])^2].$$

Recognizing both the remaining terms, we have

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

and

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

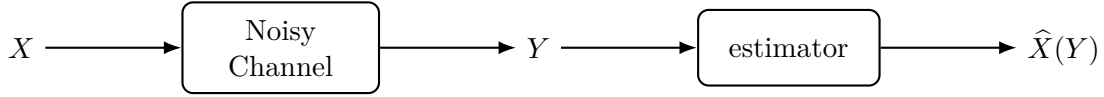
Thus,

$$\text{Var}(X) = \mathbb{E}[\text{Var}(X | Y)] + \text{Var}(\mathbb{E}[X | Y]).$$

□

The law of total variance decomposes total uncertainty into two pieces:

- $\mathbb{E}[\text{Var}(X | Y)]$: the uncertainty that remains after observing Y ,
- $\text{Var}(\mathbb{E}[X | Y])$: the variability explained by the observation Y .



MSE estimation

A common problem in statistical inference is the following: an unknown random variable X is passed through some observation mechanism, producing an observation Y , and from Y we wish to form an estimate of X . An estimator is any function of the observation,

$$\hat{X}(Y),$$

and a commonly used performance criterion is the *mean squared error* (MSE):

$$\text{MSE}(\hat{X}) \triangleq \mathbb{E}[(X - \hat{X}(Y))^2].$$

The goal is to choose the estimator that minimizes this quantity.

Theorem 3.10 (MSE decomposition). *For any estimator $\hat{X}(Y)$,*

$$\mathbb{E}[(X - \hat{X}(Y))^2] = \mathbb{E}[(X - \mathbb{E}[X | Y])^2] + \mathbb{E}[(\mathbb{E}[X | Y] - \hat{X}(Y))^2].$$

Proof. Write

$$X - \hat{X}(Y) = (X - \mathbb{E}[X | Y]) + (\mathbb{E}[X | Y] - \hat{X}(Y)).$$

Squaring and taking expectations,

$$\begin{aligned} \mathbb{E}[(X - \hat{X}(Y))^2] &= \mathbb{E}[(X - \mathbb{E}[X | Y])^2] + \mathbb{E}[(\mathbb{E}[X | Y] - \hat{X}(Y))^2] \\ &\quad + 2\mathbb{E}[(X - \mathbb{E}[X | Y])(\mathbb{E}[X | Y] - \hat{X}(Y))]. \end{aligned}$$

It remains to show that the cross term is zero. Using conditional expectation given Y ,

$$\begin{aligned} &\mathbb{E}[(X - \mathbb{E}[X | Y])(\mathbb{E}[X | Y] - \hat{X}(Y))] \\ &= \mathbb{E}[\mathbb{E}[(X - \mathbb{E}[X | Y])(\mathbb{E}[X | Y] - \hat{X}(Y)) | Y]]. \end{aligned}$$

Since $\mathbb{E}[X | Y] - \hat{X}(Y)$ is a function of Y , it may be taken outside the inner expectation:

$$= \mathbb{E}[(\mathbb{E}[X | Y] - \hat{X}(Y))\mathbb{E}[X - \mathbb{E}[X | Y] | Y]].$$

But

$$\mathbb{E}[X - \mathbb{E}[X | Y] | Y] = \mathbb{E}[X | Y] - \mathbb{E}[X | Y] = 0,$$

so the cross term vanishes. This proves the decomposition. $\square$

Theorem 3.11 (MMSE estimator). *Among all estimators of the form $\hat{X}(Y)$, the one that minimizes the mean squared error is*

$$\hat{X}_{\text{MMSE}}(Y) = \mathbb{E}[X | Y].$$

Proof. By the decomposition above,

$$\mathbb{E}[(X - \hat{X}(Y))^2] = \mathbb{E}[(X - \mathbb{E}[X | Y])^2] + \mathbb{E}[(\mathbb{E}[X | Y] - \hat{X}(Y))^2].$$

The first term does not depend on the choice of estimator, while the second term is always nonnegative. Therefore the minimum mean squared error is achieved precisely when $\hat{X}(Y) = \mathbb{E}[X | Y]$ almost surely. $\square$

The minimum achievable mean squared error is therefore

$$\text{MMSE} = \mathbb{E}[(X - \mathbb{E}[X | Y])^2] = \mathbb{E}[\text{Var}(X | Y)].$$

since,

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

Beyond just being a “useful” average, the conditional expectation is exactly the estimator that minimizes mean squared error. This is why it plays such a central role in estimation theory.

Properties of the MMSE estimator

Recall that the MMSE estimator is

$$\hat{X}(Y) = \mathbb{E}[X | Y].$$

Its estimation error is therefore

$$X - \hat{X}(Y) = X - \mathbb{E}[X | Y].$$

We have already shown that the minimum achievable mean-square error is

$$\text{MMSE} = \mathbb{E}[(X - \mathbb{E}[X | Y])^2].$$

Using the law of total variance,

$$\text{Var}(X) = \mathbb{E}[\text{Var}(X | Y)] + \text{Var}(\mathbb{E}[X | Y]),$$

we may also write

$$\text{MMSE} = \mathbb{E}[\text{Var}(X | Y)] = \text{Var}(X) - \text{Var}(\mathbb{E}[X | Y]).$$

Thus, the reduction in mean-square error obtained from observing Y is exactly $\text{Var}(\mathbb{E}[X | Y])$.

Proposition 3.12 (Marginal unbiasedness). *Let*

$$\hat{X}(Y) = \mathbb{E}[X | Y].$$

Then

$$\mathbb{E}[\hat{X}(Y)] = \mathbb{E}[X].$$

Proof. This is an immediate consequence of the law of iterated expectations:

$$\mathbb{E}[\hat{X}(Y)] = \mathbb{E}[\mathbb{E}[X | Y]] = \mathbb{E}[X].$$

$\square$

Note

The MMSE estimator is unbiased in the *marginal* sense, meaning that its overall average equals the average of X . This is weaker than saying that the estimator is unbiased pointwise for each value of Y .

Theorem 3.13 (Orthogonality principle). *Let*

$$\hat{X}(Y) = \mathbb{E}[X | Y].$$

Then for every square-integrable function $g(Y)$,

$$\mathbb{E}[(X - \hat{X}(Y))g(Y)] = 0.$$

Proof. Using conditional expectation with respect to Y ,

$$\mathbb{E}[(X - \hat{X}(Y))g(Y)] = \mathbb{E}[\mathbb{E}[(X - \hat{X}(Y))g(Y) | Y]].$$

Since $g(Y)$ is a function of Y , it can be pulled outside the inner conditional expectation:

$$= \mathbb{E}[g(Y) \mathbb{E}[X - \hat{X}(Y) | Y]].$$

But $\hat{X}(Y) = \mathbb{E}[X | Y]$, so

$$\mathbb{E}[X - \hat{X}(Y) | Y] = \mathbb{E}[X | Y] - \mathbb{E}[X | Y] = 0.$$

Hence

$$\mathbb{E}[(X - \hat{X}(Y))g(Y)] = 0.$$

□

A particularly important special case is obtained by taking $g(Y) = 1$, which gives

$$\mathbb{E}[X - \hat{X}(Y)] = 0,$$

and therefore again

$$\mathbb{E}[\hat{X}(Y)] = \mathbb{E}[X].$$

Key takeaway

The orthogonality principle says that the MMSE error is orthogonal to every random variable of the form $g(Y)$, i.e., to every candidate estimator built from the observation Y .

Geometric interpretation

The orthogonality property suggests a geometric viewpoint. Consider the collection of square-integrable random variables as an inner-product space, with:

$$\langle U, V \rangle \triangleq \mathbb{E}[UV].$$

This induces the norm

$$\|U\| \triangleq \sqrt{\mathbb{E}[U^2]},$$

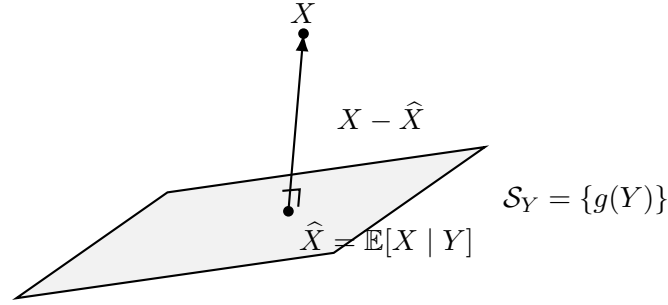


Figure 1: Geometric interpretation of MMSE estimation: $\hat{X} = \mathbb{E}[X | Y]$ is the orthogonal projection of X onto the subspace of all square-integrable functions of Y .

and hence the distance

$$d(U, V) \triangleq \|U - V\| = \sqrt{\mathbb{E}[(U - V)^2]}.$$

Now let

$$\mathcal{S}_Y \triangleq \{g(Y) : \mathbb{E}[g(Y)^2] < \infty\},$$

the subspace of all square-integrable random variables that can be expressed as functions of Y .

Since the mean-square error of an estimator $g(Y)$ is

$$\mathbb{E}[(X - g(Y))^2] = \|X - g(Y)\|^2,$$

the MMSE problem is exactly the problem of finding the element of $\mathcal{S}_Y$ closest to X .

The orthogonality principle shows that

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

is precisely the *orthogonal projection* of X onto the subspace $\mathcal{S}_Y$. Equivalently, we may decompose

$$X = \mathbb{E}[X | Y] + (X - \mathbb{E}[X | Y]),$$

where

$$\mathbb{E}[X | Y] \in \mathcal{S}_Y,$$

and the error term

$$X - \mathbb{E}[X | Y]$$

is orthogonal to all of $\mathcal{S}_Y$.

This is the estimation-theoretic analogue of projecting a vector in Euclidean space onto a subspace.

Note

This geometric picture is extremely useful. It explains both why conditional expectation is optimal under squared loss and why the estimation error is orthogonal to every function of the observation.

Lecture 4: Inner-Product Geometry, Linear Estimation, and Jointly Gaussian Random Variables

In the previous lecture, we introduced conditional expectation and saw that

$$\hat{X}_{\text{MMSE}}(Y) = \mathbb{E}[X | Y]$$

minimizes the mean squared error among all estimators that are functions of the observation Y . In this part of the lecture, we begin developing the geometric viewpoint behind MMSE estimation. The main idea is to view random variables as elements of an inner product space, so that estimation becomes a projection problem. The key idea is that estimation can be interpreted as a *projection problem* in an inner product space. Once we build the right geometric language—norm, distance, orthogonality, and projection—the form of the MMSE estimator becomes conceptually natural.

Inner Product Spaces

We begin with some linear algebra background.

Definition 4.1 (Vector space). A *vector space* $\mathcal{V}$ over $\mathbb{R}$ is a set equipped with vector addition and scalar multiplication satisfying the usual vector-space axioms. In particular, it is closed under these two operations:

1. if $v_1, v_2 \in \mathcal{V}$, then $v_1 + v_2 \in \mathcal{V}$;
2. if $v \in \mathcal{V}$ and $\alpha \in \mathbb{R}$, then $\alpha v \in \mathcal{V}$.

Thus a vector space is simply a set that is *closed under addition and scalar multiplication*.

Definition 4.2 (Inner product). An *inner product* on a real vector space $\mathcal{V}$ is a map

$$\langle \cdot, \cdot \rangle : \mathcal{V} \times \mathcal{V} \rightarrow \mathbb{R}$$

satisfying, for all $u, v, w \in \mathcal{V}$ and $\alpha \in \mathbb{R}$,

1. **symmetry:** $\langle u, v \rangle = \langle v, u \rangle$,
2. **linearity:** $\langle \alpha u + v, w \rangle = \alpha \langle u, w \rangle + \langle v, w \rangle$,
3. **positive definiteness:** $\langle u, u \rangle \geq 0$, with equality *iff* $u = 0$.

A vector space equipped with an inner product is called an *inner product space*.

Example: In Euclidean space $\mathbb{R}^n$, the standard inner product is the dot product

$$\langle u, v \rangle = u^\top v = \sum_{i=1}^n u_i v_i.$$

Norm, distance, and angle

Once an inner product is defined, it induces a notion of length.

Definition 4.3 (Norm). The *norm* of a vector $u \in \mathcal{V}$ is defined by

$$\|u\| \triangleq \sqrt{\langle u, u \rangle}.$$

Geometrically, $\|u\|$ plays the role of the length or magnitude of u .

Likewise, the distance between two vectors u and v is measured by the norm of their difference:

$$d(u, v) = \|u - v\|.$$

In Euclidean geometry, the inner product also determines angles. If $u \neq 0$ and $v \neq 0$, then

$$\langle u, v \rangle = \|u\| \|v\| \cos \phi,$$

where ϕ is the angle between u and v . Hence

$$\phi = \cos^{-1} \left(\frac{\langle u, v \rangle}{\|u\| \|v\|} \right).$$

So the inner product tells us not only how long vectors are, but encodes information about their alignment.

Orthogonality

Definition 4.4 (Orthogonality). Two vectors $u, v \in \mathcal{V}$ are said to be *orthogonal*, written $u \perp v$, if

$$\langle u, v \rangle = 0.$$

This generalizes the familiar notion of perpendicular vectors in Euclidean space.

We also need the notion of orthogonality to an entire set.

Definition 4.5 (Orthogonality to a subspace). Let $\mathcal{S} \subseteq \mathcal{V}$ be a subspace. We say that u is orthogonal to $\mathcal{S}$, written

$$u \perp \mathcal{S},$$

if

$$\langle u, v \rangle = 0 \text{ for every } v \in \mathcal{S}.$$

Thus being orthogonal to a subspace means being orthogonal to every vector in that subspace.

Projection onto a subspace

Suppose now that $\mathcal{S} \subseteq \mathcal{V}$ is a subspace and we are given a vector $v \in \mathcal{V}$. A natural question that has been studied extensively in this context is: *Which element of $\mathcal{S}$ is closest to v ?*

This is exactly the geometric problem behind estimation.

Theorem 4.6 (Projection principle). *Let $\mathcal{S}$ be a subspace of an inner product space $\mathcal{V}$, and let $v \in \mathcal{V}$. Suppose there exists $v^* \in \mathcal{S}$ such that*

$$\|v - v^*\| = \min_{u \in \mathcal{S}} \|v - u\|.$$

Then v^ is characterized by the orthogonality condition*

$$v - v^* \perp \mathcal{S}.$$

Conversely, if $v^ \in \mathcal{S}$ satisfies*

$$v - v^* \perp \mathcal{S},$$

then v^ is the unique minimizer of $\|v - u\|$ over $u \in \mathcal{S}$.*

Proof. Assume first that $v^* \in \mathcal{S}$ is the closest point to v . Take any $w \in \mathcal{S}$. Since $\mathcal{S}$ is a subspace, for any scalar $t \in \mathbb{R}$,

$$v^* + tw \in \mathcal{S}.$$

Because v^* minimizes distance to v , the function

$$\phi(t) \triangleq \|v - (v^* + tw)\|^2$$

is minimized at $t = 0$. Expanding,

$$\phi(t) = \|(v - v^*) - tw\|^2 = \|v - v^*\|^2 - 2t\langle v - v^*, w \rangle + t^2\|w\|^2.$$

For this quadratic to be minimized at $t = 0$, its linear term must vanish, so

$$\langle v - v^*, w \rangle = 0.$$

Since this holds for every $w \in \mathcal{S}$, we conclude that

$$v - v^* \perp \mathcal{S}.$$

Conversely, suppose $v^* \in \mathcal{S}$ and $v - v^* \perp \mathcal{S}$. Let any $u \in \mathcal{S}$ be given. Then $u - v^* \in \mathcal{S}$, so

$$\langle v - v^*, u - v^* \rangle = 0.$$

Hence

$$\begin{aligned} \|v - u\|^2 &= \|(v - v^*) + (v^* - u)\|^2 \\ &= \|v - v^*\|^2 + \|u - v^*\|^2 + 2\langle v - v^*, v^* - u \rangle \\ &= \|v - v^*\|^2 + \|u - v^*\|^2 \\ &\geq \|v - v^*\|^2. \end{aligned}$$

Therefore v^* indeed minimizes the distance from v to $\mathcal{S}$. Moreover, equality can hold only if $\|u - v^*\|^2 = 0$, i.e. only if $u = v^*$. Hence the minimizer is unique. $\square$

Takeaway: The closest point in the subspace is obtained by dropping a perpendicular from v onto $\mathcal{S}$. The approximation error $v - v^*$ is orthogonal to every direction available inside the subspace.

The space $L_2(\Omega, \mathcal{F}, P)$

We now move from ordinary vectors to random variables.

Let

$$L_2(\Omega, \mathcal{F}, P) = \left\{ [U] : U : \Omega \rightarrow \mathbb{R}, \mathbb{E}[U^2] < \infty \right\},$$

where two random variables are identified if they are equal almost surely. We continue to write U for the resulting equivalence class.

With this convention, $U = 0$ in L_2 precisely when $U = 0$ almost surely. Therefore,

$$\langle U, V \rangle \triangleq \mathbb{E}[UV]$$

defines an inner product on $L_2(\Omega, \mathcal{F}, P)$.

Indeed:

1. $\mathbb{E}[UV] = \mathbb{E}[VU]$, so symmetry holds;
2. $\mathbb{E}[(\alpha U + V)W] = \alpha \mathbb{E}[UW] + \mathbb{E}[VW]$, so linearity holds;
3. $\mathbb{E}[U^2] \geq 0$, with equality iff $U = 0$ almost surely.

Thus L_2 is an inner product space. Under this inner product,

$$\|U\| = \sqrt{\mathbb{E}[U^2]},$$

and the distance between U and V is

$$\|U - V\| = \sqrt{\mathbb{E}[(U - V)^2]}.$$

So in L_2 , *mean squared error is literally squared distance*.

This is the central observation behind MMSE estimation.

A geometric interpretation of estimation

Suppose $X \in L_2(\Omega, \mathcal{F}, P)$ is the random quantity we wish to estimate from an observation Y .

An estimator based on Y is any random variable of the form $g(Y)$, where g is a measurable function. Hence the class of all estimators obtainable from Y is

$$\mathcal{S} = \{g(Y) : \mathbb{E}[g(Y)^2] < \infty\}.$$

This set is a subspace of $L_2(\Omega, \mathcal{F}, P)$: sums of functions of Y are still functions of Y , and scalar multiples remain functions of Y .

Now consider the optimization problem

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

Since

$$\mathbb{E}[(X - g(Y))^2] = \|X - g(Y)\|^2,$$

this is precisely the problem of finding the point in the subspace $\mathcal{S}$ closest to X .

Therefore, by the projection principle, the optimal estimator $g^*(Y)$ is characterized by

$$X - g^*(Y) \perp \mathcal{S}.$$

That is,

$$\mathbb{E}[(X - g^*(Y))h(Y)] = 0 \text{ for every square-integrable function } h(Y).$$

Conditional expectation as orthogonal projection

We now identify the optimal projection.

Theorem 4.7 (Conditional expectation as projection). *Among all square-integrable estimators of the form $g(Y)$, the unique minimizer of*

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

is

$$g^*(Y) = \mathbb{E}[X | Y].$$

Equivalently, $\mathbb{E}[X | Y]$ is the orthogonal projection of X onto the subspace

$$\mathcal{S} = \{g(Y) : \mathbb{E}[g(Y)^2] < \infty\}.$$

Proof. We already know from MMSE estimation that $\mathbb{E}[X | Y]$ minimizes mean squared error among all functions of Y . To connect this with geometry, let $h(Y) \in \mathcal{S}$ be arbitrary. Then

$$\begin{aligned} \mathbb{E}[(X - \mathbb{E}[X | Y])h(Y)] &= \mathbb{E}[\mathbb{E}[(X - \mathbb{E}[X | Y])h(Y) | Y]] \\ &= \mathbb{E}[h(Y) \mathbb{E}[X - \mathbb{E}[X | Y] | Y]] \\ &= \mathbb{E}[h(Y)(\mathbb{E}[X | Y] - \mathbb{E}[X | Y])] \\ &= 0. \end{aligned}$$

Thus

$$X - \mathbb{E}[X | Y] \perp \mathcal{S}.$$

By the projection principle, $\mathbb{E}[X | Y]$ is exactly the orthogonal projection of X onto $\mathcal{S}$, and hence the unique minimizer. $\square$

Why this matters. This theorem gives a powerful conceptual interpretation of conditional expectation. It is not merely a probabilistic average; it is the *best approximation to X using only information contained in Y* in the mean-square sense.

Orthogonality principle

The previous theorem is often summarized as the *orthogonality principle*:

The MMSE estimation error is orthogonal to every function of the observation. That is, if

$$\hat{X}(Y) = \mathbb{E}[X | Y],$$

then

$$\mathbb{E}[(X - \hat{X}(Y))h(Y)] = 0 \text{ for all } h(Y) \in L_2.$$

This is the infinite-dimensional analogue of the orthogonality condition that appeared in linear estimation. There, one projects onto the one-dimensional subspace spanned by $Y - \mathbb{E}[Y]$; here, one projects onto the full subspace of all functions of Y .

Scalar Linear Estimation

The full MMSE estimator allows *any* square-integrable function of the observation:

$$\hat{X}_{\text{MMSE}}(Y) = \mathbb{E}[X | Y].$$

Geometrically, this is the orthogonal projection of X onto the space of all functions of Y .

In many practical settings, however, we do not know the full joint distribution of (X, Y) . Instead, we may only know the first and second moments of X and Y , namely $\mathbb{E}[X]$, $\mathbb{E}[Y]$, $\text{Var}(X)$, $\text{Var}(Y)$, $\text{Cov}(X, Y)$.

These quantities are not enough in general to determine $\mathbb{E}[X | Y]$, but they are enough to determine the *best affine estimator* of X from Y .¹

Scalar linear estimation

We now restrict attention to estimators of the form

$$\hat{X} = aY + b,$$

and seek the coefficients a, b that minimize

$$\mathbb{E}[(X - \hat{X})^2] = \mathbb{E}[(X - aY - b)^2].$$

This is often called *linear estimation*, although technically the estimator is affine because of the offset term b .

Geometric interpretation

The class of admissible estimators is now much smaller than before:

$$\mathcal{A} = \{aY + b : a, b \in \mathbb{R}\} = \text{span}\{1, Y\}.$$

Thus $\mathcal{A}$ is a linear subspace of L_2 whose elements are affine functions of Y . It is two-dimensional whenever $\text{Var}(Y) > 0$. Instead of projecting X onto the space of all functions of Y , we project it onto $\mathcal{A}$. Let

$$\hat{X}_{\text{lin}} = a^*Y + b^*$$

denote the optimal affine estimator. By the projection principle, the estimation error $X - \hat{X}_{\text{lin}}$ must be orthogonal to every element of $\mathcal{A}$.

Equivalently, it must be orthogonal to both basis directions 1 and Y . Thus

$$\mathbb{E}[X - (a^*Y + b^*)] = 0$$

and

$$\mathbb{E}[(X - (a^*Y + b^*))Y] = 0.$$

These are the normal equations for scalar linear estimation.

Solving for the optimal coefficients: Expanding the two orthogonality conditions gives

$$\mathbb{E}[X] = a^*\mathbb{E}[Y] + b^*$$

and

$$\mathbb{E}[XY] = a^*\mathbb{E}[Y^2] + b^*\mathbb{E}[Y].$$

Substituting the first equation into the second eliminates b^* :

$$\mathbb{E}[XY] = a^*\mathbb{E}[Y^2] + (\mathbb{E}[X] - a^*\mathbb{E}[Y])\mathbb{E}[Y].$$

Rearranging,

$$\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = a^*(\mathbb{E}[Y^2] - (\mathbb{E}[Y])^2).$$

¹The notes use affine/linear interchangeably.

Therefore,

$$a^* = \frac{\text{Cov}(X, Y)}{\text{Var}(Y)},$$

provided $\text{Var}(Y) > 0$.

Then

$$b^* = \mathbb{E}[X] - a^* \mathbb{E}[Y] = \mathbb{E}[X] - \frac{\text{Cov}(X, Y)}{\text{Var}(Y)} \mathbb{E}[Y].$$

So the optimal affine estimator can be written as

$$\hat{X}_{\text{lin}} = \frac{\text{Cov}(X, Y)}{\text{Var}(Y)} (Y - \mathbb{E}[Y]) + \mathbb{E}[X].$$

In the special case $\mathbb{E}[X] = \mathbb{E}[Y] = 0$, this reduces to

$$\hat{X}_{\text{lin}} = \frac{\text{Cov}(X, Y)}{\text{Var}(Y)} Y.$$

Minimum mean squared error in linear estimation

We now compute the minimum mean squared error achieved by this affine estimator.

Since

$$\hat{X}_{\text{lin}} - \mathbb{E}[X] = \frac{\text{Cov}(X, Y)}{\text{Var}(Y)} (Y - \mathbb{E}[Y]),$$

its variance is

$$\text{Var}(\hat{X}_{\text{lin}}) = \left(\frac{\text{Cov}(X, Y)}{\text{Var}(Y)} \right)^2 \text{Var}(Y) = \frac{\text{Cov}(X, Y)^2}{\text{Var}(Y)}.$$

Because $X - \hat{X}_{\text{lin}}$ is orthogonal to $\hat{X}_{\text{lin}} - \mathbb{E}[X]$, we obtain

$$\text{Var}(X) = \mathbb{E}[(X - \hat{X}_{\text{lin}})^2] + \text{Var}(\hat{X}_{\text{lin}}).$$

Hence

$$\mathbb{E}[(X - \hat{X}_{\text{lin}})^2] = \text{Var}(X) - \frac{\text{Cov}(X, Y)^2}{\text{Var}(Y)}.$$

Using the correlation coefficient

$$\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y},$$

this becomes

$$\mathbb{E}[(X - \hat{X}_{\text{lin}})^2] = \text{Var}(X)(1 - \rho_{X,Y}^2).$$

So the quality of the best linear estimator is governed entirely by how strongly X and Y are correlated.

Interpretation:

Several useful conclusions follow immediately:

- If $\rho_{X,Y} = 0$, then

$$\hat{X}_{\text{lin}} = \mathbb{E}[X].$$

In this case, the best linear estimator ignores the observation Y entirely.

- If $|\rho_{X,Y}| = 1$, then the linear estimator is perfect, and

$$\mathbb{E}[(X - \hat{X}_{\text{lin}})^2] = 0.$$

This happens exactly when $X - \mathbb{E}[X]$ is a linear function of $Y - \mathbb{E}[Y]$.

- In general, the linear estimator uses only first and second moments. The full MMSE estimator, by contrast, exploits the entire conditional distribution.

Thus the linear estimator is often easier to compute, but it need not equal the true MMSE estimator.

Linear versus (nonlinear) MMSE estimation

The geometric picture now becomes very clear.

- The full MMSE estimator projects X onto the space of *all* functions of Y .
- The linear estimator projects X only onto the much smaller linear subspace of affine functions of Y .

Because the linear subspace of affine functions of Y is much smaller, the optimal linear estimator is generally not as good as the full MMSE estimator.

Example: Let

$$Y \sim \text{Unif}[-1, 1], \quad X = Y^2.$$

Since X is exactly determined by Y , the full MMSE estimator is perfect:

$$\mathbb{E}[X | Y] = \mathbb{E}[Y^2 | Y] = Y^2 = X.$$

Therefore,

$$\mathbb{E}[(X - \mathbb{E}[X | Y])^2] = 0.$$

Now let us compute the optimal linear estimator.

First,

$$\mathbb{E}[Y] = 0.$$

Next,

$$\mathbb{E}[X] = \mathbb{E}[Y^2] = \int_{-1}^1 y^2 \cdot \frac{1}{2} dy = \frac{1}{3}.$$

Also,

$$\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \mathbb{E}[Y^3] - \frac{1}{3} \cdot 0 = 0,$$

since Y^3 is an odd function and Y is symmetric about 0.

Hence

$$\hat{X}_{\text{lin}} = \frac{\text{Cov}(X, Y)}{\text{Var}(Y)}(Y - \mathbb{E}[Y]) + \mathbb{E}[X] = \frac{1}{3}.$$

So the best linear estimator is just the constant

$$\boxed{\hat{X}_{\text{lin}} = \mathbb{E}[X] = \frac{1}{3}.$$

Its mean squared error is

$$\mathbb{E}[(X - \hat{X}_{\text{lin}})^2] = \text{Var}(X).$$

We can compute this explicitly:

$$\text{Var}(X) = \text{Var}(Y^2) = \mathbb{E}[Y^4] - (\mathbb{E}[Y^2])^2.$$

Now

$$\mathbb{E}[Y^4] = \int_{-1}^1 y^4 \cdot \frac{1}{2} dy = \frac{1}{5},$$

so

$$\text{Var}(X) = \frac{1}{5} - \left(\frac{1}{3}\right)^2 = \frac{4}{45}.$$

Thus

$$\boxed{\mathbb{E}[(X - \mathbb{E}[X | Y])^2] = 0, \text{ while } \mathbb{E}[(X - \hat{X}_{\text{lin}})^2] = \frac{4}{45} > 0.}$$

This example shows that even if X is completely determined by Y , the best affine estimator may still be strictly suboptimal when the dependence is nonlinear.

Jointly Gaussian random variables

The previous example showed that, in general, the optimal linear estimator need not coincide with the full MMSE estimator. There is, however, an extremely important class of random variables for which this gap disappears: the *jointly Gaussian* case.

Intuitively, jointly Gaussian random variables are completely described by their first and second moments. As a result, conditioning preserves a very simple form, and the conditional mean turns out to be affine in the observation.

Definition 4.8 (Jointly Gaussian random variables). Two random variables X and Y are said to be *jointly Gaussian* if

$$aX + bY$$

is Gaussian for every $a, b \in \mathbb{R}$.

In the nondegenerate case, where $\sigma_X > 0$, $\sigma_Y > 0$, and $|\rho_{X,Y}| < 1$, this is equivalent to requiring the joint density to have the form

$$f_{X,Y}(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho_{X,Y}^2}} \exp\left(-\frac{1}{2(1-\rho_{X,Y}^2)} \left[\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - 2\rho_{X,Y} \frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} \right]\right).$$

Degenerate jointly Gaussian pairs are included in the definition above, but they do not possess a two-dimensional density of this form.

A jointly Gaussian pair is specified by its mean vector and covariance matrix. In the nondegenerate bivariate case, these can equivalently be parametrized by μ_X , μ_Y , σ_X^2 , σ_Y^2 , and $\rho_{X,Y}$.

Let

$$\mu = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_X^2 & \rho_{X,Y} \sigma_X \sigma_Y \\ \rho_{X,Y} \sigma_X \sigma_Y & \sigma_Y^2 \end{pmatrix}.$$

Then

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2} \begin{pmatrix} x - \mu_X \\ y - \mu_Y \end{pmatrix}^\top \Sigma^{-1} \begin{pmatrix} x - \mu_X \\ y - \mu_Y \end{pmatrix}\right).$$

Basic properties

Jointly Gaussian random variables have several key properties that make them especially important in estimation theory.

1. The parameters μ_X , μ_Y , σ_X^2 , σ_Y^2 , $\rho_{X,Y}$ completely determine the joint distribution.
2. The marginals are Gaussian:

$$X \sim \mathcal{N}(\mu_X, \sigma_X^2), \quad Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2).$$

3. The conditional distribution is also Gaussian. More precisely, we have:

$$X \mid \{Y = y\} \sim \mathcal{N}\left(\mu_X + \rho_{X,Y} \frac{\sigma_X}{\sigma_Y} (y - \mu_Y), (1 - \rho_{X,Y}^2) \sigma_X^2\right).$$

Note the third property: it shows that conditioning on $Y = y$ shifts the mean of X by an amount that depends *linearly* on y , while the conditional variance is a constant independent of y .

Consequence for MMSE estimation

Since the conditional distribution of X given $Y = y$ is Gaussian with mean

$$\mu_X + \rho_{X,Y} \frac{\sigma_X}{\sigma_Y} (y - \mu_Y),$$

the conditional expectation is

$$\mathbb{E}[X \mid Y = y] = \mu_X + \rho_{X,Y} \frac{\sigma_X}{\sigma_Y} (y - \mu_Y).$$

Hence, as a random variable,

$$\mathbb{E}[X \mid Y] = \mu_X + \rho_{X,Y} \frac{\sigma_X}{\sigma_Y} (Y - \mu_Y).$$

Using

$$\rho_{X,Y} \frac{\sigma_X}{\sigma_Y} = \frac{\text{Cov}(X, Y)}{\text{Var}(Y)},$$

we can rewrite this as

$$\mathbb{E}[X | Y] = \mathbb{E}[X] + \frac{\text{Cov}(X, Y)}{\text{Var}(Y)}(Y - \mathbb{E}[Y]).$$

But this is exactly the optimal affine estimator derived earlier:

$$\hat{X}_{\text{lin}} = \mathbb{E}[X] + \frac{\text{Cov}(X, Y)}{\text{Var}(Y)}(Y - \mathbb{E}[Y]).$$

Therefore, in the jointly Gaussian case,

$$\hat{X}_{\text{MMSE}}(Y) = \mathbb{E}[X | Y] = \hat{X}_{\text{lin}}.$$

So for jointly Gaussian random variables, the best linear estimator is not merely the best among affine estimators; it is in fact the *true* MMSE estimator.

Minimum mean squared error

Because the conditional variance is

$$\text{Var}(X | Y = y) = (1 - \rho_{X,Y}^2)\sigma_X^2,$$

independent of y , we immediately get

$$\mathbb{E}[\text{Var}(X | Y)] = (1 - \rho_{X,Y}^2)\sigma_X^2.$$

Hence the MMSE is

$$\mathbb{E}[(X - \mathbb{E}[X | Y])^2] = (1 - \rho_{X,Y}^2)\sigma_X^2.$$

This is exactly the same expression we obtained earlier for the optimal linear estimator. Again, this reflects the special fact that, in the jointly Gaussian setting, linear estimation is already optimal.

So although nonlinear MMSE estimation can be difficult in general, the jointly Gaussian case reduces the problem to the linear-estimation formulas we already derived.

Summary

- $L_2(\Omega, \mathcal{F}, P)$ with inner product

$$\langle U, V \rangle = \mathbb{E}[UV]$$

gives a geometric view of estimation: mean squared error is squared distance.

- The MMSE estimator is the orthogonal projection of X onto the space of all square-integrable functions of Y :

$$\hat{X}_{\text{MMSE}}(Y) = \mathbb{E}[X | Y].$$

- Restricting to estimators of the form $\hat{X} = aY + b$, the optimal linear/affine estimator is

$$\hat{X}_{\text{lin}} = \mathbb{E}[X] + \frac{\text{Cov}(X, Y)}{\text{Var}(Y)}(Y - \mathbb{E}[Y]).$$

- Its mean squared error is

$$\mathbb{E}[(X - \hat{X}_{\text{lin}})^2] = \text{Var}(X) - \frac{\text{Cov}(X, Y)^2}{\text{Var}(Y)} = \text{Var}(X)(1 - \rho_{X,Y}^2).$$

- In general,

$$\hat{X}_{\text{lin}} \neq \mathbb{E}[X | Y],$$

so the best linear estimator can be worse than the full MMSE estimator.

- If (X, Y) are jointly Gaussian, then

$$X | \{Y = y\} \sim \mathcal{N}\left(\mu_X + \rho_{X,Y} \frac{\sigma_X}{\sigma_Y} (y - \mu_Y), (1 - \rho_{X,Y}^2) \sigma_X^2\right),$$

and therefore

$$\mathbb{E}[X | Y] = \mathbb{E}[X] + \frac{\text{Cov}(X, Y)}{\text{Var}(Y)} (Y - \mathbb{E}[Y]).$$

- Hence, for jointly Gaussian random variables,

$$\hat{X}_{\text{MMSE}}(Y) = \hat{X}_{\text{lin}},$$

so linear estimation is optimal.

Lecture 5: Random Vectors, Gaussian Random Vectors

In the previous lectures, we focused mainly on one random variable, or at most a pair of random variables. That was already enough to study detection, conditional expectation, MMSE estimation, and the scalar jointly Gaussian case. In many applications, however, the quantities of interest are naturally multi-dimensional: we may wish to model several random quantities jointly, estimate several unknowns from several observations, or study how a collection of variables transforms under deterministic linear operations.

We study *random vectors*, their joint distributions, characteristic functions, and covariance matrices, *Gaussian random vectors*, whose behavior is well known under linear transformations and close the lecture with *whitening* and *coloring*.

Random Vectors

Let $X_1, X_2, \dots, X_n$ be random variables defined on the same probability space. We collect them into a single object:

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}.$$

This is called a *random vector*. Unless stated otherwise, we will adopt the convention that random vectors are column vectors.

Definition 5.1 (Random Vector). A random vector is a finite collection of random variables arranged as a vector:

$$\mathbf{X} = (X_1, \dots, X_n)^\top.$$

Just as a scalar random variable is completely specified by its distribution, a random vector is completely specified by its joint distribution.

Joint CDF

The most basic way to describe the distribution of a random vector is through its joint cumulative distribution function.

Definition 5.2 (Joint CDF). Let

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}.$$

Its *joint cumulative distribution function* is

$$F_{\mathbf{X}}(\mathbf{x}) \triangleq \mathbb{P}(X_1 \leq x_1, X_2 \leq x_2, \dots, X_n \leq x_n),$$

where

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n.$$

Thus $F_{\mathbf{X}}$ records the probability that each component of the random vector falls below the corresponding coordinate threshold. As in the scalar case, the joint CDF always exists and completely characterizes the distribution. If the random vector is continuous, one may also work with a joint pdf; if it is discrete, one may work with a joint pmf. The joint CDF, however, is the most general distributional description. Marginal and conditional distributions are obtained from the joint distribution in the usual way, exactly as in the scalar and two-variable settings.

Characteristic Function of a Random Vector

Another useful way to describe the distribution of a random vector is through its characteristic function.

Definition 5.3 (Characteristic function). The *characteristic function* of the random vector $\mathbf{X}$ is defined by

$$\phi_{\mathbf{X}}(\boldsymbol{\omega}) \triangleq \mathbb{E}\left[e^{i\boldsymbol{\omega}^\top \mathbf{X}}\right], \quad \boldsymbol{\omega} \in \mathbb{R}^n.$$

Here $\boldsymbol{\omega}$ is a deterministic vector in $\mathbb{R}^n$, and $\boldsymbol{\omega}^\top \mathbf{X}$ is a scalar random variable. Therefore $e^{i\boldsymbol{\omega}^\top \mathbf{X}}$ is a complex-valued random variable, and its expectation is well defined.

The characteristic function completely determines the distribution of $\mathbf{X}$, just as in the scalar case. So there is a one-to-one correspondence between the joint distribution and the characteristic function.

Remark 5.4. When a scalar random variable X has pdf f_X , its characteristic function is

$$\phi_X(\omega) = \mathbb{E}[e^{i\omega X}] = \int_{-\infty}^{\infty} e^{i\omega x} f_X(x) dx,$$

which is essentially the Fourier transform of the pdf. Likewise, for a random vector, the characteristic function may be viewed as the multi-dimensional Fourier transform of the joint distribution.

Besides uniquely describing the distribution, characteristic functions have several useful structural properties.

Proposition 5.5 (Basic properties of characteristic functions). *Let $\mathbf{X}$ and $\mathbf{Y}$ be random vectors.*

1. **Uniqueness.** *The characteristic function $\phi_{\mathbf{X}}$ uniquely determines the distribution of $\mathbf{X}$.*
2. **Moments.** *If the relevant derivatives exist, then moments can be recovered from derivatives of $\phi_{\mathbf{X}}$ at the origin. For example,*

$$\mathbb{E}[X_j^k] = \frac{1}{i^k} \frac{\partial^k \phi_{\mathbf{X}}(\mathbf{0})}{\partial \omega_j^k}.$$

3. **Independence.** *If $\mathbf{X}$ and $\mathbf{Y}$ are independent, then*

$$\phi_{\mathbf{X}+\mathbf{Y}}(\boldsymbol{\omega}) = \phi_{\mathbf{X}}(\boldsymbol{\omega}) \phi_{\mathbf{Y}}(\boldsymbol{\omega}).$$

The third property is especially important: characteristic functions turn sums of independent random vectors into products, which is one of the main reasons they are so useful in probability theory.

Mean vector

For a random vector, expectation is defined componentwise.

Definition 5.6 (Mean Vector). Let

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}.$$

Its *mean vector* is

$$\mathbb{E}[\mathbf{X}] = \begin{bmatrix} \mathbb{E}[X_1] \\ \mathbb{E}[X_2] \\ \vdots \\ \mathbb{E}[X_n] \end{bmatrix}.$$

So, the expectation of a random vector is again a vector, obtained by taking the expectation of each component separately.

Covariance Matrix

The covariance matrix captures the second-order structure of a random vector.

Definition 5.7 (Covariance matrix). The *covariance matrix* of a random vector $\mathbf{X}$ is the $n \times n$ matrix

$$\Sigma_{\mathbf{X}} = [\Sigma_{\mathbf{X}}(i, j)]_{i, j=1}^n,$$

whose $(i, j)^{th}$ entry is

$$\Sigma_{\mathbf{X}}(i, j) = \text{Cov}(X_i, X_j).$$

Equivalently,

$$\Sigma_{\mathbf{X}} = \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top].$$

So, $(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top$ is an $n \times n$ random matrix, whose $(i, j)^{th}$ entry is $(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])$. Taking expectations entrywise gives exactly $\text{Cov}(X_i, X_j)$.

Basic properties of covariance matrices

Covariance matrices satisfy two fundamental structural properties.

Proposition 5.8. For any random vector $\mathbf{X}$, its covariance matrix $\Sigma_{\mathbf{X}}$ satisfies:

1. $\Sigma_{\mathbf{X}}$ is symmetric;
2. $\Sigma_{\mathbf{X}}$ is positive semidefinite, i.e.,

$$\mathbf{a}^\top \Sigma_{\mathbf{X}} \mathbf{a} \geq 0 \text{ for every } \mathbf{a} \in \mathbb{R}^n.$$

Proof. Symmetry follows immediately from the symmetry of covariance:

$$\Sigma_{\mathbf{X}}(i, j) = \text{Cov}(X_i, X_j) = \text{Cov}(X_j, X_i) = \Sigma_{\mathbf{X}}(j, i).$$

To prove positive semidefiniteness, let $\mathbf{a} \in \mathbb{R}^n$ be arbitrary. Using

$$\Sigma_{\mathbf{X}} = \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top],$$

we compute

$$\begin{aligned} \mathbf{a}^\top \Sigma_{\mathbf{X}} \mathbf{a} &= \mathbf{a}^\top \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top] \mathbf{a} \\ &= \mathbb{E}[\mathbf{a}^\top (\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top \mathbf{a}] \\ &= \mathbb{E}[(\mathbf{a}^\top (\mathbf{X} - \mathbb{E}[\mathbf{X}]))^2] \\ &\geq 0. \end{aligned}$$

Since the expectation of a square is nonnegative, $\Sigma_{\mathbf{X}}$ is positive semidefinite. $\square$

Remark 5.9. The positive semidefinite property is not merely algebraic. It reflects the fact that every linear combination of the components of $\mathbf{X}$ has nonnegative variance:

$$\text{Var}(\mathbf{a}^\top \mathbf{X}) = \mathbf{a}^\top \Sigma_{\mathbf{X}} \mathbf{a}.$$

Covariance under Linear Transformations

A natural question is: how does the covariance matrix change under a deterministic linear transformation?

Let $\mathbf{Y} = A\mathbf{X}$, where $\mathbf{X} \in \mathbb{R}^n$ is a random vector and A is a deterministic $m \times n$ matrix. Then $\mathbf{Y} \in \mathbb{R}^m$ is another random vector. By linearity of expectation,

$$\mathbb{E}[\mathbf{Y}] = \mathbb{E}[A\mathbf{X}] = A\mathbb{E}[\mathbf{X}].$$

The covariance matrix transforms in the expected quadratic way.

Proposition 5.10 (Covariance of a linear transformation). *If $\mathbf{Y} = A\mathbf{X}$, then $\Sigma_{\mathbf{Y}} = A\Sigma_{\mathbf{X}}A^\top$.*

Proof. Starting from the definition,

$$\Sigma_{\mathbf{Y}} = \mathbb{E}[(\mathbf{Y} - \mathbb{E}[\mathbf{Y}])(\mathbf{Y} - \mathbb{E}[\mathbf{Y}])^\top].$$

Since $\mathbf{Y} = A\mathbf{X}$ and $\mathbb{E}[\mathbf{Y}] = A\mathbb{E}[\mathbf{X}]$,

$$\begin{aligned} \Sigma_{\mathbf{Y}} &= \mathbb{E}[(A\mathbf{X} - A\mathbb{E}[\mathbf{X}])(A\mathbf{X} - A\mathbb{E}[\mathbf{X}])^\top] \\ &= \mathbb{E}[A(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top A^\top]. \end{aligned}$$

Because A and $A^\top$ are deterministic, they can be pulled outside the expectation:

$$\begin{aligned} \Sigma_{\mathbf{Y}} &= A \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top] A^\top \\ &= A\Sigma_{\mathbf{X}}A^\top. \end{aligned}$$

$\square$

This formula is one of the most important identities involving covariance matrices. It tells us exactly how second-order structure behaves under deterministic linear maps.

Gaussian Random Vectors

We now turn to one of the most important families of random vectors. There are several equivalent ways to define Gaussian random vectors. For our purposes, the cleanest and most general definition is via the characteristic function.

Definition 5.11 (Gaussian Random Vector). A random vector

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}$$

is called a *Gaussian random vector* if its characteristic function has the form

$$\phi_{\mathbf{X}}(\boldsymbol{\omega}) = \exp\left(-\frac{1}{2}\boldsymbol{\omega}^\top \Sigma \boldsymbol{\omega} + i\boldsymbol{\omega}^\top \boldsymbol{\mu}\right), \quad \boldsymbol{\omega} \in \mathbb{R}^n,$$

for some vector $\boldsymbol{\mu} \in \mathbb{R}^n$ and some symmetric positive semidefinite matrix $\Sigma \in \mathbb{R}^{n \times n}$.

In that case, $\boldsymbol{\mu}$ is the mean vector and Σ is the covariance matrix, and we write $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$. Equivalently, one often says that the components $X_1, \dots, X_n$ are *jointly Gaussian*.

Remark 5.12. This definition is especially useful because it continues to make sense even when Σ is singular. Thus it includes degenerate Gaussian random vectors as well.

Equivalent PDF form

When Σ is invertible, the characteristic-function definition is equivalent to the familiar multivariate Gaussian density.

Fact 5.13. Suppose Σ is invertible. Then $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ if and only if $\mathbf{X}$ has density

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{n/2} \det(\Sigma)^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\right).$$

So when Σ^{-1} exists, the characteristic-function definition agrees with the usual multivariate normal pdf.

Remark 5.14. The characteristic-function definition is slightly more general. If Σ is singular, a joint pdf with respect to Lebesgue measure may fail to exist, but the Gaussian random vector is still perfectly well defined through its characteristic function.

Degenerate Gaussians

A useful illustration of the previous remark is that deterministic constants are naturally included in the Gaussian family.

Example 5.15 (A constant random variable). Suppose $Y = c$ with probability one, where $c \in \mathbb{R}$ is deterministic. Then

$$\phi_Y(\omega) = \mathbb{E}[e^{i\omega Y}] = e^{i\omega c}.$$

But this may be written as

$$e^{i\omega c} = \exp\left(-\frac{1}{2}\omega \cdot 0 \cdot \omega + i\omega c\right).$$

Therefore

$$Y \sim \mathcal{N}(c, 0).$$

So even a deterministic constant can be regarded as Gaussian, with mean equal to that constant and variance 0.

Affine Transformations of Gaussian Random Vectors

A fundamental closure property of Gaussian random vectors is that deterministic affine transformations preserve Gaussianity.

Proposition 5.16. *Let $\mathbf{X} \sim \mathcal{N}(\mu, \Sigma)$, and let $\mathbf{Y} = A\mathbf{X} + b$, where A is a deterministic matrix and b is a deterministic vector of compatible dimensions. Then $\mathbf{Y}$ is also Gaussian, and*

$$\mathbf{Y} \sim \mathcal{N}(A\mu + b, A\Sigma A^\top).$$

Proof. We compute the characteristic function of $\mathbf{Y}$:

$$\phi_{\mathbf{Y}}(\omega) = \mathbb{E}[e^{i\omega^\top \mathbf{Y}}] = \mathbb{E}[e^{i\omega^\top (A\mathbf{X} + b)}] = \mathbb{E}[e^{i(A^\top \omega)^\top \mathbf{X}} e^{i\omega^\top b}] = e^{i\omega^\top b} \mathbb{E}[e^{i(A^\top \omega)^\top \mathbf{X}}] = e^{i\omega^\top b} \phi_{\mathbf{X}}(A^\top \omega).$$

Since $\mathbf{X} \sim \mathcal{N}(\mu, \Sigma)$,

$$\phi_{\mathbf{X}}(u) = \exp\left(-\frac{1}{2}u^\top \Sigma u + iu^\top \mu\right).$$

Substituting $u = A^\top \omega$, we obtain

$$\begin{aligned} \phi_{\mathbf{Y}}(\omega) &= e^{i\omega^\top b} \exp\left(-\frac{1}{2}(A^\top \omega)^\top \Sigma (A^\top \omega) + i(A^\top \omega)^\top \mu\right) \\ &= \exp\left(-\frac{1}{2}\omega^\top A\Sigma A^\top \omega + i\omega^\top (A\mu + b)\right). \end{aligned}$$

This is exactly the characteristic function of a Gaussian random vector with mean $A\mu + b$ and covariance $A\Sigma A^\top$. Therefore

$$\mathbf{Y} \sim \mathcal{N}(A\mu + b, A\Sigma A^\top).$$

□

Corollary 5.17 (Marginals of a Gaussian random vector are Gaussian). *Any subvector of a Gaussian random vector is again Gaussian. In particular, if*

$$\mathbf{Y} = \begin{bmatrix} X_{i_1} \\ X_{i_2} \\ \vdots \\ X_{i_k} \end{bmatrix}$$

is formed by selecting a subset of the components of

$$\mathbf{X} \sim \mathcal{N}(\mu, \Sigma),$$

then $\mathbf{Y}$ is Gaussian, with mean vector given by the corresponding entries of μ and covariance matrix given by the corresponding submatrix of Σ .

Proof. This is a special case of the previous proposition: take A to be the coordinate-selection matrix that extracts the desired components, and take $b = 0$. $\square$

There is another important characterization of Gaussian random vectors: a random vector is Gaussian **if and only if** every linear combination of its components is a Gaussian random variable.

Theorem 5.18. *A random vector*

$$\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$$

is Gaussian if and only if

$$v^\top \mathbf{X} = \sum_{i=1}^n v_i X_i$$

is a Gaussian random variable for every deterministic vector $v \in \mathbb{R}^n$.

Proof. The “only if” direction follows immediately from the previous proposition: if $\mathbf{X}$ is Gaussian, then every affine transformation of $\mathbf{X}$ is Gaussian, and in particular every scalar linear combination $v^\top \mathbf{X}$ is Gaussian.

For the “if” direction, assume that $\omega^\top \mathbf{X}$ is a Gaussian random variable for every $\omega \in \mathbb{R}^n$.

First, if $Z \sim \mathcal{N}(\mu, \sigma^2)$, then

$$\mathbb{E}[e^{iZ}] = \phi_Z(1) = \exp\left(-\frac{1}{2}\sigma^2 + i\mu\right).$$

Then, we also have: $\text{Var}(\omega^\top \mathbf{X}) = \omega^\top \Sigma_{\mathbf{X}} \omega$ and $\mathbb{E}[\omega^\top \mathbf{X}] = \omega^\top \mathbb{E}[\mathbf{X}] = \omega^\top \mu_{\mathbf{X}}$, where $\mu_{\mathbf{X}} = \mathbb{E}[\mathbf{X}]$.

Now compute the characteristic function of $\mathbf{X}$:

$$\phi_{\mathbf{X}}(\omega) = \mathbb{E}\left[e^{i\omega^\top \mathbf{X}}\right].$$

Since $\omega^\top \mathbf{X}$ is Gaussian, with mean $\omega^\top \mu_{\mathbf{X}}$ and variance $\omega^\top \Sigma_{\mathbf{X}} \omega$, we obtain

$$\phi_{\mathbf{X}}(\omega) = \exp\left(-\frac{1}{2}\omega^\top \Sigma_{\mathbf{X}} \omega + i\omega^\top \mu_{\mathbf{X}}\right).$$

This is exactly the characteristic function of a Gaussian random vector. Therefore

$$\mathbf{X} \sim \mathcal{N}(\mu_{\mathbf{X}}, \Sigma_{\mathbf{X}}).$$

$\square$

Remark 5.19. For Gaussian random vectors, uncorrelated components are independent. Equivalently, if $\mathbf{X} \sim \mathcal{N}(\mu, \Sigma)$ and Σ is diagonal, then $X_1, \dots, X_n$ are independent Gaussian random variables. This implication is special to the Gaussian family and does not hold for general random vectors.

Whitening of Random Vectors

We now return to general random vectors and introduce whitening.

Definition 5.20 (White random vector). A random vector $\mathbf{X}$ is said to be *white* if its covariance matrix has the form

$$\Sigma_{\mathbf{X}} = \sigma^2 I$$

for some $\sigma^2 > 0$.

Thus a white random vector has uncorrelated components, all with the same variance. In particular, the common special case

$$\Sigma_{\mathbf{X}} = I$$

corresponds to a white vector with unit component variance.

Definition 5.21 (Whitening transform). A deterministic matrix A is called a *whitening transform* for $\mathbf{X}$ if $\mathbf{Y} = A\mathbf{X}$ is white.

Remark 5.22. This terminology has origins in signal processing. Roughly speaking, whiteness corresponds to a flat spectrum, though that interpretation will be discussed more later in the course.

Square-Root Factors of a Covariance Matrix

Definition 5.23 (Square-root factor). For a covariance matrix Σ , a matrix B is called a *square-root factor* of Σ if

$$BB^{\top} = \Sigma.$$

Square-root factors exist for every covariance matrix, but they are generally not unique.

Coloring

Suppose $\mathbf{X}$ is white with

$$\Sigma_{\mathbf{X}} = I,$$

and let

$$\mathbf{Y} = B\mathbf{X},$$

where $BB^{\top} = \Sigma$. Then

$$\Sigma_{\mathbf{Y}} = B\Sigma_{\mathbf{X}}B^{\top} = BB^{\top} = \Sigma.$$

Thus B colors a white random vector to have covariance Σ .

Whitening

Conversely, suppose $\mathbf{Y}$ has covariance $\Sigma = BB^{\top}$ and B is invertible. Define

$$\mathbf{X} = B^{-1}\mathbf{Y}.$$

Then

$$\Sigma_{\mathbf{X}} = B^{-1}\Sigma(B^{-1})^{\top} = I,$$

so $\mathbf{X}$ is white.

Thus the two operations are

$$\text{coloring: } \mathbf{X} \mapsto B\mathbf{X},$$

$$\text{whitening: } \mathbf{Y} \mapsto B^{-1}\mathbf{Y}.$$

Finding $\Sigma^{1/2}$ when Σ is invertible:

Building on the above discussion, suppose Σ^{-1} exists. Since Σ is a covariance matrix, it is symmetric and positive semidefinite. Invertibility strengthens this to positive definiteness. Therefore:

1. Σ is positive definite;
2. Σ has positive eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_n > 0$;
3. Σ has an orthonormal basis of eigenvectors $u_1, u_2, \dots, u_n$.

These satisfy

$$\Sigma u_i = \lambda_i u_i, \quad i = 1, \dots, n,$$

and the orthonormality relations

$$u_i^\top u_j = 0 \text{ for } i \neq j, \quad \|u_i\| = 1.$$

Equivalently, if

$$U = \begin{bmatrix} u_1 & u_2 & \cdots & u_n \end{bmatrix}, \quad \Lambda = \text{diag}(\lambda_1, \dots, \lambda_n),$$

then

$$\Sigma = U \Lambda U^\top.$$

Since each $\lambda_i > 0$, we may define

$$\Lambda^{1/2} = \text{diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_n}),$$

and then a square root of Σ is given by

$$\Sigma^{1/2} = U \Lambda^{1/2} U^\top.$$

Geometrically, the transformation $U \Lambda^{1/2} U^\top$ rotates into the eigenbasis of Σ , scales each orthogonal direction by $\sqrt{\lambda_i}$, and then rotates back.

So,

$$\Sigma^{1/2}(\Sigma^{1/2})^\top = U \Lambda^{1/2} U^\top U \Lambda^{1/2} U^\top = U \Lambda U^\top = \Sigma.$$

Likewise,

$$\Sigma^{-1/2} = U \Lambda^{-1/2} U^\top,$$

where

$$\Lambda^{-1/2} = \text{diag}\left(\frac{1}{\sqrt{\lambda_1}}, \dots, \frac{1}{\sqrt{\lambda_n}}\right).$$

Summary

- A random vector is a finite collection of random variables stacked into a vector. Its distribution is completely specified by its joint CDF, and equivalently by its characteristic function.
- The characteristic function

$$\phi_{\mathbf{X}}(\omega) = \mathbb{E}[e^{i\omega^\top \mathbf{X}}]$$

uniquely determines the distribution. When moments exist, they can be recovered from derivatives at the origin, and independence turns sums into products:

$$\phi_{\mathbf{X}+\mathbf{Y}} = \phi_{\mathbf{X}}\phi_{\mathbf{Y}}.$$

- The mean of a random vector is the vector of componentwise expectations, and the covariance matrix is

$$\Sigma_{\mathbf{X}} = \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top].$$

- Every covariance matrix is symmetric and positive semidefinite. If $\mathbf{Y} = A\mathbf{X}$, then

$$\Sigma_{\mathbf{Y}} = A\Sigma_{\mathbf{X}}A^\top.$$

- A Gaussian random vector is defined through the quadratic-exponential form of its characteristic function:

$$\phi_{\mathbf{X}}(\omega) = \exp\left(-\frac{1}{2}\omega^\top \Sigma \omega + i\omega^\top \mu\right).$$

When Σ is invertible, this is equivalent to the usual multivariate Gaussian density.

- The characteristic-function definition naturally includes degenerate cases, such as deterministic constants, which are Gaussian with zero variance.
- Gaussian random vectors are closed under deterministic affine transformations: if

$$\mathbf{Y} = A\mathbf{X} + b,$$

then

$$\mathbf{Y} \sim \mathcal{N}(A\mu + b, A\Sigma A^\top).$$

- A random vector is Gaussian if and only if every linear combination of its components is a Gaussian random variable. In particular, every subvector (hence every marginal) of a Gaussian random vector is again Gaussian, and for Gaussian random vectors, uncorrelated components are independent.
- A random vector is *white* if its covariance matrix is of the form $\sigma^2 I$. A whitening transform is a linear map that converts a given random vector into a white one.
- If $\mathbf{X}$ is white with covariance I , then $\mathbf{Y} = \Sigma^{1/2}\mathbf{X}$ has covariance Σ ; this is called coloring. Conversely, if $\mathbf{Y}$ has covariance Σ , then $\mathbf{X} = \Sigma^{-1/2}\mathbf{Y}$ has covariance I ; this is whitening.
- When Σ is invertible, spectral decomposition gives

$$\Sigma = U\Lambda U^\top, \quad \Sigma^{1/2} = U\Lambda^{1/2}U^\top, \quad \Sigma^{-1/2} = U\Lambda^{-1/2}U^\top.$$

Lecture 6: Covariance Matrices and Linear MMSE Estimation

In the previous lecture, we introduced covariance matrices, Gaussian random vectors, and the ideas of whitening and coloring. We now make those constructions explicit by showing how to build a square-root factor of a covariance matrix from its eigendecomposition. This will then lead naturally to the matrix form of linear MMSE estimation from multiple observations.

Brief recap

- Let

$$\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$$

be a random vector. Its covariance matrix is

$$\Sigma_{\mathbf{X}} \triangleq \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top].$$

- If $\mathbf{Y} = A\mathbf{X}$, where A is deterministic, then $\Sigma_{\mathbf{Y}} = A\Sigma_{\mathbf{X}}A^\top$.
- A matrix B is called a square-root factor of Σ if $BB^\top = \Sigma$.
- If $\mathbf{X}$ is white with $\Sigma_{\mathbf{X}} = I$, then $\mathbf{Y} = B\mathbf{X}$ has covariance $\Sigma_{\mathbf{Y}} = BB^\top = \Sigma$, so B colors a white vector. Conversely, if $\mathbf{Y}$ has covariance Σ and B is invertible with $BB^\top = \Sigma$, then $\mathbf{X} = B^{-1}\mathbf{Y}$ is white since

$$\Sigma_{\mathbf{X}} = B^{-1}\Sigma(B^{-1})^\top = I.$$

Finding a square-root factor of Σ

Assume Σ^{-1} exists. Since Σ is a covariance matrix, it is symmetric and positive semidefinite; invertibility strengthens this to positive definiteness. Hence, by the spectral theorem, Σ has positive eigenvalues

$$\lambda_1, \dots, \lambda_n > 0$$

and an orthonormal basis of eigenvectors

$$u_1, \dots, u_n$$

satisfying

$$\Sigma u_i = \lambda_i u_i, \quad u_i^\top u_j = 0 \quad (i \neq j), \quad \|u_i\| = 1.$$

Let

$$U = \begin{bmatrix} u_1 & \cdots & u_n \end{bmatrix}, \quad \Lambda = \text{diag}(\lambda_1, \dots, \lambda_n).$$

Then $U^\top U = U U^\top = I$ and

$$\Sigma = U \Lambda U^\top.$$

Define

$$\Lambda^{1/2} = \text{diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_n}), \quad B \triangleq U \Lambda^{1/2}.$$

Then

$$BB^\top = (U \Lambda^{1/2})(U \Lambda^{1/2})^\top = U \Lambda U^\top = \Sigma.$$

Thus

$$B = U\Lambda^{1/2}, \quad BB^\top = \Sigma.$$

Since B is invertible,

$$B^{-1} = (U\Lambda^{1/2})^{-1} = \Lambda^{-1/2}U^\top,$$

where

$$\Lambda^{-1/2} = \text{diag}\left(\frac{1}{\sqrt{\lambda_1}}, \dots, \frac{1}{\sqrt{\lambda_n}}\right).$$

Hence, if $\mathbf{Y}$ has covariance Σ , then

$$\mathbf{X} = B^{-1}\mathbf{Y} = \Lambda^{-1/2}U^\top\mathbf{Y}$$

is white.

Remark 6.1. Square-root factors are not unique. More generally, if V is any unitary matrix (in the real case, orthogonal), then

$$\Sigma = BB^\top = BVV^\top B^\top = (BV)(BV)^\top,$$

so BV is also a square-root factor of Σ . In particular, if $\Sigma^{1/2}$ is any square-root factor, then $\Sigma^{1/2}V$ is also a square-root factor.

A particularly important distinguished choice is the symmetric positive definite square root

$$\Sigma^{1/2} = U\Lambda^{1/2}U^\top,$$

which satisfies

$$\Sigma^{1/2}\Sigma^{1/2} = \Sigma.$$

Coloring and Whitening revisited

The eigendecomposition gives an explicit recipe for both operations.

Coloring. If $\mathbf{X}$ is white with $\Sigma_{\mathbf{X}} = I$, define $\mathbf{Y} = B\mathbf{X} = U\Lambda^{1/2}\mathbf{X}$. Then $\Sigma_{\mathbf{Y}} = BB^\top = \Sigma$.

Whitening. If $\mathbf{Y}$ has covariance Σ , define $\mathbf{X} = B^{-1}\mathbf{Y} = \Lambda^{-1/2}U^\top\mathbf{Y}$. Then, $\Sigma_{\mathbf{X}} = I$.

Thus whitening and coloring amount to rotating into the eigenbasis and scaling each orthogonal direction appropriately.

Exercise 6.2. Suppose $\mathbb{E}[\mathbf{X}] = 0$ and $\det(\Sigma_{\mathbf{X}}) = 0$. Show that there exist $i \in \{1, \dots, n\}$ and constants α_j , $1 \leq j \leq n$, $j \neq i$, such that

$$X_i = \sum_{\substack{j=1 \\ j \neq i}}^n \alpha_j X_j \text{ with probability one.}$$

Linear MMSE Estimation

In principle, the minimum MSE estimator of X given $\mathbf{Y}$ is the conditional expectation $\mathbb{E}[X | \mathbf{Y}]$. However, this estimator may be difficult to compute in practice, either because the full joint distribution of X and $\mathbf{Y}$ is unknown or because evaluating the conditional expectation is computationally expensive. This motivates restricting attention to linear estimators, which depend only on means, variances, and covariances.

We now turn to the problem of estimating a scalar random variable X from observations $Y_1, \dots, Y_n$. We first assume

$$\mathbb{E}[X] = \mathbb{E}[Y_1] = \dots = \mathbb{E}[Y_n] = 0.$$

Consider linear estimators of the form

$$\hat{X} = \sum_{j=1}^n h_j Y_j.$$

If we define

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \quad \mathbf{h} = \begin{bmatrix} h_1 \\ \vdots \\ h_n \end{bmatrix},$$

then

$$\hat{X} = \mathbf{h}^\top \mathbf{Y}.$$

Geometrically, we are restricting attention to the subspace spanned by $Y_1, \dots, Y_n$, and the optimal linear estimator is the projection of X onto that subspace. By the projection principle, the estimation error must be orthogonal to every basis element:

$$\mathbb{E}[(X - \hat{X})Y_i] = 0, \quad i = 1, \dots, n.$$

Substituting $\hat{X} = \sum_{j=1}^n h_j Y_j$, we get

$$\mathbb{E} \left[\left(X - \sum_{j=1}^n h_j Y_j \right) Y_i \right] = 0, \quad i = 1, \dots, n.$$

Since all variables are zero mean, this becomes

$$\text{Cov}(X, Y_i) = \sum_{j=1}^n h_j \text{Cov}(Y_j, Y_i), \quad i = 1, \dots, n.$$

Now define the covariance matrix of $\mathbf{Y}$ and the cross-covariance vector between $\mathbf{Y}$ and X by

$$\Sigma_{\mathbf{Y}} = \begin{bmatrix} \text{Cov}(Y_1, Y_1) & \cdots & \text{Cov}(Y_1, Y_n) \\ \vdots & \ddots & \vdots \\ \text{Cov}(Y_n, Y_1) & \cdots & \text{Cov}(Y_n, Y_n) \end{bmatrix}, \quad \Sigma_{\mathbf{Y}X} = \begin{bmatrix} \text{Cov}(Y_1, X) \\ \vdots \\ \text{Cov}(Y_n, X) \end{bmatrix}.$$

Then the orthogonality conditions can be written compactly as

$$\Sigma_{\mathbf{Y}X} = \Sigma_{\mathbf{Y}} \mathbf{h}.$$

Assuming $\Sigma_{\mathbf{Y}}$ is invertible,

$$\mathbf{h} = \Sigma_{\mathbf{Y}}^{-1} \Sigma_{\mathbf{Y}X}.$$

Hence the optimal linear MSE estimator is

$$\hat{X} = (\Sigma_{\mathbf{Y}}^{-1} \Sigma_{\mathbf{Y}X})^{\top} \mathbf{Y} = \Sigma_{\mathbf{Y}X}^{\top} \Sigma_{\mathbf{Y}}^{-1} \mathbf{Y}.$$

General case: If the variables are not necessarily zero mean, we first center them and then apply the same formula. Thus

$$\hat{X} = \mathbb{E}[X] + \Sigma_{\mathbf{Y}X}^{\top} \Sigma_{\mathbf{Y}}^{-1} (\mathbf{Y} - \mathbb{E}[\mathbf{Y}]).$$

Resulting Minimum MSE:

Still in the zero-mean case, orthogonality gives

$$\mathbb{E}[(X - \hat{X})\hat{X}] = 0,$$

so

$$\text{Var}(X) = \text{Var}(\hat{X}) + \text{Var}(X - \hat{X}).$$

Therefore

$$\mathbb{E}[(X - \hat{X})^2] = \text{Var}(X) - \text{Var}(\hat{X}).$$

Using $\hat{X} = \Sigma_{\mathbf{Y}X}^{\top} \Sigma_{\mathbf{Y}}^{-1} \mathbf{Y}$, we obtain

$$\begin{aligned} \text{Var}(\hat{X}) &= \mathbb{E}[\hat{X}^2] \\ &= \mathbb{E}\left[(\Sigma_{\mathbf{Y}X}^{\top} \Sigma_{\mathbf{Y}}^{-1} \mathbf{Y})(\mathbf{Y}^{\top} \Sigma_{\mathbf{Y}}^{-1} \Sigma_{\mathbf{Y}X})\right] \\ &= \Sigma_{\mathbf{Y}X}^{\top} \Sigma_{\mathbf{Y}}^{-1} \mathbb{E}[\mathbf{Y} \mathbf{Y}^{\top}] \Sigma_{\mathbf{Y}}^{-1} \Sigma_{\mathbf{Y}X} \\ &= \Sigma_{\mathbf{Y}X}^{\top} \Sigma_{\mathbf{Y}}^{-1} \Sigma_{\mathbf{Y}} \Sigma_{\mathbf{Y}}^{-1} \Sigma_{\mathbf{Y}X} \\ &= \Sigma_{\mathbf{Y}X}^{\top} \Sigma_{\mathbf{Y}}^{-1} \Sigma_{\mathbf{Y}X}. \end{aligned}$$

Hence

$$\mathbb{E}[(X - \hat{X})^2] = \text{Var}(X) - \Sigma_{\mathbf{Y}X}^{\top} \Sigma_{\mathbf{Y}}^{-1} \Sigma_{\mathbf{Y}X}.$$

Remark 6.3 (Geometric interpretation). Geometrically, $\hat{X}$ is the orthogonal projection of X onto the subspace spanned by $Y_1, \dots, Y_n$. The error $X - \hat{X}$ is orthogonal to that subspace, which is exactly why

$$\mathbb{E}[(X - \hat{X})Y_i] = 0, \quad i = 1, \dots, n.$$

Vector-valued Estimation

The same idea extends when X is itself a random vector. Assume still that the relevant variables are zero mean, and let

$$\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_m \end{bmatrix}, \quad \mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}.$$

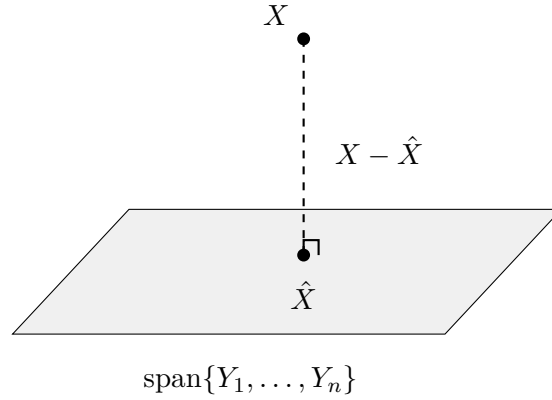


Figure 2: Geometric interpretation of linear MMSE estimation: $\hat{X}$ is the orthogonal projection of X onto $\text{span}\{Y_1, \dots, Y_n\}$, so $X - \hat{X}$ is perpendicular to that subspace.

For each component X_i , the optimal linear estimator based on $\mathbf{Y}$ is

$$\hat{X}_i = \Sigma_{\mathbf{Y}X_i}^\top \Sigma_{\mathbf{Y}}^{-1} \mathbf{Y}.$$

Stacking these estimators gives

$$\hat{\mathbf{X}} = \begin{bmatrix} \hat{X}_1 \\ \vdots \\ \hat{X}_m \end{bmatrix} = \begin{bmatrix} \Sigma_{\mathbf{Y}X_1}^\top \Sigma_{\mathbf{Y}}^{-1} \mathbf{Y} \\ \vdots \\ \Sigma_{\mathbf{Y}X_m}^\top \Sigma_{\mathbf{Y}}^{-1} \mathbf{Y} \end{bmatrix}.$$

If we define the cross-covariance matrix

$$\Sigma_{\mathbf{YX}} \triangleq \mathbb{E}[\mathbf{YX}^\top],$$

then

$$\Sigma_{\mathbf{XY}} = \Sigma_{\mathbf{YX}}^\top = \mathbb{E}[\mathbf{XY}^\top],$$

and the estimator can be written compactly as

$$\hat{\mathbf{X}} = \Sigma_{\mathbf{XY}} \Sigma_{\mathbf{Y}}^{-1} \mathbf{Y}.$$

Error covariance matrix

Define the estimation error vector

$$\mathcal{E} = \mathbf{X} - \hat{\mathbf{X}}.$$

Its covariance matrix is

$$\Sigma_{\mathcal{E}} = \mathbb{E}[(\mathbf{X} - \hat{\mathbf{X}})(\mathbf{X} - \hat{\mathbf{X}})^\top].$$

Using

$$\hat{\mathbf{X}} = \Sigma_{\mathbf{XY}} \Sigma_{\mathbf{Y}}^{-1} \mathbf{Y},$$

we obtain

$$\begin{aligned} \Sigma_{\mathcal{E}} &= \mathbb{E}[(\mathbf{X} - \Sigma_{\mathbf{XY}} \Sigma_{\mathbf{Y}}^{-1} \mathbf{Y})(\mathbf{X} - \Sigma_{\mathbf{XY}} \Sigma_{\mathbf{Y}}^{-1} \mathbf{Y})^\top] \\ &= \Sigma_{\mathbf{X}} - \Sigma_{\mathbf{XY}} \Sigma_{\mathbf{Y}}^{-1} \Sigma_{\mathbf{YX}} - \Sigma_{\mathbf{XY}} \Sigma_{\mathbf{Y}}^{-1} \Sigma_{\mathbf{YX}} + \Sigma_{\mathbf{XY}} \Sigma_{\mathbf{Y}}^{-1} \Sigma_{\mathbf{Y}} \Sigma_{\mathbf{Y}}^{-1} \Sigma_{\mathbf{YX}} \\ &= \Sigma_{\mathbf{X}} - \Sigma_{\mathbf{XY}} \Sigma_{\mathbf{Y}}^{-1} \Sigma_{\mathbf{YX}}. \end{aligned}$$

Therefore

$$\Sigma_{\mathcal{E}} = \Sigma_{\mathbf{X}} - \Sigma_{\mathbf{X}\mathbf{Y}}\Sigma_{\mathbf{Y}}^{-1}\Sigma_{\mathbf{Y}\mathbf{X}}.$$

Remark 6.4. This is the matrix analogue of the scalar formula

$$\mathbb{E}[(X - \hat{X})^2] = \text{Var}(X) - \Sigma_{\mathbf{Y}X}^{\top}\Sigma_{\mathbf{Y}}^{-1}\Sigma_{\mathbf{Y}X}.$$

Summary

1. We made the geometry of covariance matrices explicit using the eigendecomposition $\Sigma = U\Lambda U^{\top}$. This gives a concrete way to construct square-root factors of a covariance matrix, and hence gives procedures for coloring a white random vector and whitening a correlated random vector. The eigenvectors give the principal directions of variation, while the eigenvalues determine the amount of scaling along each direction.
2. We introduced linear MMSE estimation. Instead of computing the full conditional expectation $\mathbb{E}[X | \mathbf{Y}]$, we restricted attention to estimators that are linear in the observations. The optimal linear estimator is the orthogonal projection of X onto the span of $Y_1, \dots, Y_n$, giving

$$\hat{X} = \mathbb{E}[X] + \Sigma_{\mathbf{Y}X}^{\top}\Sigma_{\mathbf{Y}}^{-1}(\mathbf{Y} - \mathbb{E}[\mathbf{Y}]).$$

3. The corresponding minimum linear MSE is $\text{Var}(X) - \Sigma_{\mathbf{Y}X}^{\top}\Sigma_{\mathbf{Y}}^{-1}\Sigma_{\mathbf{Y}X}$.
4. We extended the same projection idea to vector-valued estimation. For zero-mean random vectors,

$$\hat{\mathbf{X}} = \Sigma_{\mathbf{X}\mathbf{Y}}\Sigma_{\mathbf{Y}}^{-1}\mathbf{Y},$$

with error covariance

$$\Sigma_{\mathcal{E}} = \Sigma_{\mathbf{X}} - \Sigma_{\mathbf{X}\mathbf{Y}}\Sigma_{\mathbf{Y}}^{-1}\Sigma_{\mathbf{Y}\mathbf{X}}.$$

Lecture 7: Gaussian Conditioning and Innovations

Throughout this lecture, whenever a covariance-matrix inverse appears, we assume that the corresponding covariance matrix is nonsingular. In the singular case, the formulas can be written using the Moore–Penrose pseudoinverse.

In the previous lecture, we derived the optimal linear estimator of a random vector X from another random vector Y , and we computed the corresponding error covariance matrix. In this lecture, we build on that result in two directions. First, we show that for jointly Gaussian random vectors, the conditional expectation is linear, so the linear MMSE estimator is in fact the full MMSE estimator. Second, we introduce the *innovations* representation, which sequentially decomposes a random vector into orthogonal prediction errors and leads naturally to whitening and Cholesky factorization.

Recap: optimal linear estimation for random vectors

Recall the vector-valued linear estimation problem from the previous lecture. If $X \in \mathbb{R}^m$ and $Y \in \mathbb{R}^n$ are zero-mean random vectors, then the optimal linear estimator of X based on Y , in the sense of minimizing

$$\mathbb{E}[\|X - \hat{X}\|^2],$$

is

$$\hat{X} = \Sigma_{XY} \Sigma_Y^{-1} Y,$$

where

$$\Sigma_Y = \mathbb{E}[YY^\top], \quad \Sigma_{XY} = \mathbb{E}[XY^\top].$$

The estimation error

$$E = X - \hat{X}$$

has covariance matrix

$$\Sigma_E = \Sigma_X - \Sigma_{XY} \Sigma_Y^{-1} \Sigma_{YX}.$$

In the non-zero-mean case, the corresponding formula is

$$\hat{X} = \mathbb{E}[X] + \Sigma_{XY} \Sigma_Y^{-1} (Y - \mathbb{E}[Y]).$$

A useful componentwise consequence is that

$$\mathbb{E}[(X_i - \hat{X}_i)^2] = (\Sigma_E)_{ii}.$$

Estimating a linear functional of a random vector

Let $X = (X_1, \dots, X_m)^\top$, and let

$$Z = \sum_{i=1}^m \alpha_i X_i = \alpha^\top X$$

for some deterministic vector $\alpha \in \mathbb{R}^m$. Since $\hat{X}$ is the optimal linear estimator of X based on Y , the optimal linear estimator of Z based on Y is simply

$$\hat{Z} = \alpha^\top \hat{X}.$$

Indeed, any linear estimator of Z based on Y must be of the form $b^\top Y$, and since

$$Z - \alpha^\top \hat{X} = \alpha^\top (X - \hat{X}),$$

the estimation error is obtained by applying the same linear functional to the vector error.

Its mean-squared error is therefore

$$\mathbb{E}[(Z - \hat{Z})^2] = \mathbb{E}[(\alpha^\top (X - \hat{X}))^2] = \alpha^\top \Sigma_E \alpha.$$

Using the expression for Σ_E , we get

$$\mathbb{E}[(Z - \hat{Z})^2] = \alpha^\top \left(\Sigma_X - \Sigma_{XY} \Sigma_Y^{-1} \Sigma_{YX} \right) \alpha.$$

Remark 7.1. This formula says that once we know the full vector error covariance, we automatically know the error for estimating any deterministic linear combination of the components of X .

Gaussian Random Vectors revisited

Now suppose

$$X \sim \mathcal{N}(\mu, \Sigma).$$

Recall that a Gaussian random vector is fully specified by its mean vector and covariance matrix.

We now partition X into two subvectors:

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}, \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}.$$

Here X_1 and X_2 may each be vector-valued.

The key fact is that the conditional distribution of one block given the other is again Gaussian:

$$X_2 \mid (X_1 = x_1) \sim \mathcal{N}\left(\mu_2 + \Sigma_{21} \Sigma_{11}^{-1} (x_1 - \mu_1), \Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12}\right).$$

Thus the conditional mean is

$$\mathbb{E}[X_2 \mid X_1] = \mu_2 + \Sigma_{21} \Sigma_{11}^{-1} (X_1 - \mu_1),$$

which is an affine function of X_1 .

Remark 7.2. This is exactly the same form as the optimal linear estimator from the previous lecture. Hence, for jointly Gaussian random vectors, the MMSE estimator and the linear MMSE estimator coincide.

Gaussian Conditioning as MMSE Estimation

Let us emphasize the previous point. If (X, Y) are jointly Gaussian, then

$$\mathbb{E}[X \mid Y] = \mathbb{E}[X] + \Sigma_{XY} \Sigma_Y^{-1} (Y - \mathbb{E}[Y]).$$

So for jointly Gaussian vectors, the best estimator in the mean-squared sense is linear in the observation Y .

Moreover, the conditional covariance is

$$\text{Cov}(X | Y) = \Sigma_X - \Sigma_{XY} \Sigma_Y^{-1} \Sigma_{YX}.$$

This is the same matrix that arose as the error covariance of the optimal linear estimator.

So in the jointly Gaussian case,

$$X - \mathbb{E}[X | Y]$$

is a Gaussian error vector whose covariance is exactly the linear-MMSE error covariance.

Innovations

We now turn to a particularly important sequential estimation construction.

Let

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}, \quad \mathbb{E}[X] = 0.$$

Assume that Σ_X is positive definite. For the first component, define $\hat{X}_1 = 0$, $I_1 = X_1$.

For each $i \geq 2$, define

$$X^{[i-1]} = \begin{bmatrix} X_1 \\ \vdots \\ X_{i-1} \end{bmatrix}.$$

We define $\hat{X}_i$ to be the optimal linear estimator of X_i based on the past observations $X^{[i-1]}$:

$$\hat{X}_i = \Sigma_{X_i, X^{[i-1]}} \Sigma_{X^{[i-1]}}^{-1} X^{[i-1]}.$$

The corresponding prediction error

$$I_i = X_i - \hat{X}_i$$

is called the i^{th} *innovation*.

Remark 7.3. The innovation I_i is the part of X_i that cannot be linearly predicted from the previous components $X_1, \dots, X_{i-1}$.

Orthogonality of the innovations

By the orthogonality principle, the estimation error is orthogonal to the data used to form the estimator. Therefore,

$$I_i = X_i - \hat{X}_i \perp \{X_1, \dots, X_{i-1}\}.$$

Since each earlier innovation $I_1, \dots, I_{i-1}$ is itself a linear combination of $X_1, \dots, X_{i-1}$, it follows that

$$I_i \perp \{I_1, \dots, I_{i-1}\}.$$

Hence the innovations are pairwise orthogonal. Since they are zero mean, they are uncorrelated:

$$\text{Cov}(I_i, I_j) = 0, \quad i \neq j.$$

If we normalize them by

$$W_i = \frac{I_i}{\sqrt{\text{Var}(I_i)}},$$

then

$$\mathbb{E}[W_i] = 0, \quad \text{Var}(W_i) = 1,$$

and the vector

$$W = \begin{bmatrix} W_1 \\ \vdots \\ W_n \end{bmatrix}$$

satisfies

$$\Sigma_W = I.$$

So, the normalized innovations form a white random vector.

Sequential decomposition and triangular structure

Because

$$X_i = \hat{X}_i + I_i$$

and $\hat{X}_i$ depends only on $X_1, \dots, X_{i-1}$, we can recursively express each X_i as a linear combination of the innovations $I_1, \dots, I_i$, and hence also of the normalized innovations $W_1, \dots, W_i$.

Stacking these relations gives

$$X = LW,$$

where L is a lower triangular matrix. Since W is white,

$$\Sigma_X = \mathbb{E}[XX^\top] = \mathbb{E}[LWW^\top L^\top] = L \mathbb{E}[WW^\top] L^\top = LL^\top.$$

Thus the innovations construction produces a lower-triangular square-root factor of the covariance matrix.

Remark 7.4. This is precisely the Cholesky factorization of Σ_X : for a positive definite covariance matrix, there exists a unique lower triangular matrix L with positive diagonal entries such that

$$\Sigma_X = LL^\top.$$

The innovations procedure provides a probabilistic interpretation of this factorization.

Orthogonal decomposition of estimation from independent blocks

Suppose now that

$$U = \begin{bmatrix} U_1 \\ U_2 \end{bmatrix}, \quad \Sigma_U = \begin{bmatrix} \Sigma_{U_1} & 0 \\ 0 & \Sigma_{U_2} \end{bmatrix},$$

so that U_1 and U_2 are uncorrelated. The best linear estimator of X based on U is

$$\hat{X} = \Sigma_{XU} \Sigma_U^{-1} U.$$

Writing

$$\Sigma_{XU} = \begin{bmatrix} \Sigma_{XU_1} & \Sigma_{XU_2} \end{bmatrix}, \quad \Sigma_U^{-1} = \begin{bmatrix} \Sigma_{U_1}^{-1} & 0 \\ 0 & \Sigma_{U_2}^{-1} \end{bmatrix},$$

we obtain

$$\hat{X} = \Sigma_{XU_1} \Sigma_{U_1}^{-1} U_1 + \Sigma_{XU_2} \Sigma_{U_2}^{-1} U_2.$$

So when the observation splits into orthogonal pieces, the optimal linear estimator splits into the sum of the optimal linear estimators based on each piece separately.

Remark 7.5. This is especially useful when U is formed from innovations or other whitened coordinates: orthogonal components contribute additively to the estimator.

Summary

The main takeaways from this lecture are:

- For any deterministic vector α , the best linear estimator of $Z = \alpha^\top X$ based on Y is $\hat{Z} = \alpha^\top \hat{X}$, and its error is $\alpha^\top \Sigma_E \alpha$.
- If a Gaussian vector is partitioned into two blocks, then the conditional distribution of one block given the other is again Gaussian.
- For jointly Gaussian vectors, the conditional expectation is affine, so the MMSE estimator equals the linear MMSE estimator.
- The conditional covariance in the Gaussian case is

$$\Sigma_X - \Sigma_{XY} \Sigma_Y^{-1} \Sigma_{YX},$$

the same matrix that appeared earlier as the linear estimation error covariance.

- The innovations $I_i = X_i - \hat{X}_i$ are orthogonal prediction errors.
- After normalization, the innovations form a white vector W , and the original vector can be written as

$$X = LW,$$

where L is lower triangular and satisfies

$$\Sigma_X = LL^\top.$$

Lecture 8: Kalman Filtering

Optimal Linear Estimation and Projection

Let X and Z be zero-mean random variables. We denote by $\hat{X}(Z)$ the optimal linear estimator of X based on Z , i.e.,

$$a^* = \arg \min_a \mathbb{E} \left[(X - aZ)^2 \right], \quad \hat{X}(Z) = a^* Z.$$

Equivalently, $\hat{X}(Z)$ is the orthogonal projection of X onto the linear span of Z .

Innovations and Orthogonality

Let Y_1 and Y_2 be zero-mean random variables. Define

$$\tilde{Y}_2 = Y_2 - \hat{Y}_2(Y_1).$$

Then $\tilde{Y}_2$ is orthogonal to Y_1 , so

$$\mathbb{E}[\tilde{Y}_2 Y_1] = 0.$$

Hence

$$Y_2 = \hat{Y}_2(Y_1) + \tilde{Y}_2.$$

Now consider the problem of estimating X based on (Y_1, Y_2) . Since (Y_1, Y_2) and $(Y_1, \tilde{Y}_2)$ span the same space, we have

$$\hat{X}(Y_1, Y_2) = \hat{X}(Y_1, \tilde{Y}_2).$$

Because $\tilde{Y}_2 \perp Y_1$, the projection decomposes as

$$\hat{X}(Y_1, \tilde{Y}_2) = \hat{X}(Y_1) + \hat{X}(\tilde{Y}_2).$$

Thus

$$\hat{X}(Y_1, Y_2) = \hat{X}(Y_1) + \hat{X}(\tilde{Y}_2).$$

Note: No Gaussian assumption is needed here; the derivation uses only zero means, uncorrelatedness, and second-order statistics.

Scalar State-Space Model

We now consider a sequence of random variables evolving in time according to

$$X_{i+1} = a_i X_i + U_{i+1}, \quad 0 \leq i \leq n-1,$$

$$Y_i = X_i + V_i, \quad 0 \leq i \leq n.$$

The random variables

$$X_0, U_1, \dots, U_n, V_0, \dots, V_n$$

are assumed to be zero-mean, mutually uncorrelated, and to have known variances. The coefficients $a_0, \dots, a_{n-1}$ are deterministic and known.

Estimation Objectives

For each i , define

$$Y^{[i]} = \begin{bmatrix} Y_0 \\ Y_1 \\ \vdots \\ Y_i \end{bmatrix}.$$

For convenience, whenever $j \geq i$, we write $\hat{X}_{j|i}$ for the optimal linear estimate of X_j based on $Y^{[i]}$.

In particular, we are interested in the following two cases.

Prediction:

$$\hat{X}_{i+1|i} = \hat{X}_{i+1}(Y^{[i]}).$$

Filtering:

$$\hat{X}_{i|i} = \hat{X}_i(Y^{[i]}).$$

We also define the corresponding mean-square errors

$$\sigma_{i+1|i}^2 = \text{Var}(X_{i+1} - \hat{X}_{i+1|i}),$$

$$\sigma_{i|i}^2 = \text{Var}(X_i - \hat{X}_{i|i}).$$

Sequential Innovations

For each $i \geq 0$, define the innovation

$$\tilde{Y}_i = Y_i - \hat{Y}_i(Y^{[i-1]}),$$

with the convention that $\tilde{Y}_0 = Y_0$.

By the orthogonality principle,

$$\tilde{Y}_i \perp Y^{[i-1]}.$$

Since

$$Y_i = X_i + V_i$$

and V_i is uncorrelated with all past observations, we have

$$\hat{Y}_i(Y^{[i-1]}) = \hat{X}_{i|i-1}.$$

Therefore

$$\tilde{Y}_i = Y_i - \hat{X}_{i|i-1}.$$

Prediction Recursion

Since $Y^{[i]}$ and $(Y^{[i-1]}, \tilde{Y}_i)$ span the same space,

$$\hat{X}_{i+1|i} = \hat{X}_{i+1}(Y^{[i-1]}, \tilde{Y}_i).$$

Using orthogonality, this becomes

$$\hat{X}_{i+1|i} = \hat{X}_{i+1|i-1} + \hat{X}_{i+1}(\tilde{Y}_i).$$

Now

$$X_{i+1} = a_i X_i + U_{i+1},$$

and U_{i+1} is uncorrelated with $Y^{[i-1]}$, so

$$\hat{X}_{i+1|i-1} = a_i \hat{X}_{i|i-1}.$$

Also, since $\tilde{Y}_i$ is scalar,

$$\hat{X}_{i+1}(\tilde{Y}_i) = \frac{\text{Cov}(X_{i+1}, \tilde{Y}_i)}{\text{Var}(\tilde{Y}_i)} \tilde{Y}_i.$$

Using the state equation again and the fact that U_{i+1} is uncorrelated with $\tilde{Y}_i$,

$$\text{Cov}(X_{i+1}, \tilde{Y}_i) = a_i \text{Cov}(X_i, \tilde{Y}_i).$$

Hence

$$\hat{X}_{i+1|i} = a_i \hat{X}_{i|i-1} + a_i \frac{\text{Cov}(X_i, \tilde{Y}_i)}{\text{Var}(\tilde{Y}_i)} \tilde{Y}_i.$$

We now compute the two quantities in this coefficient. First,

$$\text{Cov}(X_i, \tilde{Y}_i) = \text{Cov}(X_i, Y_i - \hat{X}_{i|i-1}).$$

Since $Y_i = X_i + V_i$ and V_i is uncorrelated with X_i and $\hat{X}_{i|i-1}$,

$$\text{Cov}(X_i, \tilde{Y}_i) = \text{Cov}(X_i, X_i - \hat{X}_{i|i-1}).$$

Because

$$\hat{X}_{i|i-1} \perp X_i - \hat{X}_{i|i-1},$$

we obtain

$$\text{Cov}(X_i, \tilde{Y}_i) = \text{Var}(X_i - \hat{X}_{i|i-1}) = \sigma_{i|i-1}^2.$$

Next,

$$\text{Var}(\tilde{Y}_i) = \text{Var}(Y_i - \hat{X}_{i|i-1}) = \text{Var}(X_i - \hat{X}_{i|i-1} + V_i).$$

Since V_i is uncorrelated with $X_i - \hat{X}_{i|i-1}$,

$$\text{Var}(\tilde{Y}_i) = \sigma_{i|i-1}^2 + \text{Var}(V_i).$$

Substituting these expressions gives

$$\hat{X}_{i+1|i} = a_i \hat{X}_{i|i-1} + \frac{a_i \sigma_{i|i-1}^2}{\sigma_{i|i-1}^2 + \text{Var}(V_i)} \tilde{Y}_i.$$

Since $\tilde{Y}_i = Y_i - \hat{X}_{i|i-1}$, we arrive at

$$\hat{X}_{i+1|i} = a_i \hat{X}_{i|i-1} + k_i (Y_i - \hat{X}_{i|i-1}),$$

where

$$k_i \triangleq \frac{a_i \sigma_{i|i-1}^2}{\sigma_{i|i-1}^2 + \text{Var}(V_i)}.$$

Prediction Error Variance Recursion

We next compute the error variance of the predictor. By definition,

$$\sigma_{i+1|i}^2 = \text{Var}(X_{i+1} - \hat{X}_{i+1|i}).$$

Using the state equation and the recursion just derived,

$$X_{i+1} - \hat{X}_{i+1|i} = a_i X_i + U_{i+1} - a_i \hat{X}_{i|i-1} - k_i(Y_i - \hat{X}_{i|i-1}).$$

Since $Y_i = X_i + V_i$, this becomes

$$X_{i+1} - \hat{X}_{i+1|i} = (a_i - k_i)(X_i - \hat{X}_{i|i-1}) - k_i V_i + U_{i+1}.$$

The three terms on the right-hand side are pairwise uncorrelated, so

$$\sigma_{i+1|i}^2 = (a_i - k_i)^2 \sigma_{i|i-1}^2 + k_i^2 \text{Var}(V_i) + \text{Var}(U_{i+1}).$$

Using

$$k_i(\sigma_{i|i-1}^2 + \text{Var}(V_i)) = a_i \sigma_{i|i-1}^2,$$

we can simplify the first two terms to obtain

$$\sigma_{i+1|i}^2 = a_i(a_i - k_i) \sigma_{i|i-1}^2 + \text{Var}(U_{i+1}).$$

Scalar Kalman Filter for Prediction

We therefore obtain the following recursion for prediction.

Initialize with

$$\hat{X}_{0|-1} = 0, \quad \sigma_{0|-1}^2 = \text{Var}(X_0).$$

Then, for $i = 0, 1, \dots, n-1$,

$$\hat{X}_{i+1|i} = a_i \hat{X}_{i|i-1} + k_i(Y_i - \hat{X}_{i|i-1}),$$

where

$$k_i = \frac{a_i \sigma_{i|i-1}^2}{\sigma_{i|i-1}^2 + \text{Var}(V_i)},$$

and

$$\sigma_{i+1|i}^2 = a_i(a_i - k_i) \sigma_{i|i-1}^2 + \text{Var}(U_{i+1}).$$

From Prediction to Filtering

We now derive the corresponding recursion for filtering. Using the same innovation decomposition,

$$\hat{X}_{i|i} = \hat{X}_i(Y^{[i-1]}, \tilde{Y}_i) = \hat{X}_{i|i-1} + \hat{X}_i(\tilde{Y}_i).$$

Since $\tilde{Y}_i$ is scalar,

$$\hat{X}_i(\tilde{Y}_i) = \frac{\text{Cov}(X_i, \tilde{Y}_i)}{\text{Var}(\tilde{Y}_i)} \tilde{Y}_i.$$

From the calculations above,

$$\text{Cov}(X_i, \tilde{Y}_i) = \sigma_{i|i-1}^2, \quad \text{Var}(\tilde{Y}_i) = \sigma_{i|i-1}^2 + \text{Var}(V_i).$$

Therefore

$$\hat{X}_{i|i} = \hat{X}_{i|i-1} + \frac{\sigma_{i|i-1}^2}{\sigma_{i|i-1}^2 + \text{Var}(V_i)} (Y_i - \hat{X}_{i|i-1}).$$

Filtering Error Variance

The filtering error variance is

$$\sigma_{i|i}^2 = \text{Var}(X_i - \hat{X}_{i|i}).$$

A direct computation yields

$$X_i - \hat{X}_{i|i} = \left(1 - \frac{\sigma_{i|i-1}^2}{\sigma_{i|i-1}^2 + \text{Var}(V_i)}\right) (X_i - \hat{X}_{i|i-1}) - \frac{\sigma_{i|i-1}^2}{\sigma_{i|i-1}^2 + \text{Var}(V_i)} V_i.$$

Since the two terms on the right-hand side are uncorrelated,

$$\sigma_{i|i}^2 = \frac{\sigma_{i|i-1}^2 \text{Var}(V_i)}{\sigma_{i|i-1}^2 + \text{Var}(V_i)}.$$

Equivalently,

$$\sigma_{i|i}^2 = \left(1 - \frac{\sigma_{i|i-1}^2}{\sigma_{i|i-1}^2 + \text{Var}(V_i)}\right) \sigma_{i|i-1}^2.$$

Filtering Form of the Recursion

Using

$$\hat{X}_{i|i-1} = a_{i-1} \hat{X}_{i-1|i-1}$$

and

$$\sigma_{i|i-1}^2 = a_{i-1}^2 \sigma_{i-1|i-1}^2 + \text{Var}(U_i),$$

we may also write the filtering recursion directly in terms of the previous filtered estimate. For $i \geq 1$,

$$\hat{X}_{i|i} = a_{i-1}(1 - \ell_i) \hat{X}_{i-1|i-1} + \ell_i Y_i,$$

where

$$\ell_i = \frac{a_{i-1}^2 \sigma_{i-1|i-1}^2 + \text{Var}(U_i)}{a_{i-1}^2 \sigma_{i-1|i-1}^2 + \text{Var}(U_i) + \text{Var}(V_i)},$$

and

$$\sigma_{i|i}^2 = (1 - \ell_i)(a_{i-1}^2 \sigma_{i-1|i-1}^2 + \text{Var}(U_i)).$$

The initialization is

$$\begin{aligned} \hat{X}_{0|0} &= \frac{\text{Var}(X_0)}{\text{Var}(X_0) + \text{Var}(V_0)} Y_0, \\ \sigma_{0|0}^2 &= \frac{\text{Var}(X_0) \text{Var}(V_0)}{\text{Var}(X_0) + \text{Var}(V_0)}. \end{aligned}$$

Summary

1. We first used the projection viewpoint of linear MMSE estimation to introduce innovations. The key idea is that each new observation can be decomposed into the part already predictable from the past and a new orthogonal component.

2. For the scalar state-space model

$$X_{i+1} = a_i X_i + U_{i+1}, \quad Y_i = X_i + V_i,$$

we derived a recursive prediction update for $\hat{X}_{i+1|i}$ using the innovation $Y_i - \hat{X}_{i|i-1}$.

3. We then derived the corresponding recursion for the prediction error variance $\sigma_{i+1|i}^2$, showing that the estimator can be updated using only the previous estimate, the previous error variance, and the noise variances.
4. Finally, we converted the prediction recursion into the filtering recursion for $\hat{X}_{i|i}$, where the current observation Y_i is incorporated to update the estimate of the current state.

Lecture 9: Kalman Filtering II

Scalar Kalman Filter Recap

Recall the scalar state-space model

$$X_{i+1} = a_i X_i + U_{i+1}, \quad 0 \leq i \leq n-1,$$

and

$$Y_i = X_i + V_i, \quad 0 \leq i \leq n.$$

The sequence X_i is the state sequence, and Y_i is the observation sequence. The random variables U_i drive the state process, while the random variables V_i are the observation noise terms.

We assume that $X_0, U_1, \dots, U_n, V_0, \dots, V_n$ are zero mean and mutually uncorrelated, with known variances. The coefficients $a_0, \dots, a_{n-1}$ are also assumed known. If the means are known but nonzero, we can subtract them off and work with the corresponding zero-mean variables, so the zero-mean assumption is mainly for notational simplicity.

The statistical information being used here is first and second-order information. Thus the object we derived is the optimal linear estimator under the given second-order model.

For each i , let

$$Y^{[i]} = \begin{bmatrix} Y_0 \\ Y_1 \\ \vdots \\ Y_i \end{bmatrix}.$$

For $j \geq i$, write $\hat{X}_{j|i}$ for the optimal linear estimator of X_j based on $Y^{[i]}$. Additionally, define the corresponding mean-square error

$$\sigma_{j|i}^2 = \mathbb{E} \left[(X_j - \hat{X}_{j|i})^2 \right] = \text{Var}(X_j - \hat{X}_{j|i}).$$

The equality with the variance follows because the estimation error has mean zero in this zero-mean linear-estimation setting.

From the previous lecture, the scalar prediction recursion is initialized by

$$\hat{X}_{0|-1} = 0, \quad \sigma_{0|-1}^2 = \text{Var}(X_0).$$

Then, for $i = 0, 1, \dots, n-1$,

$$\hat{X}_{i+1|i} = a_i \hat{X}_{i|i-1} + k_i (Y_i - \hat{X}_{i|i-1}),$$

where

$$k_i = \frac{a_i \sigma_{i|i-1}^2}{\sigma_{i|i-1}^2 + \text{Var}(V_i)}.$$

The prediction error variance satisfies

$$\sigma_{i+1|i}^2 = (a_i - k_i)^2 \sigma_{i|i-1}^2 + k_i^2 \text{Var}(V_i) + \text{Var}(U_{i+1}).$$

Using the expression for k_i , this can be simplified to

$$\sigma_{i+1|i}^2 = a_i(a_i - k_i) \sigma_{i|i-1}^2 + \text{Var}(U_{i+1}).$$

Facts and Observations

The following points are not new recursions; they are the orthogonality facts used repeatedly in deriving the Kalman filter. Throughout this section, $\perp$ denotes orthogonality, i.e., zero covariance.

1. The process noise is the innovation in the state process.

Define the state innovation by

$$\tilde{X}_i \triangleq X_i - \hat{X}_i(X^{[i-1]}),$$

where $X^{[i-1]} = (X_0, \dots, X_{i-1})$. Since

$$X_i = a_{i-1}X_{i-1} + U_i,$$

we have

$$\hat{X}_i(X^{[i-1]}) = a_{i-1}X_{i-1}.$$

Therefore

$$\tilde{X}_i = X_i - \hat{X}_i(X^{[i-1]}) = U_i.$$

Thus, U_i is precisely the new part of X_i not explained by the past state sequence. The term $a_{i-1}X_{i-1}$ is the predictable part, while U_i is the fresh randomness added at time i .

2. The observation innovation gives the recursive update.

Define the observation innovation by

$$\tilde{Y}_i \triangleq Y_i - \hat{Y}_i(Y^{[i-1]}).$$

Since $Y_i = X_i + V_i$ and V_i is uncorrelated with all past observations, we have

$$\hat{Y}_i(Y^{[i-1]}) = \hat{X}_{i|i-1}.$$

Hence

$$\tilde{Y}_i = Y_i - \hat{X}_{i|i-1}.$$

The variables $Y^{[i]}$ and $(Y^{[i-1]}, \tilde{Y}_i)$ span the same linear space. Indeed, given $Y^{[i]}$, we can form $Y^{[i-1]}$ and $\tilde{Y}_i$; conversely, given $Y^{[i-1]}$ and $\tilde{Y}_i$, we can recover Y_i by adding back $\hat{X}_{i|i-1}$, which is itself a linear function of $Y^{[i-1]}$. Therefore

$$\hat{X}_{i|i} = \hat{X}_i(Y^{[i]}) = \hat{X}_i(Y^{[i-1]}, \tilde{Y}_i).$$

Since $\tilde{Y}_i$ is orthogonal to $Y^{[i-1]}$, the projection onto the sum of these two orthogonal subspaces is the sum of the projections:

$$\hat{X}_{i|i} = \hat{X}_i(Y^{[i-1]}) + \hat{X}_i(\tilde{Y}_i).$$

Therefore

$$\hat{X}_{i|i} = \hat{X}_{i|i-1} + \hat{X}_i(\tilde{Y}_i).$$

3. U_{i+1} is orthogonal to the observations up to time i .

Intuitively, $Y^{[i]}$ consists of noisy observations of the state sequence only up to time i , while U_{i+1} is the new randomness that drives the next state. More formally,

$$\text{span}\{Y_0, \dots, Y_i\} \subseteq \text{span}\{X_0, U_1, \dots, U_i, V_0, \dots, V_i\}.$$

Since U_{i+1} is uncorrelated with every variable on the right-hand side, it is orthogonal to every observation up to time i . Therefore,

$$U_{i+1} \perp Y_j \text{ for } j = 0, \dots, i.$$

4. The prediction error is orthogonal to the predictor.

By the orthogonality principle,

$$\hat{X}_{i|i-1} \perp (X_i - \hat{X}_{i|i-1}).$$

Consequently,

$$\text{Cov}(X_i, X_i - \hat{X}_{i|i-1}) = \text{Cov}(X_i - \hat{X}_{i|i-1}, X_i - \hat{X}_{i|i-1}).$$

Hence

$$\text{Cov}(X_i, X_i - \hat{X}_{i|i-1}) = \text{Var}(X_i - \hat{X}_{i|i-1}) = \sigma_{i|i-1}^2.$$

5. The prediction error is orthogonal to the current observation noise.

Since X_i depends only on $X_0, U_1, \dots, U_i$, and since $\hat{X}_{i|i-1}$ is a linear function of $Y^{[i-1]}$, we have

$$X_i - \hat{X}_{i|i-1} \in \text{span}(X_0, U_1, \dots, U_i, V_0, \dots, V_{i-1}).$$

Since V_i is uncorrelated with all variables in this span,

$$X_i - \hat{X}_{i|i-1} \perp V_i.$$

Therefore

$$\text{Var}(\tilde{Y}_i) = \text{Var}(X_i - \hat{X}_{i|i-1} + V_i) = \sigma_{i|i-1}^2 + \text{Var}(V_i).$$

A special case: So far, we have only used first and second-order information, and the estimator is optimal among linear estimators. If, in addition, $X_0, U_1, \dots, U_n, V_0, \dots, V_n$ are *jointly Gaussian*, then uncorrelatedness implies independence. Since all variables in the state-space model are obtained from these Gaussian variables by linear transformations, the variables involved are jointly Gaussian.

For jointly Gaussian random variables, conditional expectations are linear functions of the observations. Therefore, in the Gaussian case, the Kalman filter derived above is not only the optimal linear filter, but the optimal filter among all estimators.

Example: Repeated Noisy Measurements of a Fixed Random Variable

Consider the special case $a_i = 1$, $\text{Var}(U_i) = 0$, $1 \leq i \leq n$. Then $X_0 = X_1 = \dots = X_n = X$, where $\text{Var}(X) = P$. Assume also that $\text{Var}(V_i) = \sigma^2$ for all i . Thus we repeatedly observe

$$Y_i = X + V_i.$$

This is the problem of estimating a fixed random amplitude X from repeated noisy measurements.

The Kalman filter is initialized as

$$\hat{X}_{0|-1} = 0, \sigma_{0|-1}^2 = P.$$

For this model, the recursion becomes

$$\hat{X}_{i+1|i} = \hat{X}_{i|i-1} + k_i (Y_i - \hat{X}_{i|i-1}),$$

with

$$k_i = \frac{\sigma_{i|i-1}^2}{\sigma_{i|i-1}^2 + \sigma^2}.$$

The error variance recursion becomes

$$\sigma_{i+1|i}^2 = (1 - k_i) \sigma_{i|i-1}^2.$$

Substituting the expression for k_i gives

$$\sigma_{i+1|i}^2 = \frac{\sigma^2 \sigma_{i|i-1}^2}{\sigma_{i|i-1}^2 + \sigma^2}.$$

Equivalently,

$$\frac{1}{\sigma_{i+1|i}^2} = \frac{1}{\sigma_{i|i-1}^2} + \frac{1}{\sigma^2}.$$

This shows that the inverse error variance accumulates additively. The quantity $1/\sigma_{i|i-1}^2$ may be interpreted as the current precision of the estimate, and each new observation contributes an additional precision $1/\sigma^2$. In terms of the number of measurements incorporated, if P_m denotes the error variance after m noisy measurements, then

$$\frac{1}{P_m} = \frac{1}{P} + \frac{m}{\sigma^2}.$$

Therefore

$$P_m = \frac{P\sigma^2}{mP + \sigma^2}.$$

The gain used on the next measurement is

$$K_m = \frac{P_m}{P_m + \sigma^2} = \frac{P}{(m+1)P + \sigma^2}.$$

Equivalently, if K_i denotes the gain on the i -th measurement, with $i = 1$ for the first measurement, then

$$K_i = \frac{P}{iP + \sigma^2}.$$

Thus, as more measurements are observed, the error variance decreases like

$$P_m = \frac{P\sigma^2}{mP + \sigma^2}.$$

In particular, the error decreases on the order of $1/m$. The Kalman gain also decreases over time: once many comparable measurements have already been seen, one additional measurement contributes less to the estimate.

Vector Kalman Filter

We now move from the scalar model to the vector state-space model

$$X_{i+1} = A_i X_i + U_{i+1}, \quad 0 \leq i \leq n-1,$$

and

$$Y_i = X_i + V_i, \quad 0 \leq i \leq n.$$

Here X_i, Y_i, U_i, V_i are random vectors. We assume that $X_0, U_1, \dots, U_n, V_0, \dots, V_n$ are zero mean and mutually uncorrelated random vectors, with known covariance matrices. The matrices $A_0, \dots, A_{n-1}$ are deterministic and known.

As before, let $Y^{[i]} = (Y_0, \dots, Y_i)$. Define $\hat{X}_{i+1|i}$ to be the optimal linear estimator of X_{i+1} based on $Y^{[i]}$. Define the prediction error covariance matrix by

$$\Sigma_{i+1|i} = \text{Cov} \left(X_{i+1} - \hat{X}_{i+1|i} \right).$$

The vector Kalman filter is initialized as

$$\hat{X}_{0|-1} = 0, \quad \Sigma_{0|-1} = \Sigma_{X_0}.$$

For $i = 0, 1, \dots, n-1$, the recursion is

$$\hat{X}_{i+1|i} = A_i \hat{X}_{i|i-1} + K_i \left(Y_i - \hat{X}_{i|i-1} \right),$$

where

$$K_i = A_i \Sigma_{i|i-1} \left(\Sigma_{i|i-1} + \Sigma_{V_i} \right)^{-1}.$$

The prediction error covariance satisfies

$$\Sigma_{i+1|i} = (A_i - K_i) \Sigma_{i|i-1} A_i^T + \Sigma_{U_{i+1}}.$$

This is the direct vector analogue of the scalar Kalman filter. The scalar variance $\sigma_{i|i-1}^2$ is replaced by the error covariance matrix $\Sigma_{i|i-1}$, and division by the innovation variance becomes multiplication by the inverse innovation covariance matrix. In practice, vector states allow us to track several coupled quantities at once, such as the spatial position of an object, or both its position and velocity.

We assume here that the innovation covariance $\Sigma_{i|i-1} + \Sigma_{V_i}$ is nonsingular. If it is singular, then there are degenerate directions in the observation, and one can reduce to the nonsingular observation subspace or handle the trivial directions separately.

Deriving the Vector Recursion

In this lecture, we do not repeat the full scalar derivation in the vector case; the point is that the same innovation argument carries over almost verbatim with vectors and matrices replacing scalars.

Define the observation innovation by

$$\tilde{Y}_i = Y_i - \hat{Y}_i(Y^{[i-1]}).$$

Since $Y_i = X_i + V_i$ and V_i is uncorrelated with all past observations,

$$\tilde{Y}_i = Y_i - \hat{X}_{i|i-1}.$$

As in the scalar case, $Y^{[i]}$ and $(Y^{[i-1]}, \tilde{Y}_i)$ span the same linear space, while $\tilde{Y}_i$ is orthogonal to $Y^{[i-1]}$. Therefore the new observation affects the estimate only through the innovation $Y_i - \hat{X}_{i|i-1}$.

For a vector observation, the optimal linear estimate of one vector based on another uses the matrix projection formula

$$\hat{X}(Z) = \text{Cov}(X, Z) \text{Cov}(Z)^{-1} Z.$$

Applying this with $X = X_{i+1}$ and $Z = \tilde{Y}_i$ gives

$$\hat{X}_{i+1|i} = A_i \hat{X}_{i|i-1} + \text{Cov}(X_{i+1}, \tilde{Y}_i) \text{Cov}(\tilde{Y}_i)^{-1} \tilde{Y}_i.$$

The same covariance computations as in the scalar case give

$$\text{Cov}(X_{i+1}, \tilde{Y}_i) = A_i \Sigma_{i|i-1},$$

and

$$\text{Cov}(\tilde{Y}_i) = \Sigma_{i|i-1} + \Sigma_{V_i}.$$

Thus

$$K_i = A_i \Sigma_{i|i-1} \left(\Sigma_{i|i-1} + \Sigma_{V_i} \right)^{-1}.$$

The covariance recursion is obtained by writing the prediction error as

$$X_{i+1} - \hat{X}_{i+1|i} = (A_i - K_i)(X_i - \hat{X}_{i|i-1}) - K_i V_i + U_{i+1},$$

and then using the same orthogonality assumptions as before. After simplifying using the expression for K_i , this gives

$$\Sigma_{i+1|i} = (A_i - K_i) \Sigma_{i|i-1} A_i^T + \Sigma_{U_{i+1}}.$$

Toward Random Processes

The Kalman filter naturally suggests moving beyond a fixed finite-dimensional random vector. The recursion does not require us to know a final time in advance: another observation arrives, the estimate is updated, and the procedure can continue. This motivates the next topic, **random processes**, where we study collections of random variables indexed by time. In this sense, Kalman filtering provides a natural bridge from finite-dimensional linear estimation to estimation and prediction for time-indexed stochastic systems.

Lecture 10: Random Processes I

The Kalman filter naturally leads us from finite-dimensional random vectors to collections of random variables indexed by time. This motivates the study of random processes.

Definition

A *random process*, also called a *stochastic process*, is a collection of random variables

$$\{X(t), t \in \mathcal{T}\},$$

where t is the index and $\mathcal{T}$ is the index set.

The index set $\mathcal{T}$ is often interpreted as time. Some common examples are

$$\mathcal{T} = \{1, 2, 3, \dots\},$$

$$\mathcal{T} = \{\dots, -2, -1, 0, 1, 2, \dots\},$$

$$\mathcal{T} = [0, T],$$

$$\mathcal{T} = [0, \infty),$$

and

$$\mathcal{T} = \mathbb{R}.$$

When $\mathcal{T}$ is countable, we usually call the process a *discrete-time process*. When $\mathcal{T}$ is an interval of real numbers, we usually call the process a *continuous-time process*.

A realization of the random process is called a *sample path*. More explicitly, if ω is the outcome of the underlying experiment, then

$$t \mapsto X(t, \omega)$$

is one sample path of the process.

Example 1: A Discrete-Time Process

Let

$$U \sim \text{Unif}[0, 1],$$

and define

$$X(n) = U^n, \quad n = 1, 2, \dots$$

Here

$$\mathcal{T} = \{1, 2, \dots\}.$$

This is a discrete-time random process. The randomness comes from the single random variable U . Once U is chosen, the whole sample path is determined.

For example, if $U = \frac{1}{2}$, then the sample path is

$$X(n) = \left(\frac{1}{2}\right)^n.$$

If $U = \frac{1}{\sqrt{2}}$, then the sample path is

$$X(n) = \left(\frac{1}{\sqrt{2}}\right)^n.$$

For each fixed n , $X(n)$ is a random variable. Its CDF is

$$F_{X(n)}(x) = \Pr(X(n) \leq x) = \Pr(U^n \leq x).$$

Since $U \in [0, 1]$, for $0 \leq x \leq 1$ this becomes

$$F_{X(n)}(x) = \Pr(U \leq x^{1/n}) = x^{1/n}.$$

Thus

$$F_{X(n)}(x) = \begin{cases} 0, & x < 0, \\ x^{1/n}, & 0 \leq x \leq 1, \\ 1, & x > 1. \end{cases}$$

Notice that the random variables $X(1), X(2), \dots$ are not independent. They are all functions of the same underlying random variable U .

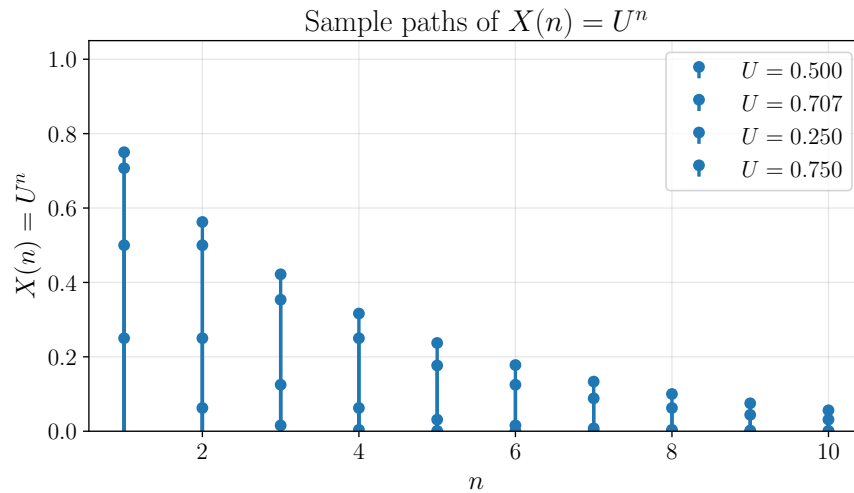


Figure 3: Sample paths of the discrete-time process $X(n) = U^n$. Once U is chosen, the whole path is determined.

Example 2: A Continuous-Time Process

Let

$$\Theta \sim \text{Unif}[0, 2\pi],$$

and define

$$X(t) = \cos(t + \Theta), \quad t \geq 0.$$

Here

$$\mathcal{T} = [0, \infty).$$

This is a continuous-time random process. Again, once the random phase Θ is chosen, the whole sample path is determined.

For example, if $\Theta = 0$, then

$$X(t) = \cos t.$$

If $\Theta = \frac{\pi}{2}$, then

$$X(t) = \cos\left(t + \frac{\pi}{2}\right).$$

So the sample paths are shifted cosine waves. The randomness is in the initial phase.

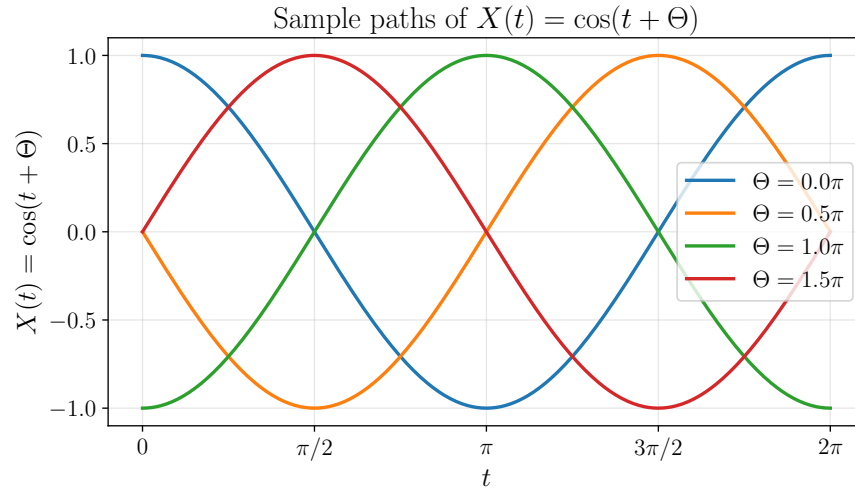


Figure 4: Sample paths of the random phase cosine process $X(t) = \cos(t + \Theta)$. The randomness is only in the phase shift.

Specification of a Random Process

A random process is specified by all of its finite-dimensional CDFs. That is, for every n and every choice of time indices $t_1, t_2, \dots, t_n \in \mathcal{T}$, we specify the joint CDF

$$F_{t_1, \dots, t_n}(x_1, \dots, x_n) = \Pr(X(t_1) \leq x_1, X(t_2) \leq x_2, \dots, X(t_n) \leq x_n).$$

These finite-dimensional distributions describe the joint behavior of the process at any finite collection of time points.

There is a consistency requirement: the finite-dimensional distributions must agree with one another when variables are reordered or marginalized out. Under the appropriate consistency conditions, Kolmogorov's extension theorem says that such a family of finite-dimensional distributions defines a random process.

Important Random Processes

1. IID Process

A discrete-time process

$$\{X(n), n \in \mathbb{N}\}$$

is an *IID process* if the random variables are independent and identically distributed.

Equivalently, for any n ,

$$X(1), X(2), \dots, X(n)$$

are IID with common marginal CDF F_X .

Thus the joint CDF factors as

$$\Pr(X(1) \leq x_1, \dots, X(n) \leq x_n) = \prod_{i=1}^n F_X(x_i).$$

2. Random Walk

Let

$$Z(1), Z(2), \dots$$

be IID random variables with

$$Z(n) = \begin{cases} +1, & \text{with probability } 1/2, \\ -1, & \text{with probability } 1/2. \end{cases}$$

Define the random walk process by

$$X(0) = 0,$$

and for $n \geq 1$,

$$X(n) = \sum_{i=1}^n Z(i).$$

Equivalently,

$$X(n+1) = X(n) + Z(n+1).$$

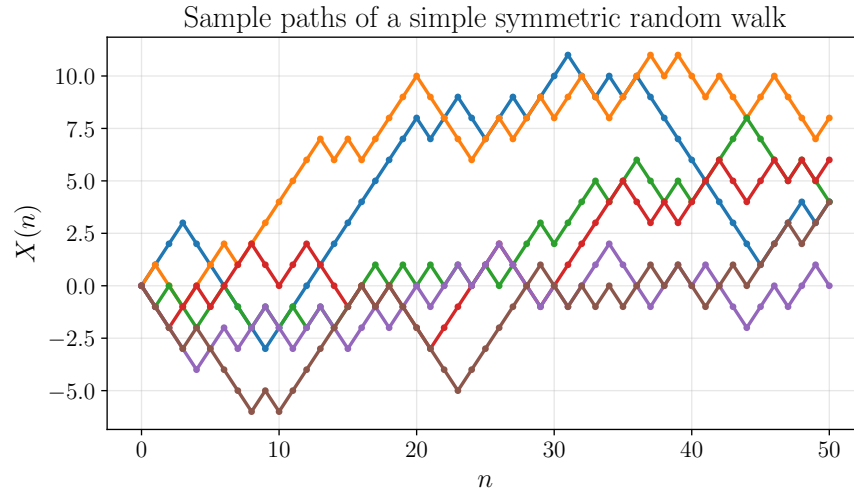


Figure 5: Sample paths of a simple symmetric random walk. At each step, the process moves by $+1$ or -1 .

The process moves one step to the right or one step to the left at each time. Since the increments are independent, the future movement after time n is determined by the current position $X(n)$ plus fresh randomness.

3. Markov Process

A discrete-time process

$$\{X(n), n \in \mathbb{N}\}$$

is a *Markov process* if the future depends on the past only through the present.

Formally, for every n and every set A ,

$$\Pr(X(n+1) \in A \mid X(1), X(2), \dots, X(n)) = \Pr(X(n+1) \in A \mid X(n)),$$

whenever the conditional probabilities are well-defined.

Equivalently, conditioned on $X(n)$, the random variable $X(n+1)$ is independent of the earlier history $X(1), \dots, X(n-1)$.

Thus, conditioned on the present state, the next state is independent of the earlier past.

Random Walk as a Markov Process

The random walk is a Markov process. Recall that

$$X(n+1) = X(n) + Z(n+1),$$

where $Z(n+1)$ is independent of

$$X(1), X(2), \dots, X(n).$$

Therefore, for any possible values $x(1), \dots, x(n+1)$,

$$\Pr(X(n+1) = x(n+1) \mid X(n) = x(n), X(n-1) = x(n-1), \dots, X(1) = x(1))$$

is equal to

$$\Pr(X(n) + Z(n+1) = x(n+1) \mid X(n) = x(n), X(n-1) = x(n-1), \dots, X(1) = x(1)).$$

Given $X(n) = x(n)$, this becomes

$$\Pr(Z(n+1) = x(n+1) - x(n) \mid X(n) = x(n), X(n-1) = x(n-1), \dots, X(1) = x(1)).$$

Since $Z(n+1)$ is independent of the past, this is

$$\Pr(Z(n+1) = x(n+1) - x(n)).$$

Thus

$$\Pr(X(n+1) = x(n+1) \mid X(n) = x(n), X(n-1) = x(n-1), \dots, X(1) = x(1))$$

depends on the past only through $x(n)$. Hence the random walk is Markov.

Independent Increment Processes

A discrete-time process

$$\{X(n), n \in \mathbb{N}\}$$

is an *independent increment process* if, for every

$$n_1 < n_2 < \dots < n_k,$$

the random variables

$$X(n_1), X(n_2) - X(n_1), \dots, X(n_k) - X(n_{k-1})$$

are independent.

The idea is that the changes of the process over disjoint time intervals are independent.

The random walk is an example. Indeed, if

$$X(n) = \sum_{i=1}^n Z(i),$$

then

$$X(n_1) = \sum_{i=1}^{n_1} Z(i),$$

$$X(n_2) - X(n_1) = \sum_{i=n_1+1}^{n_2} Z(i),$$

and more generally

$$X(n_j) - X(n_{j-1}) = \sum_{i=n_{j-1}+1}^{n_j} Z(i).$$

These are sums over disjoint blocks of IID random variables, so the increments are independent.

Exercise: Independent Increments Imply Markov

An independent increment process is a Markov process.

To see the idea, write

$$X(n+1) = X(n) + (X(n+1) - X(n)).$$

The increment

$$X(n+1) - X(n)$$

is independent of the past values

$$X(1), X(2), \dots, X(n).$$

Therefore, once we know the present value $X(n)$, the distribution of $X(n+1)$ no longer depends on the earlier past. Hence

$$\Pr(X(n+1) \in A \mid X(1), \dots, X(n)) = \Pr(X(n+1) \in A \mid X(n)).$$

So the process is Markov.

Counting Processes

A continuous-time random process

$$\{N(t) : t \geq 0\}$$

is called a *counting process* if $N(0) = 0$, $N(t)$ is nonnegative and integer-valued, and

$$N(s) \leq N(t)$$

whenever $s \leq t$.

The quantity $N(t) - N(s)$ counts the number of events occurring in $(s, t]$. A typical sample path is a nondecreasing, right-continuous step function. In a general counting process, a jump may have any positive integer size. In a *simple counting process*, every jump has size one.

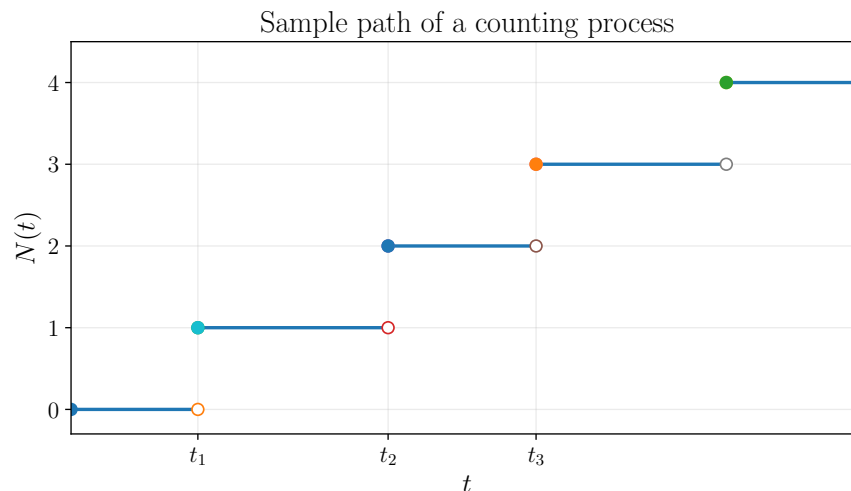


Figure 6: A sample path of a counting process. The process stays constant between arrivals and jumps when an event occurs.

Poisson Distribution Reminder

A random variable X has a Poisson distribution with parameter λ if

$$\Pr(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k = 0, 1, 2, \dots$$

In this case,

$$\mathbb{E}[X] = \lambda,$$

and

$$\text{Var}(X) = \lambda.$$

Poisson as a Law of Small Numbers

One reason the Poisson distribution appears so often is the following limiting phenomenon. Suppose

$$Y_n \sim \text{Binomial}\left(n, \frac{\lambda}{n}\right).$$

Then, for each fixed $k = 0, 1, 2, \dots$,

$$\Pr(Y_n = k) = \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}.$$

Rewrite this as

$$\Pr(Y_n = k) = \frac{n(n-1)\cdots(n-k+1)}{n^k} \frac{\lambda^k}{k!} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k}.$$

As $n \rightarrow \infty$, the first factor goes to 1, the middle factor stays as $\lambda^k/k!$, and

$$\left(1 - \frac{\lambda}{n}\right)^n \rightarrow e^{-\lambda}.$$

Also,

$$\left(1 - \frac{\lambda}{n}\right)^{-k} \rightarrow 1.$$

Therefore

$$\Pr(Y_n = k) \rightarrow e^{-\lambda} \frac{\lambda^k}{k!}.$$

Thus

$$Y_n \Rightarrow \text{Poisson}(\lambda).$$

This is sometimes called the *law of small numbers*: if we have many independent trials, each with a very small probability of success, but with the total expected number of successes remaining of order one, then the total number of successes is approximately Poisson.

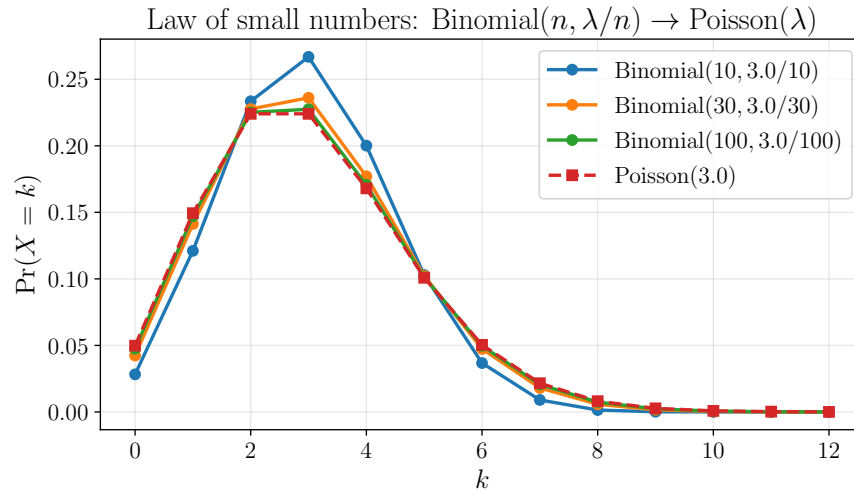


Figure 7: The binomial distribution with many trials and small success probability approaches a Poisson distribution.

Poisson Process

A counting process

$$\{N(t), t \geq 0\}$$

is a *Poisson process* with rate $\lambda > 0$ if:

$$N(0) = 0,$$

$\{N(t), t \geq 0\}$ has independent increments,

and for every $s < t$,

$$N(t) - N(s) \sim \text{Poisson}(\lambda(t - s)).$$

Thus, the number of events in an interval depends only on the length of the interval. Also, counts in disjoint intervals are independent.

The rate λ has a simple interpretation. Since

$$\mathbb{E}[N(t) - N(s)] = \lambda(t - s),$$

the process has, on average, λ arrivals per unit time. Thus λ controls the time scale of the process: larger λ means arrivals tend to occur more frequently.

The Poisson process is more than just a counting process. A general counting process only needs to be nonnegative, integer-valued, and nondecreasing. A Poisson process additionally says that counts over disjoint intervals are independent, and that the distribution of the count in an interval depends only on the length of the interval.

For

$$0 \leq t_1 \leq t_2 \leq \dots \leq t_k,$$

we can compute the joint distribution of

$$N(t_1), N(t_2), \dots, N(t_k)$$

by rewriting the event in terms of increments. If

$$0 \leq n_1 \leq n_2 \leq \dots \leq n_k,$$

then

$$\Pr(N(t_1) = n_1, N(t_2) = n_2, \dots, N(t_k) = n_k)$$

equals

$$\Pr(N(t_1) = n_1, N(t_2) - N(t_1) = n_2 - n_1, \dots, N(t_k) - N(t_{k-1}) = n_k - n_{k-1}).$$

By independent increments, this factors as

$$e^{-\lambda t_1} \frac{(\lambda t_1)^{n_1}}{n_1!} \prod_{j=2}^k e^{-\lambda(t_j - t_{j-1})} \frac{(\lambda(t_j - t_{j-1}))^{n_j - n_{j-1}}}{(n_j - n_{j-1})!}.$$

If the integers are not ordered as

$$0 \leq n_1 \leq n_2 \leq \dots \leq n_k,$$

then the probability is zero, because a counting process cannot decrease.

Arrival Time Process

Associated with any counting process is its *arrival time process* $\{T(n), n \geq 0\}$. Here $T(0) = 0$, and for $n \geq 1$, $T(n)$ is the time of the n^{th} event of the counting process $N(t)$. In particular, $T(1)$ is the time of the first arrival. If $N(t)$ is a Poisson process with rate λ , then

$$F_{T(1)}(t) = \Pr(T(1) \leq t).$$

The event $T(1) > t$ means that no arrival has occurred by time t , so

$$\Pr(T(1) > t) = \Pr(N(t) = 0).$$

Since $N(t) \sim \text{Poisson}(\lambda t)$, we get

$$\Pr(N(t) = 0) = e^{-\lambda t}.$$

Therefore

$$F_{T(1)}(t) = 1 - e^{-\lambda t}, \quad t \geq 0.$$

Thus

$$T(1) \sim \text{Exp}(\lambda).$$

Inter-Arrival Times

Define the inter-arrival time process by

$$X(n) = T(n) - T(n-1), \quad n \geq 1.$$

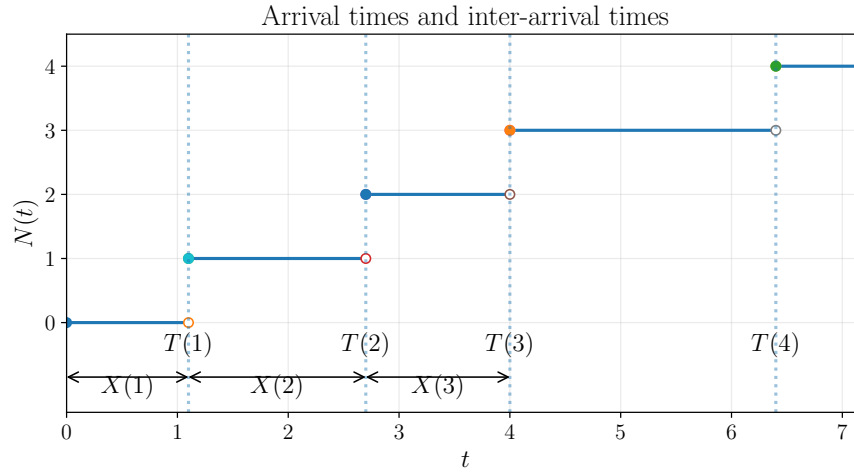


Figure 8: Arrival times $T(n)$ and inter-arrival times $X(n) = T(n) - T(n-1)$ for a counting process.

Thus $X(1)$ is the waiting time until the first arrival, $X(2)$ is the waiting time from the first arrival to the second arrival, and so on. For a Poisson process with rate λ , the inter-arrival times are IID exponential random variables:

$$X(1), X(2), \dots \stackrel{\text{iid}}{\sim} \text{Exp}(\lambda).$$

The intuition is that after an arrival occurs, the process starts afresh. Because the Poisson process has independent increments and the distribution of future counts depends only on the length of the future interval, the waiting time until the next arrival has the same distribution as the waiting time until the first arrival. Since

$$T(n) = \sum_{i=1}^n X(i),$$

the n^{th} arrival time is a sum of n IID exponential random variables. Therefore

$$T(n) \sim \text{Gamma}(\lambda, n),$$

with density

$$f_{T(n)}(t) = \frac{\lambda^n t^{n-1}}{(n-1)!} e^{-\lambda t}, \quad t \geq 0.$$

More generally, for $n_2 > n_1$,

$$T(n_2) - T(n_1) = \sum_{i=n_1+1}^{n_2} X(i) \sim \text{Gamma}(\lambda, n_2 - n_1).$$

Poisson Process as a Limit of Binomial Processes

Another way to understand the Poisson process is as a limit of finer and finer discrete-time counting processes. Fix a small time step $0 < \Delta \leq 1/\lambda$, and allow arrivals only at the grid times $\Delta, 2\Delta, 3\Delta, \dots$. At each grid time, independently, an arrival occurs with probability $\lambda\Delta$.

For fixed $s < t$, define the exact number of grid points in $(s, t]$ by

$$m_{\Delta}(s, t) = \left\lfloor \frac{t}{\Delta} \right\rfloor - \left\lfloor \frac{s}{\Delta} \right\rfloor.$$

It follows that

$$N_{\Delta}(t) - N_{\Delta}(s) \sim \text{Binomial}(m_{\Delta}(s, t), \lambda\Delta).$$

As $\Delta \rightarrow 0$,

$$m_{\Delta}(s, t)\lambda\Delta \rightarrow \lambda(t - s),$$

and therefore

$$N_{\Delta}(t) - N_{\Delta}(s) \Rightarrow \text{Poisson}(\lambda(t - s)).$$

Thus the Poisson process can be viewed as the continuous-time limit of a process where many tiny time slots each independently have a small probability of an arrival. This viewpoint explains why the Poisson process is such a natural model for arrivals in many settings: photons arriving at a detector, packets arriving in a network, customers arriving at a queue, and similar systems where there are many small independent opportunities for an event to occur.

Binomial-process viewpoint: arrivals on a finer and finer time grid

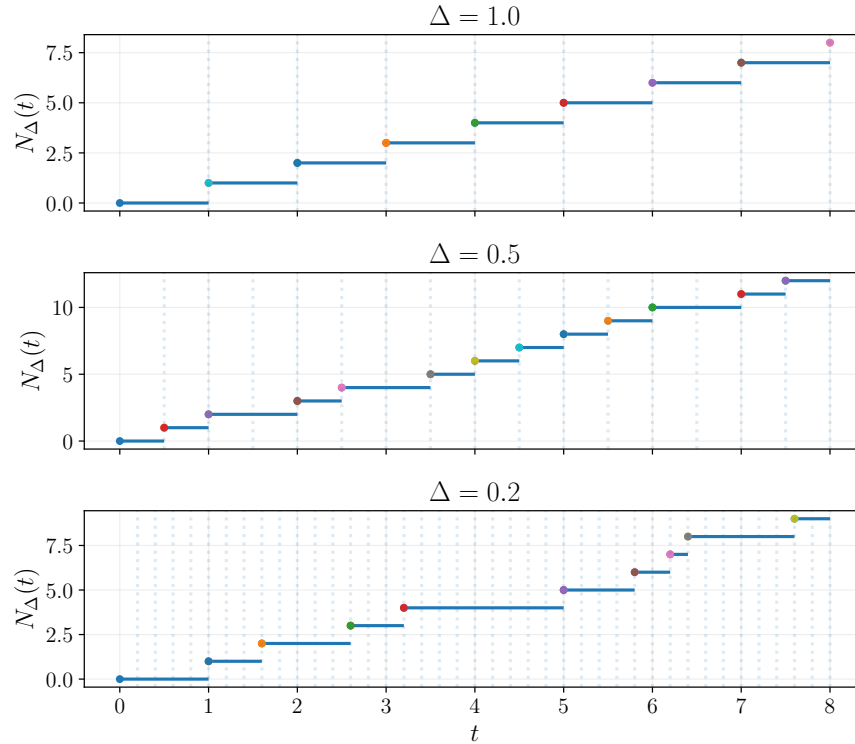


Figure 9: A binomial-process viewpoint: arrivals can occur only on a time grid, and as the grid becomes finer the process approaches a Poisson process.

Mean and Autocorrelation Functions

A full specification of a random process requires all finite-dimensional distributions. However, in many problems, especially linear estimation problems, first- and second-order information is already very useful. This motivates the mean and autocorrelation functions of a process.

For a random process

$$\{X(t), t \in \mathcal{T}\},$$

the *mean function* is

$$\mu_X(t) = \mathbb{E}[X(t)].$$

The *autocorrelation function* is

$$R_X(t_1, t_2) = \mathbb{E}[X(t_1)X(t_2)].$$

The *autocovariance function* is

$$C_X(t_1, t_2) = \text{Cov}(X(t_1), X(t_2)).$$

Thus

$$C_X(t_1, t_2) = R_X(t_1, t_2) - \mu_X(t_1)\mu_X(t_2).$$

In general, the mean and autocorrelation functions do not fully specify a random process. However, they are often enough for linear estimation problems, since linear estimators depend only on first- and second-order information. For Gaussian random processes, these functions will in fact give a complete specification.

Example: Random Phase Cosine

Recall the process

$$X(t) = \cos(t + \Theta),$$

where

$$\Theta \sim \text{Unif}[0, 2\pi].$$

The mean function is

$$\mu_X(t) = \mathbb{E}[\cos(t + \Theta)] = \frac{1}{2\pi} \int_0^{2\pi} \cos(t + \theta) d\theta = 0.$$

The autocorrelation function is

$$R_X(t_1, t_2) = \mathbb{E}[\cos(t_1 + \Theta) \cos(t_2 + \Theta)].$$

Using

$$\cos a \cos b = \frac{1}{2} \cos(a - b) + \frac{1}{2} \cos(a + b),$$

we get

$$R_X(t_1, t_2) = \frac{1}{2} \cos(t_1 - t_2).$$

Since the mean is zero, the autocovariance function is the same:

$$C_X(t_1, t_2) = \frac{1}{2} \cos(t_1 - t_2).$$

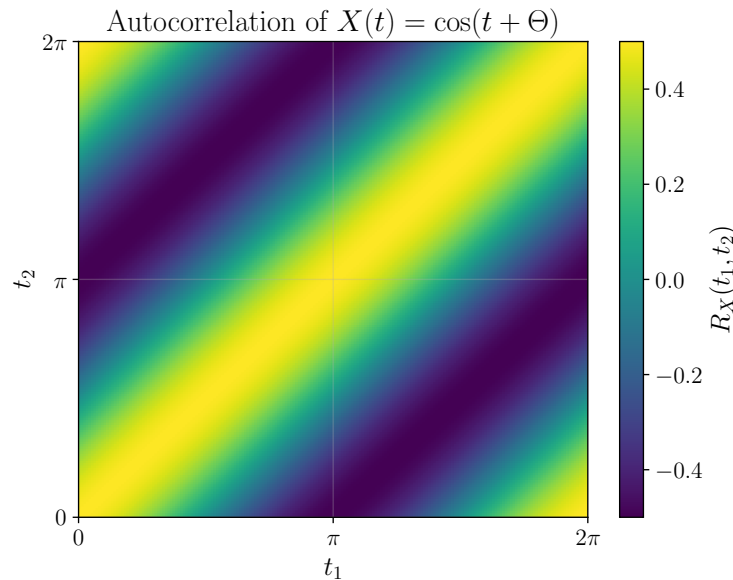


Figure 10: Autocorrelation function $R_X(t_1, t_2) = \frac{1}{2} \cos(t_1 - t_2)$ for the random phase cosine process.

Example: Poisson Process

Let

$$\{N(t), t \geq 0\}$$

be a Poisson process with rate λ . Since

$$N(t) \sim \text{Poisson}(\lambda t),$$

the mean function is

$$\mu_N(t) = \mathbb{E}[N(t)] = \lambda t.$$

Now assume $t_1 \leq t_2$. Then

$$N(t_2) = N(t_1) + (N(t_2) - N(t_1)),$$

so

$$R_N(t_1, t_2) = \mathbb{E}[N(t_1)N(t_2)] = \mathbb{E}[N(t_1)^2] + \mathbb{E}[N(t_1)(N(t_2) - N(t_1))].$$

The two increments $N(t_1)$ and $N(t_2) - N(t_1)$ are independent. Also,

$$N(t_1) \sim \text{Poisson}(\lambda t_1),$$

so

$$\mathbb{E}[N(t_1)^2] = \text{Var}(N(t_1)) + \mathbb{E}[N(t_1)]^2 = \lambda t_1 + \lambda^2 t_1^2.$$

Therefore

$$R_N(t_1, t_2) = \lambda t_1 + \lambda^2 t_1^2 + (\lambda t_1)(\lambda(t_2 - t_1)).$$

Thus

$$R_N(t_1, t_2) = \lambda t_1 + \lambda^2 t_1 t_2.$$

For general $t_1, t_2 \geq 0$,

$$R_N(t_1, t_2) = \lambda \min(t_1, t_2) + \lambda^2 t_1 t_2.$$

Since

$$\mu_N(t_1)\mu_N(t_2) = \lambda^2 t_1 t_2,$$

the autocovariance function is

$$C_N(t_1, t_2) = \lambda \min(t_1, t_2).$$

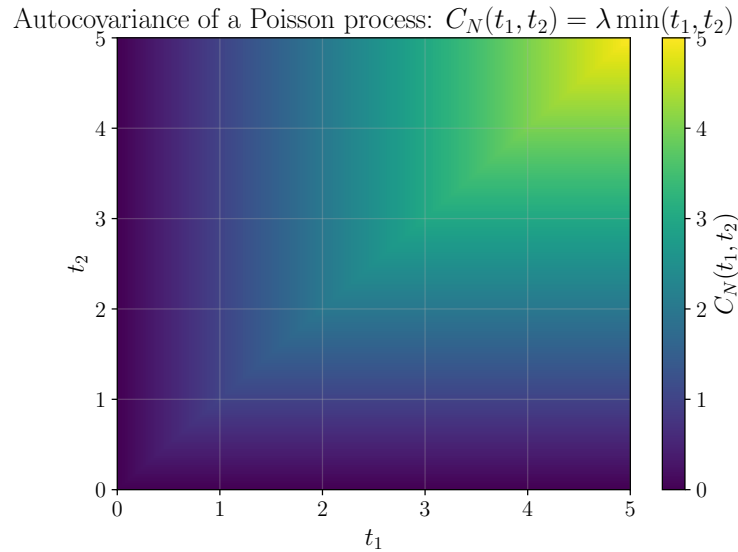


Figure 11: Autocovariance function $C_N(t_1, t_2) = \lambda \min(t_1, t_2)$ for a Poisson process.

Gaussian Random Processes

A random process

$$\{X(t), t \in \mathcal{T}\}$$

is a *Gaussian random process* if, for every finite collection of time points

$$t_1, \dots, t_k \in \mathcal{T},$$

the random vector

$$(X(t_1), \dots, X(t_k))$$

is jointly Gaussian.

Equivalently, every finite linear combination

$$\sum_{i=1}^k a_i X(t_i)$$

is a Gaussian random variable.

A Gaussian random process is completely specified by its mean function and autocovariance function. This is because every finite-dimensional distribution is Gaussian, and a Gaussian random vector is determined by its mean vector and covariance matrix.

Example: Gauss-Markov Process

Let

$$Z(1), Z(2), \dots$$

be IID Gaussian random variables with

$$Z(n) \sim \mathcal{N}(0, \sigma^2).$$

This is often called *white Gaussian noise*.

Define a process by

$$X(1) = Z(1),$$

and for $n \geq 2$,

$$X(n) = \alpha X(n-1) + Z(n),$$

where α is a fixed deterministic constant.

This process is called a *Gauss-Markov process*. It is Gaussian because every $X(n)$ is a linear combination of the Gaussian variables

$$Z(1), Z(2), \dots, Z(n).$$

Therefore any finite linear combination of

$$X(1), X(2), \dots, X(k)$$

is also a linear combination of Gaussian random variables, and hence is Gaussian.

It is also Markov. The recursion

$$X(n) = \alpha X(n-1) + Z(n)$$

shows that, given $X(n-1)$, the next value $X(n)$ depends on the past only through $X(n-1)$ and fresh noise $Z(n)$, which is independent of the past.

Thus this process lies at the intersection of two important families: Gaussian processes and Markov processes.

Lecture 11: Random Processes II

We continue our discussion of random processes. Today we revisit Poisson processes and their arrival-time description, mean and autocorrelation functions, and then use these ideas to define Gaussian random processes. *This lecture roughly follows Lecture Notes 6 of Abbas El Gamal.*

Reminder: Binomial Distribution

Recall that

$$X \sim \text{Binomial}(n, q)$$

means that X counts the number of successes in n independent trials, where each trial has success probability q .

Thus

$$\Pr(X = k) = \binom{n}{k} q^k (1 - q)^{n-k}, \quad 0 \leq k \leq n.$$

Also,

$$\mathbb{E}[X] = nq.$$

Now consider

$$Y_n \sim \text{Binomial}\left(n, \frac{\lambda}{n}\right).$$

Then

$$\mathbb{E}[Y_n] = n \frac{\lambda}{n} = \lambda.$$

For fixed k ,

$$\Pr(Y_n = k) = \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}.$$

Expanding,

$$\Pr(Y_n = k) = \frac{n!}{(n-k)!k!} \frac{\lambda^k}{n^k} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k}.$$

Equivalently,

$$\Pr(Y_n = k) = \frac{n(n-1)\cdots(n-k+1)}{n^k} \frac{\lambda^k}{k!} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k}.$$

As $n \rightarrow \infty$,

$$\frac{n(n-1)\cdots(n-k+1)}{n^k} \rightarrow 1, \text{ so } \left(1 - \frac{\lambda}{n}\right)^n \rightarrow e^{-\lambda}, \left(1 - \frac{\lambda}{n}\right)^{-k} \rightarrow 1.$$

Therefore

$$\Pr(Y_n = k) \rightarrow e^{-\lambda} \frac{\lambda^k}{k!}.$$

So, $Y_n \Rightarrow \text{Poisson}(\lambda)$.

This limiting relationship is one of the main reasons the Poisson distribution appears naturally: many independent trials, each with very small success probability, produce a Poisson number of successes when the total expected number of successes remains fixed.

Counting Processes

A continuous-time random process

$$\{N(t), t \geq 0\}$$

is called a *counting process* if

$$N(t) \geq 0,$$

$N(t)$ is integer-valued,

and if $s \leq t$, then $N(s) \leq N(t)$.

The interpretation is that $N(t)$ counts the number of events or arrivals that have occurred up to time t . For $s < t$, $N(t) - N(s)$ is the number of events that occurred during the interval $(s, t]$.

A sample path of a counting process is a nondecreasing step function. It stays constant between arrivals and jumps up when an arrival occurs.

Poisson Process

A counting process $\{N(t), t \geq 0\}$ is a *Poisson process* with rate $\lambda > 0$ if:

- $N(0) = 0$,
- $\{N(t), t \geq 0\}$ has independent increments,
- and for every $s < t$,

$$N(t) - N(s) \sim \text{Poisson}(\lambda(t - s)).$$

Thus, counts in disjoint intervals are independent, and the distribution of the count in an interval depends only on the length of that interval.

Since a Poisson random variable has mean equal to its parameter,

$$\mathbb{E}[N(t) - N(s)] = \lambda(t - s).$$

Thus λ is the average number of arrivals per unit time.

Joint Distributions from Independent Increments

For $0 \leq t_1 \leq t_2 \leq \dots \leq t_k$, we can compute the joint distribution of $N(t_1), N(t_2), \dots, N(t_k)$, by rewriting everything in terms of increments.

If $0 \leq n_1 \leq n_2 \leq \dots \leq n_k$, then $\Pr(N(t_1) = n_1, N(t_2) = n_2, \dots, N(t_k) = n_k) = \Pr(N(t_1) = n_1, N(t_2) - N(t_1) = n_2 - n_1, \dots, N(t_k) - N(t_{k-1}) = n_k - n_{k-1})$. By using independent increments, this factors as

$$e^{-\lambda t_1} \frac{(\lambda t_1)^{n_1}}{n_1!} \prod_{j=2}^k e^{-\lambda(t_j - t_{j-1})} \frac{(\lambda(t_j - t_{j-1}))^{n_j - n_{j-1}}}{(n_j - n_{j-1})!}.$$

If the integers are not ordered as $0 \leq n_1 \leq n_2 \leq \dots \leq n_k$, then the probability is zero, since a counting process cannot decrease.

Arrival Time Process

For any counting process, we can define the associated *arrival time process*

$$\{T(n), n \geq 0\}.$$

Here $T(0) = 0$, and for $n \geq 1$, $T(n)$ is the time of the n^{th} arrival.

For a Poisson process with rate λ , the first arrival time satisfies

$$F_{T(1)}(t) = \Pr(T(1) \leq t).$$

The event $T(1) > t$ means that no arrival occurred by time t . Hence

$$\Pr(T(1) > t) = \Pr(N(t) = 0).$$

Since

$$N(t) \sim \text{Poisson}(\lambda t),$$

we have

$$\Pr(N(t) = 0) = e^{-\lambda t}.$$

Therefore

$$F_{T(1)}(t) = 1 - e^{-\lambda t}, \quad t \geq 0.$$

Thus

$$T(1) \sim \text{Exp}(\lambda).$$

Inter-Arrival Time Process

The *inter-arrival time process* is defined by

$$X(n) = T(n) - T(n-1), \quad n \geq 1.$$

Thus $X(n)$ is the amount of time between the $(n-1)$ st arrival and the n th arrival.

When $N(t)$ is a Poisson process with rate λ , $X(1), X(2), \dots$ are IID exponential random variables:

$$X(n) \sim \text{Exp}(\lambda).$$

The intuition is that after each arrival, the process starts afresh. The future arrivals are independent of the past, and their distribution depends only on the future time duration.

Since

$$T(n) = \sum_{i=1}^n X(i),$$

the arrival time process is an independent increment process.

For $n_2 > n_1 \geq 0$,

$$T(n_2) - T(n_1) = \sum_{i=n_1+1}^{n_2} X(i).$$

This is a sum of $n_2 - n_1$ IID exponential random variables. Therefore

$$T(n_2) - T(n_1) \sim \text{Gamma}(\lambda, n_2 - n_1),$$

where we use the rate-parameter convention. Its density is

$$f_{T(n_2)-T(n_1)}(t) = \frac{\lambda^{n_2-n_1} t^{n_2-n_1-1}}{(n_2-n_1-1)!} e^{-\lambda t}, \quad t \geq 0.$$

In particular,

$$T(n) \sim \text{Gamma}(\lambda, n).$$

Poisson Process as a Limit of Binomial Processes

Another way to understand the Poisson process is as a limit of discrete-time binomial counting processes.

Fix a small time step $\Delta > 0$. Define a counting process $\{N_\Delta(t), t \geq 0\}$ where arrivals can occur only at times $\Delta, 2\Delta, 3\Delta, \dots$.

At each such time point, independently,

$$\Pr(\text{one arrival}) = \lambda\Delta + o(\Delta),$$

$$\Pr(\text{no arrival}) = 1 - \lambda\Delta + o(\Delta),$$

and

$$\Pr(\text{more than one arrival}) = o(\Delta).$$

In the simplest version, at each grid point we take

$$\Pr(\text{one arrival}) = \lambda\Delta,$$

and

$$\Pr(\text{no arrival}) = 1 - \lambda\Delta.$$

For each fixed Δ , $N_\Delta(0) = 0$, and $N_\Delta(t)$ has independent increments, because disjoint time intervals contain disjoint collections of independent Bernoulli trials.

For $s < t$, the number of grid points in $(s, t]$ is approximately $\frac{t-s}{\Delta}$. Therefore, $N_\Delta(t) - N_\Delta(s)$ is approximately Binomial $(\frac{t-s}{\Delta}, \lambda\Delta)$.

Using the Binomial-to-Poisson limit,

$$\Pr(N_\Delta(t) - N_\Delta(s) = k) \rightarrow e^{-\lambda(t-s)} \frac{(\lambda(t-s))^k}{k!}.$$

Thus, as $\Delta \rightarrow 0$,

$$N_\Delta(t) - N_\Delta(s) \Rightarrow \text{Poisson}(\lambda(t-s)).$$

So as the time grid becomes finer and finer, the binomial counting process converges to a Poisson process with rate λ .

Mean and Autocorrelation Functions

For a random process

$$\{X(t), t \in \mathcal{T}\},$$

the *mean function* is

$$\mu_X(t) = \mathbb{E}[X(t)].$$

The *autocorrelation function* is

$$R_X(t_1, t_2) = \mathbb{E}[X(t_1)X(t_2)].$$

The *autocovariance function* is

$$C_X(t_1, t_2) = \text{Cov}(X(t_1), X(t_2)).$$

Therefore

$$C_X(t_1, t_2) = R_X(t_1, t_2) - \mu_X(t_1)\mu_X(t_2).$$

The word “auto” emphasizes that we are correlating the process with itself at two possibly different time points.

Example 1: Random Phase Cosine

Let

$$X(t) = \cos(t + \Theta),$$

where

$$\Theta \sim \text{Unif}[0, 2\pi].$$

The mean function is

$$\mu_X(t) = \mathbb{E}[\cos(t + \Theta)].$$

Since Θ is uniform on $[0, 2\pi]$,

$$\mu_X(t) = \frac{1}{2\pi} \int_0^{2\pi} \cos(t + \theta) d\theta = 0.$$

Now compute the autocorrelation function:

$$R_X(t_1, t_2) = \mathbb{E}[\cos(t_1 + \Theta) \cos(t_2 + \Theta)].$$

Thus

$$R_X(t_1, t_2) = \frac{1}{2\pi} \int_0^{2\pi} \cos(t_1 + \theta) \cos(t_2 + \theta) d\theta.$$

Using

$$\cos a \cos b = \frac{1}{2} \cos(a - b) + \frac{1}{2} \cos(a + b),$$

we get

$$R_X(t_1, t_2) = \frac{1}{2\pi} \int_0^{2\pi} \left[\frac{1}{2} \cos(t_1 - t_2) + \frac{1}{2} \cos(t_1 + t_2 + 2\theta) \right] d\theta.$$

The second term integrates to zero over $[0, 2\pi]$, so

$$R_X(t_1, t_2) = \frac{1}{2} \cos(t_1 - t_2).$$

Since $\mu_X(t) = 0$,

$$C_X(t_1, t_2) = R_X(t_1, t_2) = \frac{1}{2} \cos(t_1 - t_2).$$

Example 2: Poisson Process

Let $\{N(t), t \geq 0\}$ be a Poisson process with rate λ . Since

$$N(t) - N(0) \sim \text{Poisson}(\lambda t),$$

we have

$$\mu_N(t) = \mathbb{E}[N(t)] = \lambda t.$$

Now assume $t_1 \leq t_2$. Then

$$N(t_2) = N(t_1) + (N(t_2) - N(t_1)).$$

Therefore

$$R_N(t_1, t_2) = \mathbb{E}[N(t_1)N(t_2)]$$

equals

$$\mathbb{E}[N(t_1)(N(t_1) + N(t_2) - N(t_1))].$$

So

$$R_N(t_1, t_2) = \mathbb{E}[N(t_1)^2] + \mathbb{E}[N(t_1)(N(t_2) - N(t_1))].$$

The two random variables $N(t_1)$ and $N(t_2) - N(t_1)$ are independent. Also,

$$N(t_1) \sim \text{Poisson}(\lambda t_1).$$

Thus

$$\mathbb{E}[N(t_1)^2] = \text{Var}(N(t_1)) + \mathbb{E}[N(t_1)]^2 = \lambda t_1 + \lambda^2 t_1^2.$$

And

$$\mathbb{E}[N(t_1)(N(t_2) - N(t_1))] = (\lambda t_1)(\lambda(t_2 - t_1)).$$

Therefore

$$R_N(t_1, t_2) = \lambda t_1 + \lambda^2 t_1^2 + \lambda^2 t_1(t_2 - t_1).$$

So

$$R_N(t_1, t_2) = \lambda t_1 + \lambda^2 t_1 t_2.$$

For general $t_1, t_2 \geq 0$,

$$R_N(t_1, t_2) = \lambda \min\{t_1, t_2\} + \lambda^2 t_1 t_2.$$

Since

$$\mu_N(t_1)\mu_N(t_2) = \lambda^2 t_1 t_2,$$

the autocovariance function is

$$C_N(t_1, t_2) = \lambda \min\{t_1, t_2\}.$$

Gaussian Random Processes

A random process $\{X(t), t \in \mathcal{T}\}$ is called a *Gaussian random process* if, for every k and every choice of time points $t_1, t_2, \dots, t_k \in \mathcal{T}$, the random vector

$$\begin{bmatrix} X(t_1) \\ X(t_2) \\ \vdots \\ X(t_k) \end{bmatrix}$$

is a Gaussian random vector.

Equivalently, every finite linear combination of process values is Gaussian. That is, for any $a_1, \dots, a_k \in \mathbb{R}$,

$$\sum_{i=1}^k a_i X(t_i)$$

is a Gaussian random variable.

A Gaussian random process is specified by its mean and autocorrelation functions. Indeed, for any finite collection $t_1, \dots, t_k$, the vector $(X(t_1), \dots, X(t_k))^T$ is Gaussian, so its distribution is determined by its mean vector and covariance matrix. The mean vector is determined by $\mu_X(t)$, and the covariance matrix is determined by $C_X(t_i, t_j)$, or equivalently by μ_X and R_X .

Example: Gauss-Markov Process

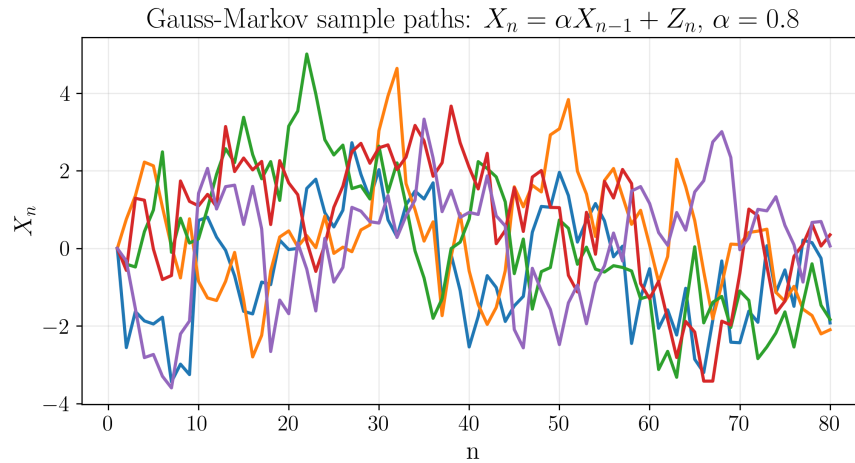


Figure 12: Sample paths of the Gauss-Markov process $X_n = \alpha X_{n-1} + Z_n$.

Let $\{Z_n, n \geq 1\}$ be white Gaussian noise. That is, $Z_1, Z_2, \dots$ are IID and

$$Z_n \sim \mathcal{N}(0, \sigma^2).$$

Define the process $\{X_n, n \geq 1\}$ by

$$X_1 = Z_1,$$

and for $n > 1$,

$$X_n = \alpha X_{n-1} + Z_n,$$

where α is a deterministic constant.

This is called a *Gauss-Markov process*.

Exercise

Show that $\{X_n, n \geq 1\}$ is: (a) a Gaussian random process, and (b) a Markov process.

An extension of this also appears in HW6.

Lecture 12: Stationarity and Power Spectral Density

These notes correspond closely to Lecture Note 7 of Abbas El Gamal (AEG).

Recap: Mean and Autocorrelation Functions

Let $X = \{X(t), -\infty < t < \infty\}$ be a continuous-time random process. The *mean function* and *autocorrelation function* are

$$\mu_X(t) = \mathbb{E}[X(t)], \quad R_X(t_1, t_2) = \mathbb{E}[X(t_1)X(t_2)].$$

Example: Random Phase Cosine

Let $X(t) = \cos(t + \Theta)$, where $\Theta \sim \text{Unif}[0, 2\pi)$. Then

$$\mu_X(t) = \mathbb{E}[\cos(t + \Theta)] = 0.$$

Also,

$$\begin{aligned} R_X(t_1, t_2) &= \mathbb{E}[\cos(t_1 + \Theta) \cos(t_2 + \Theta)] \\ &= \frac{1}{2} \cos(t_1 - t_2). \end{aligned}$$

Thus the autocorrelation depends only on the time difference $t_1 - t_2$.

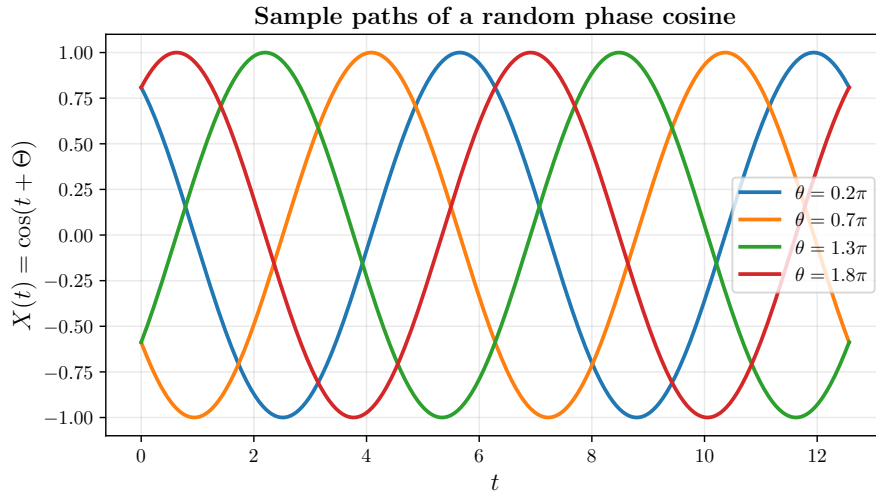


Figure 13: Sample paths of the random phase cosine process $X(t) = \cos(t + \Theta)$ for several fixed values of the phase Θ . Once the phase is chosen, the entire sample path is deterministic.

Stationarity

We write $X = \{X(t), -\infty < t < \infty\}$ for a continuous-time random process, or $X = \{X(n), -\infty < n < \infty\}$ for a discrete-time random process. For random variables U and V , we write $U \stackrel{d}{=} V$ if they have the same distribution, equivalently $F_U = F_V$.

Definition: Stationarity

A random process X is *stationary* if, for every k , every $t_1, \dots, t_k$, and every shift s ,

$$(X(t_1), X(t_2), \dots, X(t_k)) \stackrel{d}{=} (X(t_1 + s), X(t_2 + s), \dots, X(t_k + s)).$$

That is, shifting time does not change the joint distribution of any finite collection of samples.

Example: Random Phase Cosine is Stationary

Again let $X(t) = \cos(t + \Theta)$, where $\Theta \sim \text{Unif}[0, 2\pi)$. For any $t_1, \dots, t_k$ and shift s ,

$$(X(t_1 + s), \dots, X(t_k + s)) = (\cos(t_1 + s + \Theta), \dots, \cos(t_k + s + \Theta)).$$

Let $\Theta' = (\Theta + s) \bmod 2\pi$. Since Θ is uniform on $[0, 2\pi)$, we also have $\Theta' \stackrel{d}{=} \Theta$. Therefore,

$$(X(t_1 + s), \dots, X(t_k + s)) \stackrel{d}{=} (\cos(t_1 + \Theta), \dots, \cos(t_k + \Theta)) = (X(t_1), \dots, X(t_k)).$$

Hence $X(t) = \cos(t + \Theta)$ is stationary.

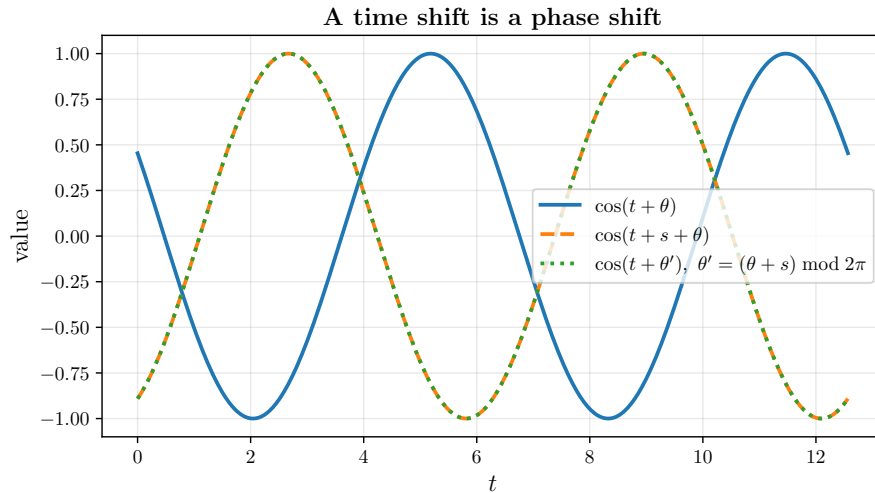


Figure 14: A time shift of the random phase cosine is equivalent to a phase shift. Since $(\Theta + s) \bmod 2\pi$ has the same distribution as Θ when $\Theta \sim \text{Unif}[0, 2\pi)$, the finite-dimensional distributions are unchanged by time shifts.

Wide-Sense Stationarity

A random process X is *wide-sense stationary*, abbreviated WSS, if

$$\mathbb{E}[X(t)] = \mu \text{ for all } t, \quad R_X(t, t + s) \text{ is a function of } s \text{ only}, \quad \mathbb{E}[X(t)^2] < \infty.$$

For a WSS process, we write

$$R_X(\tau) = R_X(t, t + \tau) = \mathbb{E}[X(t)X(t + \tau)].$$

Properties of the Autocorrelation Function

For a WSS process, the autocorrelation function satisfies the following properties.

1. Evenness

$$R_X(\tau) = R_X(-\tau).$$

Indeed,

$$R_X(\tau) = \mathbb{E}[X(t)X(t+\tau)] = \mathbb{E}[X(t+\tau)X(t)] = R_X(-\tau),$$

where the last equality uses WSS.

2. Maximum at Zero

$$R_X(0) \geq |R_X(\tau)|.$$

This follows from Cauchy-Schwarz. For any random variables X and Y ,

$$(\mathbb{E}[XY])^2 \leq \mathbb{E}[X^2]\mathbb{E}[Y^2].$$

Taking $X = X(t)$ and $Y = X(t+\tau)$ gives

$$\begin{aligned} (R_X(\tau))^2 &= (\mathbb{E}[X(t)X(t+\tau)])^2 \\ &\leq \mathbb{E}[X(t)^2]\mathbb{E}[X(t+\tau)^2] \\ &= R_X(0)^2. \end{aligned}$$

Hence $|R_X(\tau)| \leq R_X(0)$.

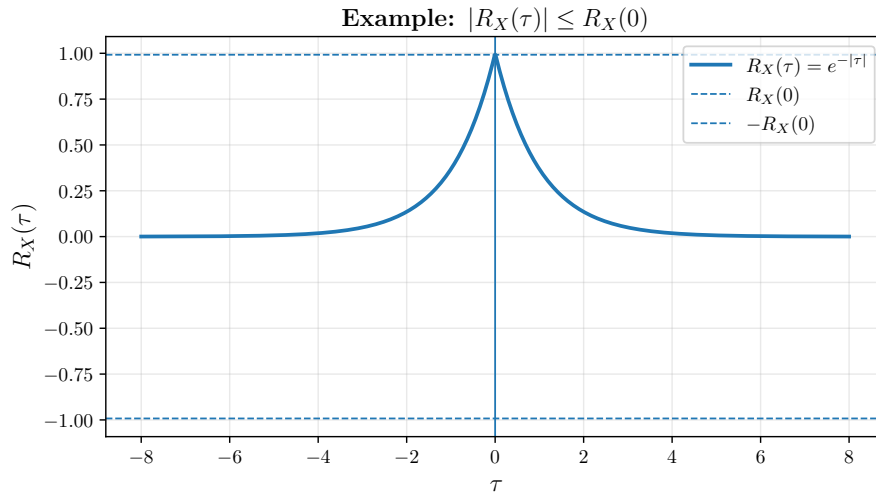


Figure 15: Example of a valid autocorrelation function illustrating the bound $|R_X(\tau)| \leq R_X(0)$. The value at zero equals the average power and is the largest possible autocorrelation magnitude.

3. Nonnegative Definiteness

For every n , every $t_1, \dots, t_n$, and every $a_1, \dots, a_n$,

$$\sum_{i=1}^n \sum_{j=1}^n a_i a_j R_X(t_i - t_j) \geq 0.$$

To see this, define

$$Y = \begin{bmatrix} X(t_1) \\ \vdots \\ X(t_n) \end{bmatrix}.$$

Then

$$\Sigma_Y(i, j) = \mathbb{E}[X(t_i)X(t_j)] = R_X(t_j - t_i).$$

Therefore,

$$\sum_{i=1}^n \sum_{j=1}^n a_i a_j R_X(t_j - t_i) = a^T \Sigma_Y a = \mathbb{E}[(a^T Y)^2] \geq 0.$$

Thus R_X is nonnegative definite.

When the Autocorrelation Achieves Its Maximum Away From Zero

Suppose that for some $T \neq 0$,

$$R_X(T) = R_X(0).$$

We claim that $R_X(\tau)$ is periodic with period T . We also claim that, for every fixed t ,

$$X(t) = X(t + T)$$

almost surely.

For any WSS process and any τ ,

$$\begin{aligned} (R_X(\tau) - R_X(\tau + T))^2 &= \left(\mathbb{E}[X(t)X(t + \tau)] - \mathbb{E}[X(t)X(t + \tau + T)] \right)^2 \\ &= \left(\mathbb{E}[X(t)(X(t + \tau) - X(t + \tau + T))] \right)^2 \\ &\leq \mathbb{E}[X(t)^2] \mathbb{E}[(X(t + \tau) - X(t + \tau + T))^2] \\ &= R_X(0)(2R_X(0) - 2R_X(T)). \end{aligned}$$

If $R_X(T) = R_X(0)$, then the right-hand side is 0, so

$$R_X(\tau) = R_X(\tau + T) \quad \text{for all } \tau.$$

Also,

$$\mathbb{E}[(X(t) - X(t + T))^2] = 2R_X(0) - 2R_X(T) = 0,$$

Therefore, for every fixed t ,

$$X(t) = X(t + T)$$

almost surely. This pointwise almost-sure statement does not by itself imply the existence of a single probability-one event on which the equality holds simultaneously for every real t . Such a sample-path statement requires additional regularity assumptions.

For the random phase cosine $X(t) = \cos(t + \Theta)$, we found $R_X(\tau) = \frac{1}{2} \cos \tau$ (cf. Figure 16). Indeed, $X(t) = X(t + 2\pi)$ with probability 1.

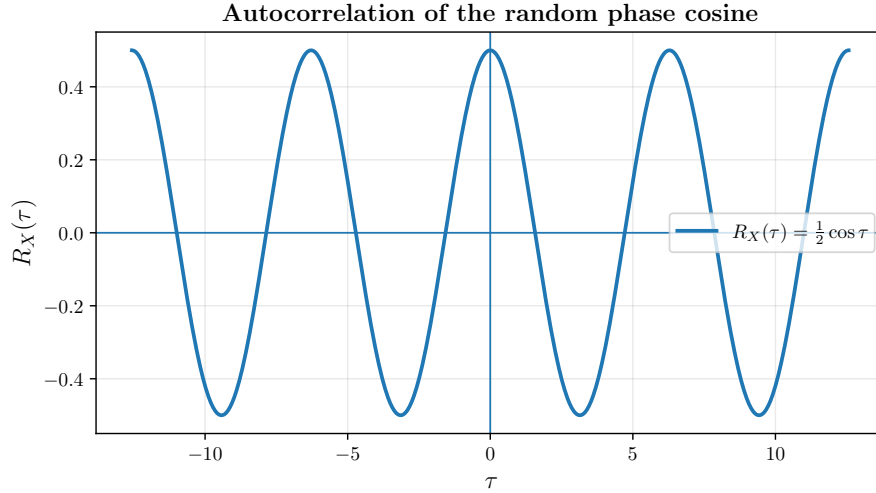


Figure 16: Autocorrelation function of the random phase cosine process. For $X(t) = \cos(t + \Theta)$ with $\Theta \sim \text{Unif}[0, 2\pi)$, the autocorrelation is $R_X(\tau) = \frac{1}{2} \cos \tau$, which is periodic in τ .

Power Spectral Density

The *power spectral density*, or PSD, of a WSS process is the Fourier transform of its autocorrelation function.

Continuous Time

For continuous time,

$$S_X(f) = \mathcal{F}\{R_X(\tau)\} = \int_{-\infty}^{\infty} R_X(\tau) e^{-i2\pi f \tau} d\tau.$$

Conversely,

$$R_X(\tau) = \int_{-\infty}^{\infty} S_X(f) e^{i2\pi f \tau} df.$$

For the random phase cosine, we have

$$R_X(\tau) = \frac{1}{2} \cos \tau.$$

Under the Fourier convention used here, this corresponds to spectral lines at

$$f = \pm \frac{1}{2\pi}.$$

Discrete Time

For discrete time,

$$S_X(f) = \sum_{n=-\infty}^{\infty} R_X(n) e^{-i2\pi n f}, \quad -\frac{1}{2} \leq f < \frac{1}{2}.$$

Conversely,

$$R_X(n) = \int_{-1/2}^{1/2} S_X(f) e^{i2\pi n f} df.$$

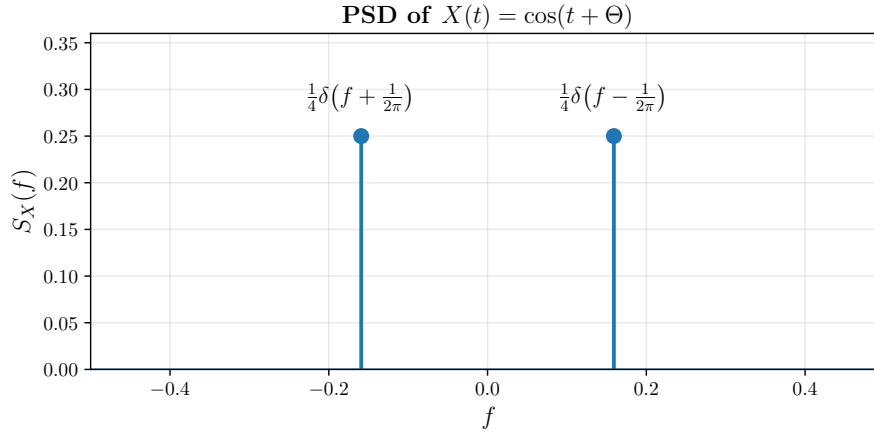


Figure 17: PSD of the random phase cosine process. Since $R_X(\tau) = \frac{1}{2} \cos \tau$, the spectrum consists of impulses at $f = \pm \frac{1}{2\pi}$ under the Fourier convention $e^{-i2\pi f\tau}$.

Properties of the PSD

For a WSS process, the PSD has the following properties.

1. Real and Even

For a real-valued WSS process, $S_X(f)$ is real and even, i.e., $S_X(f) = S_X(-f)$.

2. Total Power

In continuous time,

$$\int_{-\infty}^{\infty} S_X(f) df = R_X(0) = \mathbb{E}[X(t)^2].$$

Thus the area under the PSD equals the average power of the process.

3. Output Power Through an Ideal LTI Filter

If a WSS process X is passed through an LTI filter, the output process Y is also WSS. For an ideal bandpass filter passing frequencies $f_1 \leq |f| \leq f_2$, we will see that

$$\mathbb{E}[Y(t)^2] = 2 \int_{f_1}^{f_2} S_X(f) df.$$

The factor of 2 comes from the positive and negative frequency bands.

4. Nonnegativity

$$S_X(f) \geq 0.$$

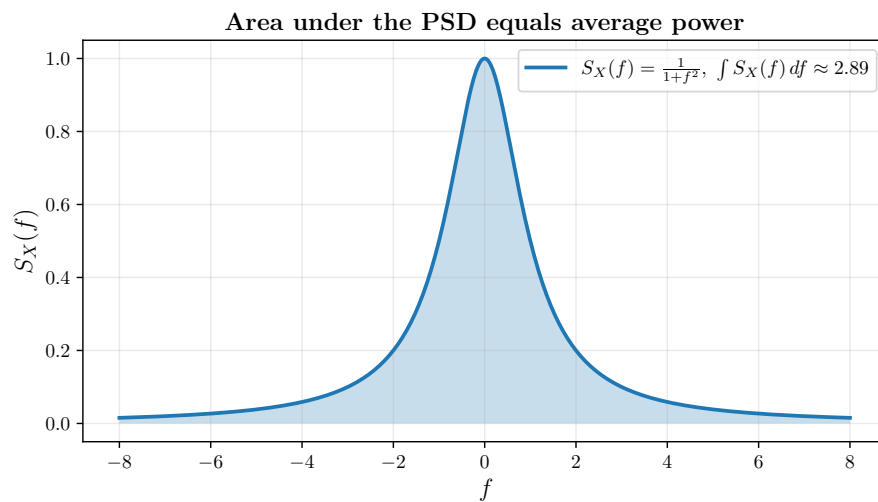


Figure 18: The area under the power spectral density equals the average power of the WSS process. In continuous time, $\int_{-\infty}^{\infty} S_X(f) df = R_X(0) = \mathbb{E}[X(t)^2]$.

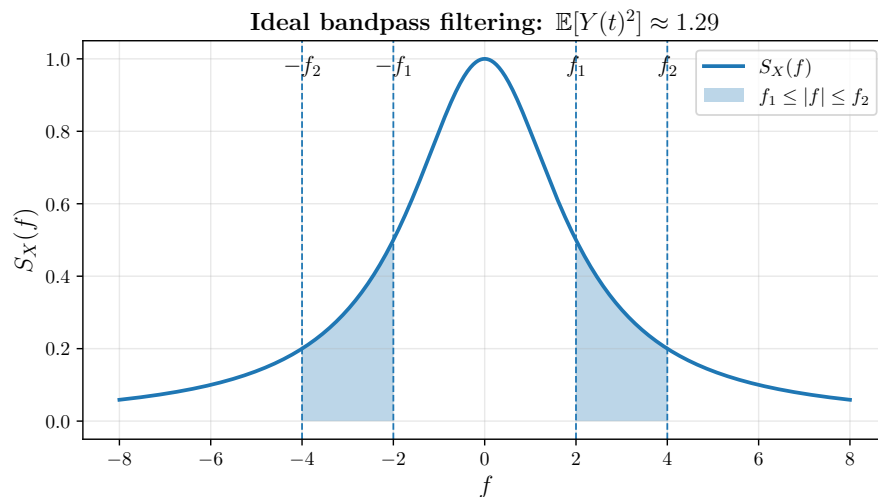


Figure 19: PSD interpretation of ideal bandpass filtering. The output power is the area of the input PSD over the passed frequency bands. For a real WSS process with an even PSD, the two symmetric bands contribute $2 \int_{f_1}^{f_2} S_X(f) df$.

Lecture 13: Linear Systems with Random Process Inputs

Today's lecture roughly follows parts of Lecture Notes 7 and 8 of Abbas El Gamal (AEG).

Recap: Mean, Autocorrelation, WSS, and PSD

For a continuous-time random process $X = \{X(t), -\infty < t < \infty\}$, the mean function and autocorrelation function are

$$\mu_X(t) = \mathbb{E}[X(t)], \quad R_X(t_1, t_2) = \mathbb{E}[X(t_1)X(t_2)].$$

For example, if $X(t) = \cos(t + \Theta)$ with $\Theta \sim \text{Unif}[0, 2\pi)$, then

$$\mu_X(t) = 0, \quad R_X(t_1, t_2) = \frac{1}{2} \cos(t_1 - t_2).$$

A process X is WSS if

$$\mu_X(t) = \mu \text{ for all } t, \quad R_X(t + \tau, t) \text{ is a function only of } \tau, \quad \mathbb{E}[X(t)^2] < \infty.$$

In this case, we write

$$R_X(\tau) = R_X(t + \tau, t).$$

For the random phase cosine example above,

$$R_X(\tau) = \frac{1}{2} \cos \tau.$$

For a WSS process, the autocorrelation function satisfies

$$R_X(\tau) = R_X(-\tau), \quad R_X(0) \geq |R_X(\tau)|, \quad R_X(\tau) \text{ is nonnegative definite.}$$

The power spectral density, or PSD, is defined as the Fourier transform of the autocorrelation function:

$$S_X(f) = \mathcal{F}\{R_X(\tau)\}.$$

The PSD satisfies

$$S_X(f) \text{ is real and even,} \quad \int_{-\infty}^{\infty} S_X(f) df = R_X(0), \quad S_X(f) \geq 0.$$

Which Functions Can Be Autocorrelation Functions?

Not every function can be the autocorrelation function of a WSS process. Recall that a valid autocorrelation function $R_X(\tau)$ must satisfy

$$R_X(\tau) = R_X(-\tau), \quad R_X(0) \geq |R_X(\tau)| \text{ for all } \tau, \quad S_X(f) = \mathcal{F}\{R_X(\tau)\} \geq 0.$$

These conditions let us quickly rule out many candidate functions.

A one-sided decaying function cannot be an autocorrelation function, since it is not even. For example,

$$R(\tau) = e^{-\tau} \mathbf{1}_{\{\tau \geq 0\}}$$

Which functions can be autocorrelation functions?

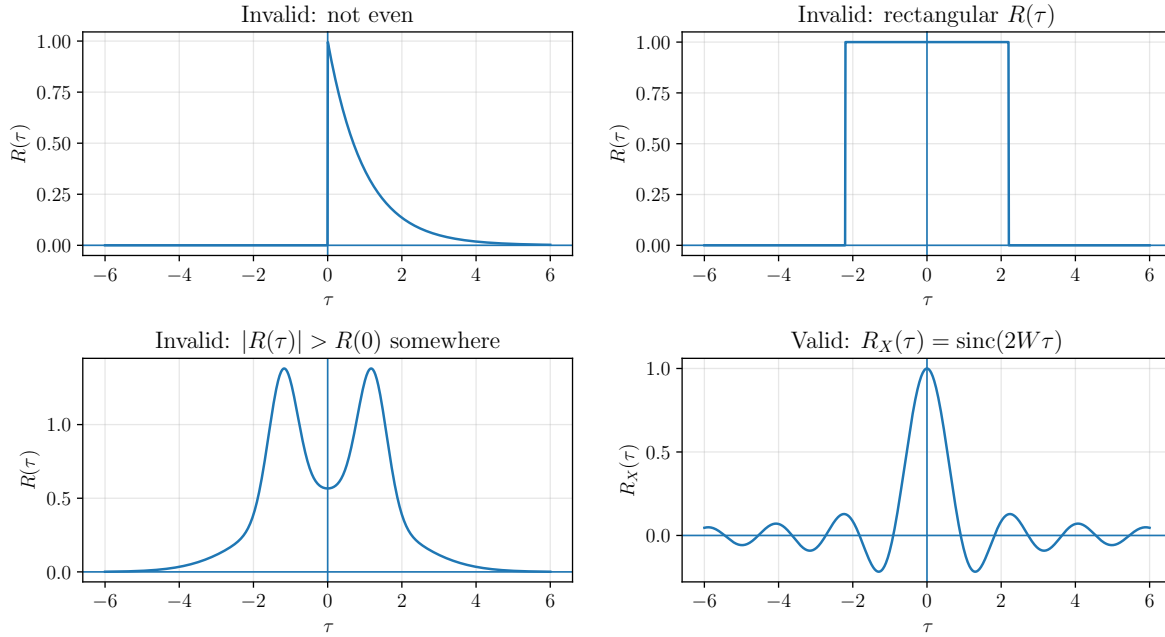


Figure 20: Examples of candidate autocorrelation functions. A valid WSS autocorrelation function must be even, must satisfy $R_X(0) \geq |R_X(\tau)|$ for all τ , and must have a nonnegative Fourier transform $S_X(f)$.

fails the condition $R(\tau) = R(-\tau)$.

A function can also fail because its magnitude away from the origin exceeds its value at the origin. If for some τ_0 ,

$$|R(\tau_0)| > R(0),$$

then R cannot be an autocorrelation function, since every WSS autocorrelation must satisfy

$$R_X(0) \geq |R_X(\tau)|.$$

Evenness and having a maximum at the origin are still not sufficient. Consider the rectangular function

$$R(\tau) = \mathbf{1}_{\{|\tau| \leq T\}}.$$

It is even and is maximized at $\tau = 0$, but its Fourier transform is proportional to

$$S(f) \propto 2T \operatorname{sinc}(2Tf),$$

which takes both positive and negative values. Since a PSD must satisfy $S_X(f) \geq 0$, this function cannot be a valid autocorrelation function.

On the other hand, a sinc function can be a valid autocorrelation function. Using the normalized convention

$$\operatorname{sinc}(x) = \frac{\sin(\pi x)}{\pi x},$$

let

$$R_X(\tau) = \operatorname{sinc}(2W\tau).$$

PSD-based justification of the last two examples

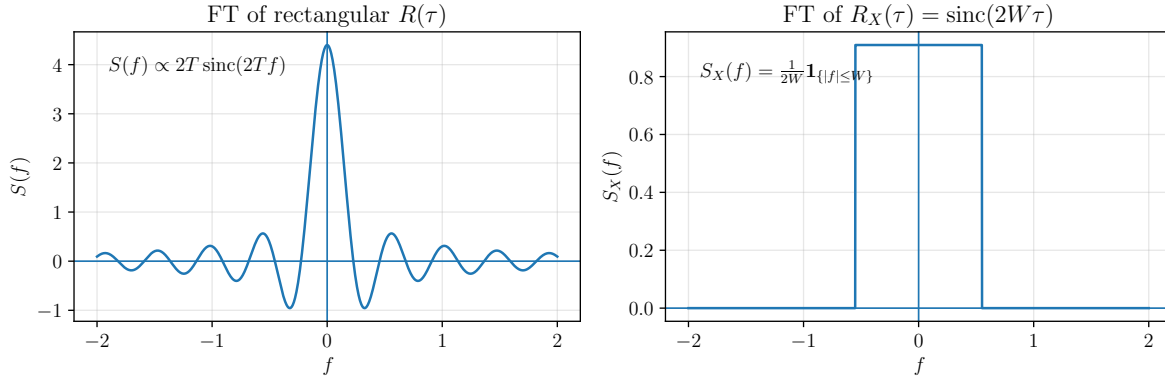


Figure 21: PSD-based justification for the last two examples. A rectangular function in τ has a sinc Fourier transform that takes negative values, so it cannot be a PSD. On the other hand, a sinc autocorrelation has a rectangular nonnegative PSD, so it is valid.

Its Fourier transform is

$$S_X(f) = \frac{1}{2W} \mathbf{1}_{\{|f| \leq W\}},$$

which is nonnegative for all f . Hence this is a valid autocorrelation function.

Switching Order of Expectation and Integration

Suppose V is a discrete random variable with possible values v_i and

$$\sum_i \Pr(V = v_i) = 1.$$

Let

$$X(t) = f(t, V).$$

Then

$$\begin{aligned} \mathbb{E} \left[\int_{-\infty}^{\infty} X(t) dt \right] &= \mathbb{E} \left[\int_{-\infty}^{\infty} f(t, V) dt \right] \\ &= \sum_i \Pr(V = v_i) \int_{-\infty}^{\infty} f(t, v_i) dt \\ &= \int_{-\infty}^{\infty} \left[\sum_i \Pr(V = v_i) f(t, v_i) \right] dt \\ &= \int_{-\infty}^{\infty} \mathbb{E}[X(t)] dt. \end{aligned}$$

Thus, under the benign regularity conditions we assume in this course,

$$\mathbb{E} \left[\int_{-\infty}^{\infty} X(t) dt \right] = \int_{-\infty}^{\infty} \mathbb{E}[X(t)] dt.$$



Figure 22: A general linear system and the LTI special case with random process input $X(t)$ and output $Y(t)$.

Linear Systems with Random Process Inputs

Now suppose a random process $X(t)$ is passed through a linear system to produce an output process $Y(t)$.

For a general linear time-varying system,

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

For an LTI system with impulse response $h(\cdot)$,

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

Given $\mu_X(t)$, $R_X(t_1, t_2)$, and the system kernel h , we want to find

$$\mu_Y(t), \quad R_Y(t_1, t_2), \quad R_{XY}(t_1, t_2).$$

The cross-correlation function between X and Y is

$$R_{XY}(t_1, t_2) = \mathbb{E}[X(t_1)Y(t_2)].$$

Also,

$$R_{XY}(t_1, t_2) = \mathbb{E}[X(t_1)Y(t_2)] = \mathbb{E}[Y(t_2)X(t_1)] = R_{YX}(t_2, t_1).$$

For the output mean,

$$\begin{aligned} \mu_Y(t) &= \mathbb{E}[Y(t)] \\ &= \mathbb{E} \left[\int_{-\infty}^{\infty} h(t, t - \tau) X(\tau) d\tau \right] \\ &= \int_{-\infty}^{\infty} h(t, t - \tau) \mathbb{E}[X(\tau)] d\tau \\ &= \int_{-\infty}^{\infty} h(t, t - \tau) \mu_X(\tau) d\tau. \end{aligned}$$

For the LTI special case,

$$\mu_Y(t) = \int_{-\infty}^{\infty} h(t - \tau) \mu_X(\tau) d\tau.$$

Cross-Correlation Between Input and Output

Recall that for a general linear system,

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

The cross-correlation from Y to X is

$$R_{YX}(t_1, t_2) = \mathbb{E}[Y(t_1)X(t_2)].$$

Using the system expression for $Y(t_1)$,

$$\begin{aligned} R_{YX}(t_1, t_2) &= \mathbb{E} \left[\left(\int_{-\infty}^{\infty} h(t_1, t_1 - \tau) X(\tau) d\tau \right) X(t_2) \right] \\ &= \int_{-\infty}^{\infty} h(t_1, t_1 - \tau) \mathbb{E}[X(\tau)X(t_2)] d\tau \\ &= \int_{-\infty}^{\infty} h(t_1, t_1 - \tau) R_X(\tau, t_2) d\tau. \end{aligned}$$

Similarly, the output autocorrelation is

$$R_Y(t_1, t_2) = \mathbb{E}[Y(t_1)Y(t_2)].$$

Substituting the linear-system expression for both $Y(t_1)$ and $Y(t_2)$ gives

$$\begin{aligned} R_Y(t_1, t_2) &= \mathbb{E} \left[\left(\int_{-\infty}^{\infty} h(t_1, t_1 - \tau_1) X(\tau_1) d\tau_1 \right) \left(\int_{-\infty}^{\infty} h(t_2, t_2 - \tau_2) X(\tau_2) d\tau_2 \right) \right] \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(t_1, t_1 - \tau_1) h(t_2, t_2 - \tau_2) R_X(\tau_1, \tau_2) d\tau_1 d\tau_2. \end{aligned}$$

Joint Wide-Sense Stationarity

Two random processes $X(t)$ and $Y(t)$ are *jointly WSS* if each process is WSS and the cross-correlation depends only on the time difference. That is,

$$R_{XY}(t + \tau, t)$$

is a function of τ only, not of t . In this case we write

$$R_{XY}(\tau) = R_{XY}(t + \tau, t).$$

The *cross power spectral density* of jointly WSS processes X and Y is

$$S_{XY}(f) = \mathcal{F}\{R_{XY}(\tau)\}.$$

LTI System with WSS Input

Now suppose $X(t)$ is WSS and is passed through an LTI system with impulse response $h(\cdot)$:

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

Since X is WSS, $R_X(\tau_1, \tau_2) = R_X(\tau_1 - \tau_2)$. Therefore,

$$R_{YX}(t_1, t_2) = \int_{-\infty}^{\infty} h(t_1 - \tau) R_X(\tau - t_2) d\tau.$$

Let $s = t_1 - \tau$. Then $\tau = t_1 - s$, so

$$\begin{aligned} R_{YX}(t_1, t_2) &= \int_{-\infty}^{\infty} h(s) R_X(t_1 - t_2 - s) ds \\ &= (h * R_X)(t_1 - t_2). \end{aligned}$$

Thus $R_{YX}(t_1, t_2)$ depends only on $t_1 - t_2$.

Next, for the output autocorrelation,

$$R_Y(t_1, t_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(t_1 - \tau_1) h(t_2 - \tau_2) R_X(\tau_1 - \tau_2) d\tau_1 d\tau_2.$$

For fixed τ_2 , the inner integral is

$$\int_{-\infty}^{\infty} h(t_1 - \tau_1) R_X(\tau_1 - \tau_2) d\tau_1 = (h * R_X)(t_1 - \tau_2).$$

Therefore,

$$R_Y(t_1, t_2) = \int_{-\infty}^{\infty} h(t_2 - \tau_2) (h * R_X)(t_1 - \tau_2) d\tau_2.$$

Let $s = t_2 - \tau_2$. Then

$$R_Y(t_1, t_2) = \int_{-\infty}^{\infty} h(s) (h * R_X)(t_1 - t_2 + s) ds.$$

Equivalently,

$$R_Y(t_1, t_2) = [h(-\tau) * h(\tau) * R_X(\tau)](t_1 - t_2).$$

Hence Y is WSS, with autocorrelation

$$R_Y(\tau) = h(-\tau) * h(\tau) * R_X(\tau).$$

Lecture 14: Cross-Correlation, Cross-Spectral Density, and White Noise

Recap: Random Processes, WSS, and Joint WSS

Let $X(t)$ be a random process. Its mean function and autocorrelation function are

$$\mu_X(t) = \mathbb{E}[X(t)], \quad R_X(t_1, t_2) = \mathbb{E}[X(t_1)X(t_2)].$$

The process X is WSS if

$$\mu_X(t) = \mu \text{ for all } t, \quad R_X(t + \tau, t) = R_X(\tau), \quad \mathbb{E}[X(t)^2] < \infty.$$

Now suppose $Y(t)$ is also a random process. The cross-correlation between X and Y is

$$R_{XY}(t_1, t_2) = \mathbb{E}[X(t_1)Y(t_2)].$$

The processes X and Y are *jointly WSS* if each is WSS and the cross-correlation depends only on the time difference:

$$R_{XY}(t + \tau, t) = R_{XY}(\tau).$$

Sum of jointly WSS processes

Suppose X and Y are jointly WSS, and define

$$Z(t) = X(t) + Y(t).$$

Then

$$\begin{aligned} R_Z(\tau) &= \mathbb{E}[Z(t + \tau)Z(t)] \\ &= \mathbb{E}[(X(t + \tau) + Y(t + \tau))(X(t) + Y(t))] \\ &= R_X(\tau) + R_{XY}(\tau) + R_{YX}(\tau) + R_Y(\tau). \end{aligned}$$

Using $R_{YX}(\tau) = R_{XY}(-\tau)$, this becomes

$$R_Z(\tau) = R_X(\tau) + R_Y(\tau) + R_{XY}(\tau) + R_{XY}(-\tau).$$

Uncorrelated zero-mean jointly WSS processes

Suppose X and Y are zero-mean jointly WSS processes. They are called *uncorrelated* if

$$R_{XY}(\tau) = 0 \text{ for all } \tau.$$

Therefore, if $Z(t) = X(t) + Y(t)$, then

$$R_Z(\tau) = R_X(\tau) + R_Y(\tau).$$

Power Spectral Density and Cross Power Spectral Density

For a WSS process X , the power spectral density is

$$S_X(f) = \mathcal{F}\{R_X(\tau)\}.$$

For jointly WSS processes X and Y , the *cross power spectral density* is

$$S_{XY}(f) = \mathcal{F}\{R_{XY}(\tau)\}.$$

Example: PSD of Discrete-Time White Noise

Let $W(n)$ be discrete-time white noise with

$$\mathbb{E}[W(n)] = 0, \quad \text{Var}(W(n)) = \sigma^2.$$

Then

$$R_W(k) = \mathbb{E}[W(n+k)W(n)] = \begin{cases} \sigma^2, & k = 0, \\ 0, & k \neq 0. \end{cases}$$

Equivalently,

$$R_W(k) = \sigma^2 \delta(k),$$

where $\delta(k)$ is the Kronecker delta function. Therefore,

$$S_W(f) = \mathcal{F}\{R_W(k)\} = \sigma^2 \mathcal{F}\{\delta(k)\} = \sigma^2 \quad \text{for all } f.$$

Thus white noise has a flat PSD.

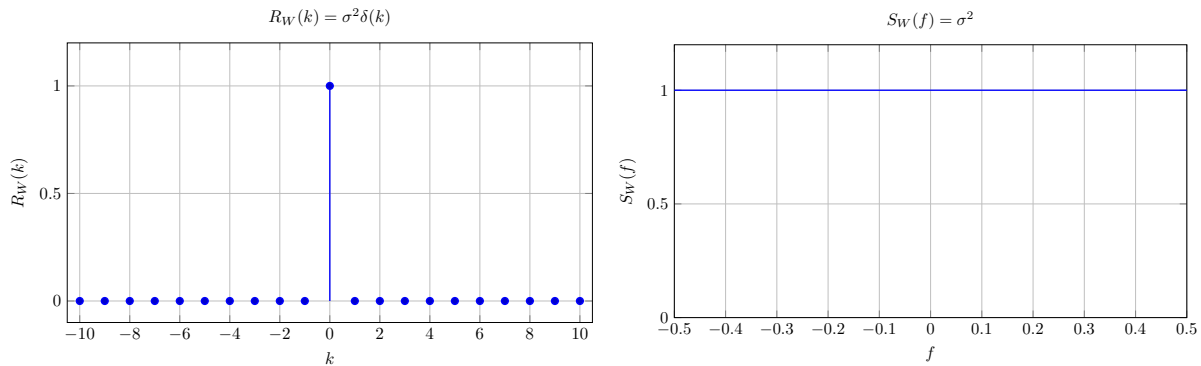


Figure 23: Autocorrelation and PSD of discrete-time white noise. Since $R_W(k) = \sigma^2 \delta(k)$, its PSD is flat: $S_W(f) = \sigma^2$ for all $-\frac{1}{2} \leq f < \frac{1}{2}$.

LTI system with WSS input

Suppose a WSS process $X(t)$ is passed through an LTI system with impulse response $h(t)$:

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

From the previous lecture, we saw that X and Y are jointly WSS with

$$R_{YX}(\tau) = h(\tau) * R_X(\tau),$$

and

$$R_Y(\tau) = h(-\tau) * h(\tau) * R_X(\tau).$$

Then, in the spectral domain,

$$S_{YX}(f) = H(f)S_X(f),$$

where

$$H(f) = \mathcal{F}\{h(t)\}.$$

Recall that

$$\mathcal{F}\{h(-t)\} = H^*(f).$$

From the previous result,

$$R_Y(\tau) = h(-\tau) * h(\tau) * R_X(\tau).$$

Taking Fourier transforms gives

$$S_Y(f) = H^*(f)H(f)S_X(f) = |H(f)|^2 S_X(f).$$

Proof that PSD is nonnegative

We now use the LTI filtering result to give intuition for why

$$S_X(f) \geq 0.$$

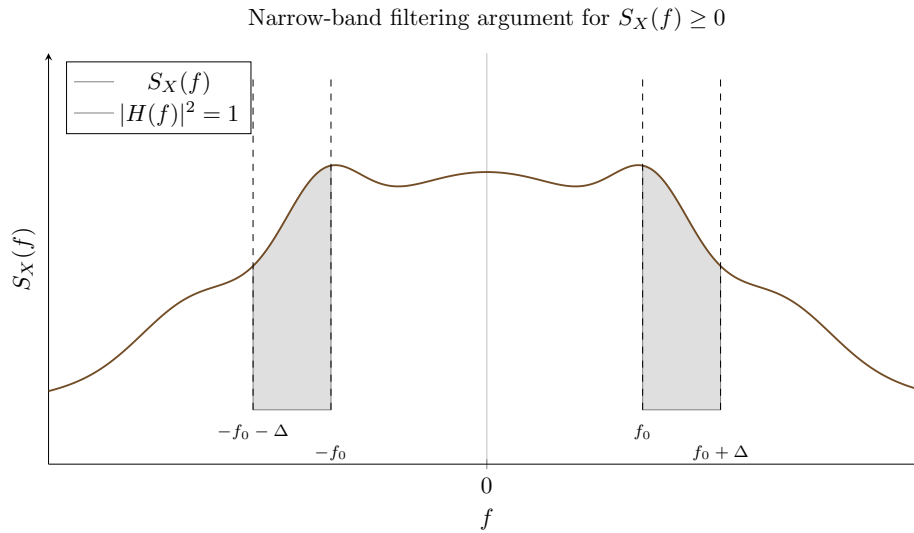


Figure 24: Narrow-band filtering argument for PSD nonnegativity. Passing a WSS process through an ideal LTI filter that keeps only two small bands around $\pm f_0$ gives an output whose average power is $2 \int_{f_0}^{f_0+\Delta} S_X(f) df \geq 0$. Since the band is arbitrary, the PSD must be nonnegative.

Consider a WSS input $X(t)$ passed through an ideal LTI filter whose frequency response has gain 1 on two narrow bands around $\pm f_0$:

$$f_0 \leq f \leq f_0 + \Delta, \quad -f_0 - \Delta \leq f \leq -f_0,$$

and gain 0 elsewhere. Let the output be $Y(t)$. Since $Y(t)^2 \geq 0$, we have

$$0 \leq \mathbb{E}[Y(t)^2] = R_Y(0).$$

Using the PSD-total-power identity and the LTI output PSD formula,

$$\begin{aligned} 0 \leq R_Y(0) &= \int_{-\infty}^{\infty} S_Y(f) df \\ &= \int_{-\infty}^{\infty} |H(f)|^2 S_X(f) df \\ &= 2 \int_{f_0}^{f_0+\Delta} S_X(f) df. \end{aligned}$$

Since f_0 and Δ are arbitrary, every small frequency band must have nonnegative total power. Hence the PSD must satisfy

$$S_X(f) \geq 0.$$

Hand-Wavy Interpretation of the PSD

A useful way to think about the PSD is to decompose it into narrow frequency bands. Suppose X is WSS with PSD $S_X(f)$. We can imagine slicing the PSD into many small pieces:

$$S_X(f) = \sum_i S_i^\Delta(f),$$

where $S_i^\Delta(f)$ is the contribution from a narrow band near f_i and, by symmetry, also near $-f_i$.

For very small Δ , each slice is approximately a pair of impulses:

$$S_i^\Delta(f) \approx \Delta S_X(f_i) \delta(f - f_i) + \Delta S_X(f_i) \delta(f + f_i).$$

This corresponds in the time domain to a random-phase sinusoidal component at frequency f_i . Thus, informally, a WSS process can be viewed as a superposition of uncorrelated random harmonics at different frequencies, where the power assigned to frequency f_i is proportional to $S_X(f_i)$.

This is the sense in which the PSD tells us how the total average power of the process is distributed across frequencies. Low-frequency components correspond to slowly varying structure or long time-scale memory, while high-frequency components correspond to rapidly varying structure or short time-scale memory.

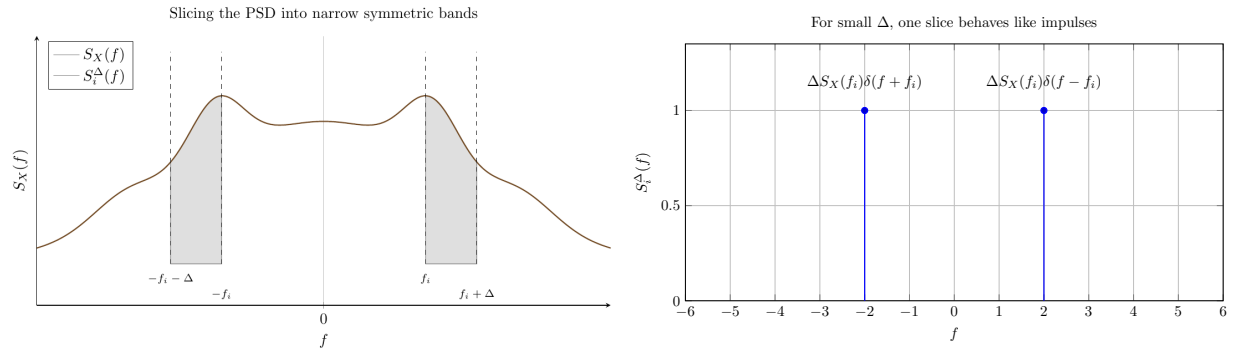


Figure 25: Frequency-domain intuition for PSD. The PSD can be decomposed into narrow symmetric slices, and for small Δ , each slice behaves approximately like impulses at $\pm f_i$. This motivates viewing a WSS process as a superposition of uncorrelated random harmonics at different frequencies.

Example: ARMA Processes

Let $W(n)$ be white noise with

$$\mathbb{E}[W(n)] = 0, \quad \text{Var}(W(n)) = \sigma^2.$$

An autoregressive moving-average process, abbreviated ARMA(p, q), is defined by

$$X(n) = - \sum_{k=1}^p \alpha_k X(n-k) + W(n) + \sum_{\ell=1}^q \beta_\ell W(n-\ell).$$

Equivalently,

$$X(n) + \sum_{k=1}^p \alpha_k X(n-k) = W(n) + \sum_{\ell=1}^q \beta_\ell W(n-\ell).$$

Define sequences $a(n)$ and $b(n)$ by

$$a(n) = \begin{cases} 1, & n = 0, \\ \alpha_n, & 1 \leq n \leq p, \\ 0, & \text{otherwise,} \end{cases} \quad b(n) = \begin{cases} 1, & n = 0, \\ \beta_n, & 1 \leq n \leq q, \\ 0, & \text{otherwise.} \end{cases}$$

Then the ARMA equation can be written as

$$X(n) * a(n) = W(n) * b(n).$$

Let

$$A(f) = \mathcal{F}\{a(n)\} = 1 + \sum_{k=1}^p \alpha_k e^{-j2\pi f k}, \quad B(f) = \mathcal{F}\{b(n)\} = 1 + \sum_{\ell=1}^q \beta_\ell e^{-j2\pi f \ell}.$$

Thus X is obtained by passing white noise W through the LTI filter with frequency response $\frac{B(f)}{A(f)}$. Since $S_W(f) = \sigma^2$, the PSD of X is

$$S_X(f) = S_W(f) \left| \frac{B(f)}{A(f)} \right|^2 = \sigma^2 \left| \frac{1 + \sum_{\ell=1}^q \beta_\ell e^{-j2\pi f \ell}}{1 + \sum_{k=1}^p \alpha_k e^{-j2\pi f k}} \right|^2.$$

Recall from EE102A that stability requires the roots of the AR polynomial

$$1 + \sum_{k=1}^p \alpha_k z^k \text{ to lie outside the unit circle.}$$

Example: AR(1)

(Check Figure 26 to see how PSD of the AR(1) process looks for different stable values of α .)

The AR(1), equivalently ARMA(1,0), process is

$$X(n) = -\alpha X(n-1) + W(n).$$

The AR polynomial is $1 + \alpha z$. Its root is

$$z = -\frac{1}{\alpha}.$$

Stability requires

$$\left| \frac{1}{\alpha} \right| > 1,$$

or equivalently

$$-1 < \alpha < 1.$$

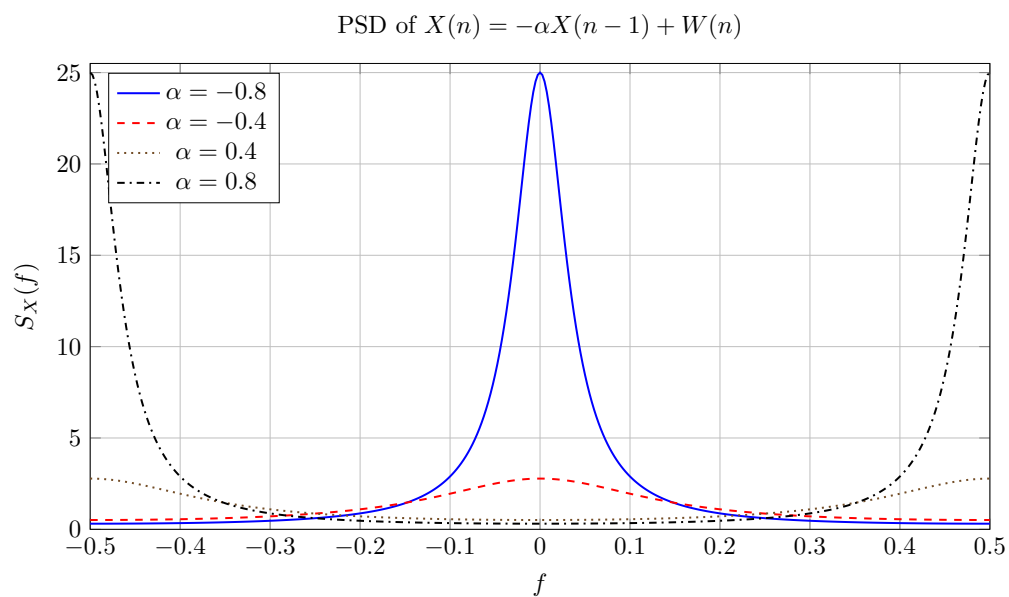


Figure 26: PSD of the AR(1) process $X(n) = -\alpha X(n-1) + W(n)$ for different stable values of α . Positive and negative α emphasize different parts of the discrete-time frequency range, while stability requires $|\alpha| < 1$.

Lecture 15: Process Linear Estimation, Wiener Filtering, and Spectral Factorization

Process Linear Estimation

Let $X(t)$ and $Y(t)$ be zero-mean jointly WSS processes. We seek the MMSE linear estimate of $X(t)$ based on observations

$$\{Y(s) : t - a \leq s \leq t + b\}.$$

Equivalently, we estimate $X(t)$ using a linear filter of the form

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

where h is chosen to minimize

$$\mathbb{E}[(X(t) - \hat{X}(t))^2].$$

By the orthogonality principle, the estimation error must be orthogonal to every observation used in the estimate:

$$\mathbb{E}[(X(t) - \hat{X}(t))Y(t - \tau)] = 0, \quad -b \leq \tau \leq a.$$

Thus we seek h such that

$$R_{XY}(\tau) = \mathbb{E}[\hat{X}(t)Y(t - \tau)].$$

Now

$$\begin{aligned} \mathbb{E}[\hat{X}(t)Y(t - \tau)] &= \mathbb{E}\left[\left(\int_{-b}^a h(\alpha)Y(t - \alpha) d\alpha\right)Y(t - \tau)\right] \\ &= \int_{-b}^a h(\alpha)\mathbb{E}[Y(t - \alpha)Y(t - \tau)] d\alpha \\ &= \int_{-b}^a h(\alpha)R_Y(\tau - \alpha) d\alpha. \end{aligned}$$

Therefore the optimal filter must satisfy

$$R_{XY}(\tau) = \int_{-b}^a h(\alpha)R_Y(\tau - \alpha) d\alpha, \quad -b \leq \tau \leq a.$$

We will consider two important cases:

infinite smoothing: $a = b = \infty$, filtering: $a = \infty$, $b = 0$.

Infinite Smoothing and the Non-Causal Wiener Filter

For infinite smoothing, $a = b = \infty$, so the estimate can use all past and future observations:

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

The normal equation becomes

$$R_{XY}(\tau) = \int_{-\infty}^{\infty} h(\alpha)R_Y(\tau - \alpha) d\alpha = (h * R_Y)(\tau).$$

Taking Fourier transforms gives

$$S_{XY}(f) = H(f)S_Y(f).$$

Hence the optimal non-causal Wiener filter is

$$H(f) = \frac{S_{XY}(f)}{S_Y(f)}.$$

Equivalently,

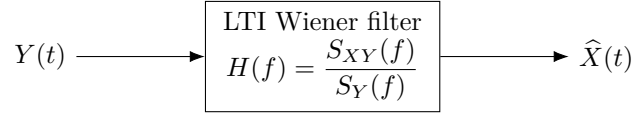


Figure 27: Non-causal Wiener filter for estimating $X(t)$ from the observation process $Y(t)$.

The resulting linear MMSE is

$$\begin{aligned} \mathbb{E}[(X(t) - \hat{X}(t))^2] &= \mathbb{E}[X(t)^2] - \mathbb{E}[\hat{X}(t)^2] \\ &= \int_{-\infty}^{\infty} S_X(f) df - \int_{-\infty}^{\infty} S_Y(f) \left| \frac{S_{XY}(f)}{S_Y(f)} \right|^2 df \\ &= \int_{-\infty}^{\infty} \left(S_X(f) - \frac{|S_{XY}(f)|^2}{S_Y(f)} \right) df. \end{aligned}$$

Example: Continuous-Time Additive White Noise Channel

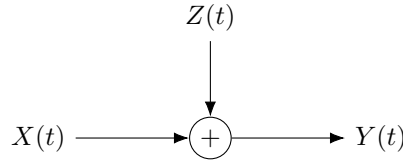


Figure 28: Continuous-time additive white noise channel, $Y(t) = X(t) + Z(t)$.

Consider

$$Y(t) = X(t) + Z(t),$$

where $X(t)$ and $Z(t)$ are zero-mean uncorrelated WSS processes, and $Z(t)$ is white noise with

$$S_Z(f) = \frac{N}{2}, \text{ for all } f.$$

Then

$$S_Y(f) = S_X(f) + S_Z(f) = S_X(f) + \frac{N}{2}.$$

Also,

$$\begin{aligned} R_{XY}(\tau) &= \mathbb{E}[X(t + \tau)Y(t)] \\ &= \mathbb{E}[X(t + \tau)(X(t) + Z(t))] \\ &= R_X(\tau), \end{aligned}$$

because X and Z are uncorrelated. Therefore,

$$S_{XY}(f) = S_X(f).$$

Thus the non-causal Wiener filter is

$$H(f) = \frac{S_X(f)}{S_X(f) + N/2}.$$

The corresponding MSE is

$$\begin{aligned} \text{MSE} &= \int_{-\infty}^{\infty} \left(S_X(f) - \frac{S_X(f)^2}{S_X(f) + N/2} \right) df \\ &= \frac{N}{2} \int_{-\infty}^{\infty} \frac{S_X(f)}{S_X(f) + N/2} df. \end{aligned}$$

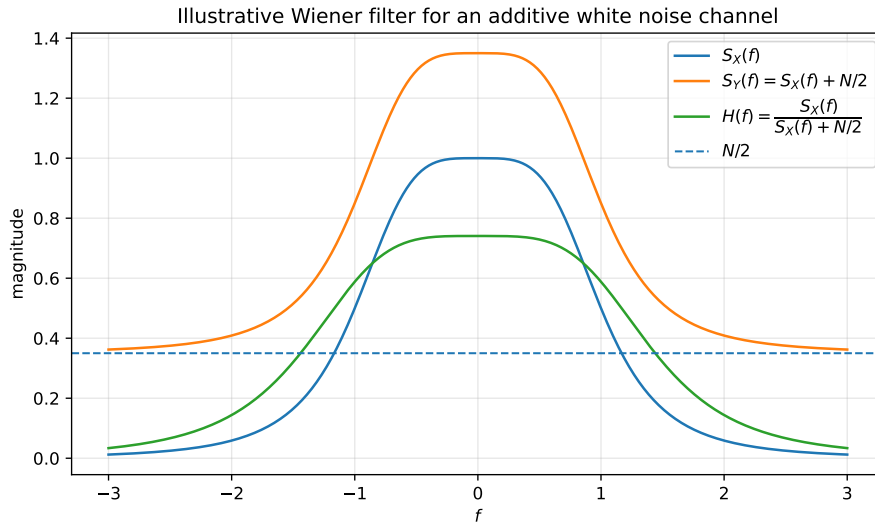


Figure 29: Illustrative Wiener filtering in an additive white noise channel. Frequencies where the signal PSD is large relative to the noise floor are passed with gain closer to 1, while frequencies with low signal-to-noise ratio are attenuated.

Spectral Factorization

We now assume discrete time. Given a WSS process $X(n)$ with PSD

$$S_X(f) = S(f),$$

we can find a frequency response $H(f)$ such that

$$S(f) = H(f)H^*(f) = |H(f)|^2.$$

For example, one possible choice is

$$H(f) = \sqrt{S(f)}.$$

Such an H is called a *square root* of S , or a *spectral factor*.

This gives a coloring interpretation. If $W(n)$ is white noise with

$$S_W(f) = 1,$$

then passing $W(n)$ through an LTI system with response $H(f)$ produces $Y(n)$ with

$$S_Y(f) = |H(f)|^2 S_W(f) = S(f).$$

So H colors white noise into a process with PSD S .

Conversely, suppose $Y(n)$ has PSD $S(f)$, that $H(f) \neq 0$ almost everywhere, and that $1/H(f)$ defines a well-posed stable inverse filter. Passing Y through this inverse filter gives

$$S_W(f) = \left| \frac{1}{H(f)} \right|^2 S_Y(f) = 1$$

almost everywhere. This is the whitening interpretation.

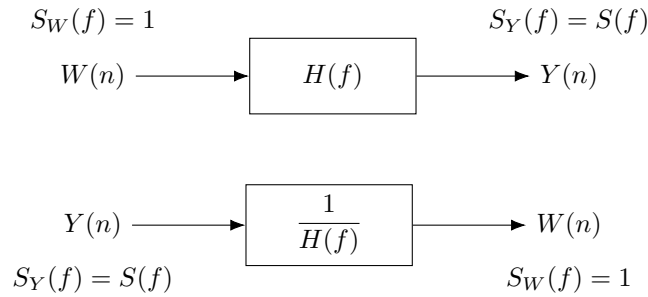


Figure 30: Coloring and whitening interpretations of spectral factorization.

Linear Prediction and the Innovations Process

Let $\hat{X}(n|n-1)$ denote the best MSE estimator of $X(n)$ of the form

$$\hat{X}(n|n-1) = \sum_{k=1}^{\infty} a(k)X(n-k).$$

Define the prediction error, or innovation,

$$U(n) = X(n) - \hat{X}(n|n-1).$$

By the orthogonality principle,

$$U(n) \perp X(n-k), \quad k \geq 1.$$

Since $U(n-k)$ is generated from the past samples $X(n-k), X(n-k-1), \dots$, this implies

$$U(n) \perp U(n-k), \quad k \geq 1.$$

Thus $\{U(n)\}$ is white noise. It is called the *innovation process*.

We can write

$$X(n) = \hat{X}(n|n-1) + U(n),$$

so

$$U(n) = X(n) - \sum_{k=1}^{\infty} a(k)X(n-k).$$

Equivalently,

$$U(n) = X(n) * g(n),$$

where

$$g(n) = \begin{cases} 1, & n = 0, \\ -a(n), & n \geq 1, \\ 0, & n < 0. \end{cases}$$

Thus g is causal. In the spectral domain, if $G(f) = \mathcal{F}\{g(n)\}$, then

$$S_U(f) = |G(f)|^2 S_X(f).$$

Since $U(n)$ is white noise,

$$S_U(f) = \sigma_U^2.$$

Therefore,

$$S_X(f) = \frac{\sigma_U^2}{|G(f)|^2}.$$

Assuming $1/G$ exists and is stable, we can also generate $X(n)$ from the innovation process:

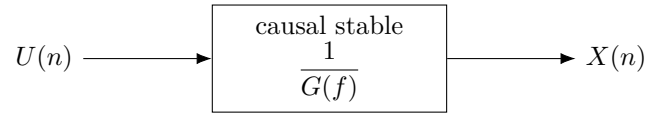


Figure 31: Reconstruction of $X(n)$ from the innovation process through a causal stable filter.

Causal Spectral Factorization

The previous discussion suggests a stronger version of spectral factorization. Under suitable regularity conditions, we can find a causal stable filter $S^+(f)$ such that, if $W(n)$ is white noise with

$$S_W(f) = 1,$$

then passing $W(n)$ through this filter produces a process $X(n)$ with

$$S_X(f) = S(f).$$

Equivalently,

$$S(f) = |S^+(f)|^2,$$

where $S^+(f)$ is chosen to correspond to a causal stable filter. This is the causal spectral factor.

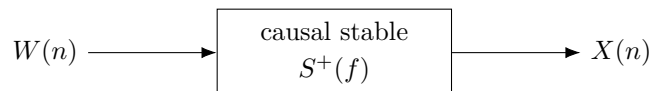


Figure 32: Causal spectral factorization: passing white noise through a causal stable spectral factor produces a process with the desired PSD.

Lecture 16: Spectral Factorization and Causal Wiener Filtering

Today's lecture very roughly follows Lecture Note 8 of Abbas El Gamal (AEG).

Roadmap and Recap

Today's lecture continues the frequency-domain viewpoint for WSS processes and LTI systems. The main topics covered are: i) a short extension of the jointly WSS/LTI calculations, ii) a recap of the non-causal Wiener filter, iii) spectral factorization and why causal spectral factors are useful, iv) the discrete-time spectral factorization theorem, v) the ARMA example, and vi) a proof sketch of the theorem.

The correspondence with AEG Lecture Note 8 is only rough: the goal here is to isolate the parts of the classical material that we will actually use.

Filtering One Process and Cross-Spectra

Suppose X and Y are jointly WSS, and let Z be obtained by passing Y through an LTI system with impulse response g :

$$Z(t) = \int_{-\infty}^{\infty} g(\alpha)Y(t - \alpha) d\alpha, \quad G(f) = \mathcal{F}\{g(t)\}.$$

Here lowercase g denotes the impulse response, while uppercase G denotes the frequency response. We want the cross-correlation and cross-spectrum between X and the filtered process Z .

By definition,

$$\begin{aligned} R_{XZ}(\tau) &= \mathbb{E}[X(t + \tau)Z(t)] \\ &= \mathbb{E}\left[X(t + \tau) \int_{-\infty}^{\infty} g(\alpha)Y(t - \alpha) d\alpha\right] \\ &= \int_{-\infty}^{\infty} g(\alpha) \mathbb{E}[X(t + \tau)Y(t - \alpha)] d\alpha \\ &= \int_{-\infty}^{\infty} g(\alpha) R_{XY}(\tau + \alpha) d\alpha. \end{aligned}$$

Let $\beta = -\alpha$. Then

$$R_{XZ}(\tau) = \int_{-\infty}^{\infty} g(-\beta) R_{XY}(\tau - \beta) d\beta.$$

Equivalently, if $\tilde{g}(\tau) = g(-\tau)$, then

$$R_{XZ}(\tau) = (\tilde{g} * R_{XY})(\tau).$$

Taking Fourier transforms gives

$$S_{XZ}(f) = G^*(f)S_{XY}(f),$$

where for real-valued impulse responses $\mathcal{F}\{g(-t)\} = G^*(f)$.

Recall also that if a WSS process X is passed through an LTI system H , producing Y , then X and Y are jointly WSS with

$$S_{XY}(f) = H^*(f)S_X(f), \quad S_{YX}(f) = H(f)S_X(f),$$

and

$$S_Y(f) = |H(f)|^2 S_X(f).$$

Recap: Non-Causal Wiener Filter

For zero-mean jointly WSS processes X and Y , the infinite-smoothing, non-causal Wiener filter estimates $X(t)$ from all of Y using

$$Y(t) \longrightarrow \boxed{\text{LTI } H(f) = \frac{S_{XY}(f)}{S_Y(f)}} \longrightarrow \hat{X}(t).$$

The resulting MSE is

$$\begin{aligned} \mathbb{E}[(X(t) - \hat{X}(t))^2] &= \mathbb{E}[X(t)^2] - \mathbb{E}[\hat{X}(t)^2] \\ &= \int_{-\infty}^{\infty} S_X(f) df - \int_{-\infty}^{\infty} S_Y(f) \left| \frac{S_{XY}(f)}{S_Y(f)} \right|^2 df \\ &= \int_{-\infty}^{\infty} \left(S_X(f) - \frac{|S_{XY}(f)|^2}{S_Y(f)} \right) df. \end{aligned}$$

This expression is just the frequency-domain version of the projection picture behind linear MMSE estimation. The estimate $\hat{X}(t)$ is the orthogonal projection of $X(t)$ onto the linear space generated by the observations of Y . Therefore $X(t) - \hat{X}(t)$ is orthogonal to $\hat{X}(t)$, and the Pythagorean identity gives

$$\mathbb{E}[(X(t) - \hat{X}(t))^2] = \mathbb{E}[X(t)^2] - \mathbb{E}[\hat{X}(t)^2].$$

The first term is the total power of X , and the second term is the power captured by the optimal linear estimate.

Spectral Factorization: Recap

Assume discrete time. Given a WSS process $X(n)$ with PSD

$$S_X(f) = S(f),$$

a function H is called a *spectral factor* of S if

$$S(f) = H(f)H^*(f) = |H(f)|^2.$$

A spectral factor always exists in a very basic sense; for example, one may take $H(f) = \sqrt{S(f)}$. But this factor need not correspond to a causal, stable, invertible filter, which is the more important question for prediction and causal filtering.

The reason spectral factors are useful is the coloring/whitening interpretation. If $W(n)$ is white noise with $S_W(f) = 1$, then passing W through an LTI system with response H produces a process with PSD

$$S_X(f) = |H(f)|^2 S_W(f) = S(f).$$

Thus H colors white noise into a process with spectrum S . Conversely, passing a process with PSD S through $1/H$ whitens it, provided $1/H$ is stable.

Linear Prediction and Innovations

One way to see why a causal spectral factor should exist is through linear prediction. Let $\hat{X}(n|n-1)$ denote the optimal linear estimate of $X(n)$ based on the past samples

$$X(n-1), X(n-2), \dots$$

Thus

$$\hat{X}(n|n-1) = \sum_{k=1}^{\infty} a(k)X(n-k).$$

Define the prediction error, or innovation,

$$U(n) = X(n) - \hat{X}(n|n-1).$$

By the orthogonality principle, $U(n)$ is orthogonal to every past sample $X(n-k)$, $k \geq 1$. Since $U(n-k)$ is generated from the samples $X(n-k), X(n-k-1), \dots$, this implies

$$U(n) \perp U(n-k), \quad k \geq 1.$$

Thus the innovation process is white:

$$S_U(f) = \sigma_U^2,$$

where σ_U^2 is the innovation variance. Equivalently, σ_U^2 is the MSE of the optimal linear one-step predictor.

We can write the innovation as the output of a causal filter:

$$U(n) = X(n) - \sum_{k=1}^{\infty} a(k)X(n-k) = X(n) * g(n),$$

where

$$g(n) = \begin{cases} 1, & n = 0, \\ -a(n), & n \geq 1, \\ 0, & n < 0. \end{cases}$$

Thus

$$X(n) \longrightarrow \boxed{\text{causal } G(f)} \longrightarrow U(n),$$

where $G(f) = \mathcal{F}\{g(n)\}$. Assuming $1/G$ exists and is stable, we can recover X from the innovations:

$$U(n) \longrightarrow \boxed{\text{causal stable } \frac{1}{G(f)}} \longrightarrow X(n).$$

Therefore, under appropriate regularity conditions, we can find a causal stable filter $H = 1/G$ such that

$$W(n) \longrightarrow \boxed{H(f)} \longrightarrow X(n), \quad S_W(f) = 1, \quad S_X(f) = S(f).$$

The informal picture is therefore: from X to its innovations U , there is a causal whitening filter G ; if G is invertible with stable inverse, then $1/G$ is a causal coloring filter from white noise back to X . The spectral factorization theorem guarantees a causal spectral factor under the Paley-Wiener condition. BIBO stability and stable invertibility require additional assumptions.

Spectral Factorization Theorem

We now state the discrete-time spectral factorization theorem. Suppose $S(f)$ is a nonnegative, integrable spectral density satisfying the Paley-Wiener condition

$$\int_{-1/2}^{1/2} \log S(f) df > -\infty.$$

Under this condition, there exists a causal spectral factor $S^+(f)$ such that

$$S(f) = S^+(f)S^-(f),$$

where

$$S^-(f) = \left(S^+(f)\right)^*.$$

Equivalently,

$$S(f) = \left|S^+(f)\right|^2.$$

Here, causal means that the inverse Fourier series of $S^+(f)$ has no negative-time coefficients. Since S is integrable, the impulse response corresponding to S^+ is square summable. The Paley-Wiener condition alone does not guarantee BIBO stability of this filter or of its inverse. BIBO stability means that the corresponding impulse response is absolutely summable. Whenever a causal BIBO-stable factor with a causal BIBO-stable inverse is required, those properties must be assumed separately or established using stronger regularity conditions.

Example: ARMA Processes and Spectral Factorization

Consider the ARMA(p, q) process

$$X(n) = -\sum_{k=1}^p \alpha_k X(n-k) + W(n) + \sum_{\ell=1}^q \beta_\ell W(n-\ell),$$

where $W(n)$ is zero-mean white noise with variance σ^2 . In the frequency domain, X is obtained by passing W through the LTI system

$$H(f) = \frac{1 + \sum_{\ell=1}^q \beta_\ell e^{-j2\pi f\ell}}{1 + \sum_{k=1}^p \alpha_k e^{-j2\pi fk}}.$$

Thus

$$W(n) \longrightarrow \boxed{\text{LTI } H(f)} \longrightarrow X(n),$$

and

$$S_X(f) = \sigma^2 H(f) H^*(f).$$

Assume that the numerator and denominator of H have no common factors. Define

$$\begin{aligned} A(z) &= 1 + \sum_{k=1}^p \alpha_k z^{-k}, \\ B(z) &= 1 + \sum_{\ell=1}^q \beta_\ell z^{-\ell}, \\ H(z) &= \frac{B(z)}{A(z)}. \end{aligned}$$

For a causal rational system, BIBO stability requires all poles of H , equivalently all zeros of $A(z)$, to lie strictly inside the unit circle in the z -plane.

If $w = z^{-1}$ and

$$\tilde{A}(w) = 1 + \sum_{k=1}^p \alpha_k w^k,$$

then all zeros of $\tilde{A}(w)$ lie strictly outside the unit circle in the w -plane.

If, in addition, all zeros of H , equivalently all zeros of $B(z)$, lie strictly inside the unit circle in the z -plane, then H is minimum phase and both H and $1/H$ are causal and BIBO stable. In terms of the delay variable $w = z^{-1}$, this means that all zeros of

$$\tilde{B}(w) = 1 + \sum_{\ell=1}^q \beta_{\ell} w^{\ell}$$

lie strictly outside the unit circle in the w -plane.

Under these assumptions, a causal BIBO-stable spectral factor with a causal BIBO-stable inverse is

$$S_X^+(f) = \sigma H(f),$$

because

$$S_X(f) = \sigma^2 H(f) H^*(f) = (\sigma H(f)) (\sigma H(f))^*.$$

This rational case is especially useful because the causal spectral factor can often be read off directly from the transfer function.

Proof Sketch of Spectral Factorization

Define

$$L(f) = \log S(f).$$

Since $S(f)$ is real, nonnegative, and symmetric, $L(f)$ is real and symmetric. The Paley-Wiener condition ensures that the inverse Fourier coefficients exist:

$$\ell(n) = \mathcal{F}^{-1}\{L(f)\}.$$

Because $L(f)$ is real and symmetric, $\ell(n)$ is real and symmetric:

$$\ell(n) = \ell(-n).$$

Now split $\ell(n)$ into causal and anti-causal parts. Define

$$\ell_c(n) = \begin{cases} \ell(0)/2, & n = 0, \\ \ell(n), & n > 0, \\ 0, & n < 0, \end{cases} \quad \ell_{ac}(n) = \begin{cases} \ell(0)/2, & n = 0, \\ \ell(n), & n < 0, \\ 0, & n > 0. \end{cases}$$

Since $\ell(n)$ is symmetric,

$$\ell_{ac}(n) = \ell_c(-n), \quad \ell(n) = \ell_c(n) + \ell_c(-n).$$

Taking Fourier transforms,

$$L(f) = L_c(f) + L_c^*(f).$$

Therefore,

$$\begin{aligned} S(f) &= \exp\{L(f)\} \\ &= \exp\{L_c(f)\} \exp\{L_c^*(f)\}. \end{aligned}$$

Define

$$S^+(f) = \exp\{L_c(f)\}, \quad S^-(f) = \exp\{L_c^*(f)\}.$$

Then

$$S(f) = S^+(f)S^-(f),$$

where S^+ is causal and S^- is anti-causal. More explicitly,

$$S^+(f) = \exp\left\{\frac{\ell(0)}{2} + \sum_{n=1}^{\infty} \ell(n)e^{-j2\pi fn}\right\}.$$

Using the exponential power series,

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!},$$

we see that $S^+(f)$ has only nonnegative-time Fourier coefficients, so it corresponds to a causal filter. Similarly, $S^-(f) = (S^+(f))^*$ is anti-causal. This gives the desired causal spectral factorization.

This proof is constructive: it tells us how to obtain the causal factor from the spectrum by taking $L(f) = \log S(f)$, keeping the causal half of its inverse Fourier coefficients, and exponentiating. In general the resulting expression is not particularly pleasant to work with, but in structured cases such as ARMA processes, the causal factor can often be identified directly.

Looking Ahead: Causal Wiener Filtering

The point of causal spectral factorization is that it lets us reduce causal estimation from a correlated observation process to estimation from innovations. When the chosen spectral factor and its inverse are both causal and BIBO stable, the whitening transformation is causal and invertible, so the past of the innovations contains the same linear information as the past of Y . Since the innovations are white, estimation from them becomes much simpler. This is the key idea behind the causal Wiener filter.

Lecture 17: Causal Wiener Filtering, Lookahead, and Linear Prediction

Roadmap

Today's lecture uses the tools from the last few lectures to derive the causal Wiener filter. We then extend the same idea to filtering with lookahead, and finally specialize it to optimal one-step linear prediction.

Causal Wiener Filter

Let $X(n)$ and $Y(n)$ be zero-mean jointly WSS discrete-time processes, with autocorrelations and cross-correlation $R_X(n)$, $R_{XY}(n)$, and $R_Y(n)$. We seek the optimal, in MSE sense, linear estimator $\hat{X}(n)$ of $X(n)$ based only on the present and past observations

$$\{Y(n-k) : k \geq 0\}.$$

Equivalently, we seek an estimator of the form

$$\hat{X}(n) = Y(n) * h(n),$$

where h is causal.

Assume that $S_Y(f)$ admits a spectral factorization

$$S_Y(f) = S_Y^+(f)S_Y^-(f), \quad S_Y^-(f) = \left(S_Y^+(f)\right)^*,$$

such that $S_Y^+(f)$ and $1/S_Y^+(f)$ are both causal and BIBO stable. Pass Y through the causal whitening filter $1/S_Y^+(f)$:

$$Y(n) \longrightarrow \boxed{\frac{1}{S_Y^+(f)}} \longrightarrow \tilde{Y}(n).$$

Then

$$S_{\tilde{Y}}(f) = \left| \frac{1}{S_Y^+(f)} \right|^2 S_Y(f) = 1.$$

Thus $\tilde{Y}(n)$ is white with unit variance. The key point is that this whitening step does not lose causal linear information. Since $1/S_Y^+$ is causal and stable, and its inverse S_Y^+ is also causal and stable, the samples $\{Y(k) : k \leq n\}$ and $\{\tilde{Y}(k) : k \leq n\}$ contain the same linearly recoverable information. Therefore, instead of estimating from the correlated observations Y , we may first causally whiten them and estimate from the white innovations $\tilde{Y}$.

Since the components of $\tilde{Y}$ are uncorrelated and have unit variance, the optimal causal linear estimate of $X(n)$ from $\{\tilde{Y}(n-k) : k \geq 0\}$ is obtained by projecting onto each component separately:

$$\begin{aligned} \hat{X}(n) &= \sum_{k=0}^{\infty} \tilde{Y}(n-k) \frac{\mathbb{E}[X(n)\tilde{Y}(n-k)]}{\text{Var}(\tilde{Y}(n-k))} \\ &= \sum_{k=0}^{\infty} \tilde{Y}(n-k) R_{X\tilde{Y}}(k). \end{aligned}$$

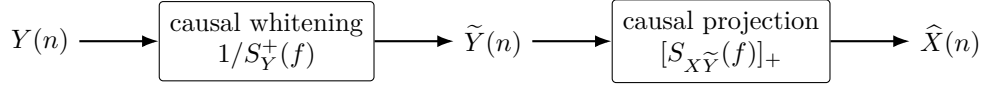


Figure 33: Causal Wiener filtering: causally whiten the observation process and then keep only the causal part of the optimal estimator in the whitened domain.

Define the causal truncation operation

$$[h(n)]_+ = \begin{cases} h(n), & n \geq 0, \\ 0, & n < 0, \end{cases} \quad [H(f)]_+ = \mathcal{F}\{[h(n)]_+\}.$$

The operation $[\cdot]_+$ keeps the causal part of a filter and kills the anti-causal part. In the whitened domain, the optimal causal estimator therefore keeps the causal part of the cross-spectrum:

$$\tilde{Y}(n) \longrightarrow [S_{X\tilde{Y}}(f)]_+ \longrightarrow \hat{X}(n).$$

Because $\tilde{Y}$ is obtained from Y by passing it through $1/S_Y^+$, we have

$$S_{X\tilde{Y}}(f) = \left(\frac{1}{S_Y^+(f)} \right)^* S_{XY}(f) = \frac{S_{XY}(f)}{S_Y^-(f)}.$$

Therefore the causal Wiener filter from Y to $\hat{X}$ is

$$H_{\text{CWF}}(f) = \frac{1}{S_Y^+(f)} \left[\frac{S_{XY}(f)}{S_Y^-(f)} \right]_+.$$

Compare this with the non-causal Wiener filter

$$H_{\text{WF}}(f) = \frac{S_{XY}(f)}{S_Y(f)} = \frac{1}{S_Y^+(f)} \frac{S_{XY}(f)}{S_Y^-(f)}.$$

Both filters first whiten the observations. The non-causal Wiener filter then uses all components of the whitened process, past and future, while the causal Wiener filter uses only the components available up to the present. This is exactly why the causal filter has the extra projection $[\cdot]_+$.

Estimation with Lookahead

Now suppose we allow d steps of lookahead. We seek an estimator of the form

$$\hat{X}_d(n) = \sum_{k=-d}^{\infty} h_d(k) Y(n-k),$$

so the estimator may use $Y(n+d), Y(n+d-1), \dots, Y(n)$ and all past observations. Positive d means lookahead, while negative d can be interpreted as lookback.

Define

$$Z(n) = Y(n+d).$$

Then estimating $X(n)$ from $\{Y(n-k) : k \geq -d\}$ is equivalent to estimating $X(n)$ causally from $\{Z(n-k) : k \geq 0\}$. Since Z is just a time shift of Y ,

$$S_Z(f) = S_Y(f).$$

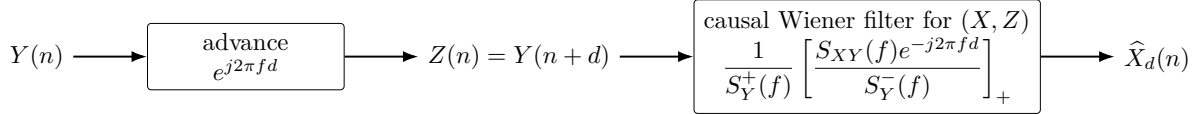


Figure 34: Filtering with d steps of lookahead. Shifting Y to form $Z(n) = Y(n+d)$ turns the lookahead problem into an ordinary causal Wiener filtering problem.

Also, because $Z(n) = Y(n+d)$ corresponds to the frequency response $e^{j2\pi fd}$ from Y to Z ,

$$S_{XZ}(f) = e^{-j2\pi fd} S_{XY}(f).$$

Applying the causal Wiener filter formula to the pair (X, Z) gives the optimal causal filter from Z to $\hat{X}_d$:

$$\frac{1}{S_Y^+(f)} \left[\frac{S_{XY}(f)e^{-j2\pi fd}}{S_Y^-(f)} \right]_+.$$

As a filter from the original observation process Y to $\hat{X}_d$, we must also include the advance from Y to Z , so

$$H_d(f) = e^{j2\pi fd} \frac{1}{S_Y^+(f)} \left[\frac{S_{XY}(f)e^{-j2\pi fd}}{S_Y^-(f)} \right]_+.$$

The middle operation has a simple time-domain meaning. If $G(f)$ has inverse Fourier transform $g(n)$, then

$$\mathcal{F}^{-1} \left\{ e^{j2\pi fd} [G(f)e^{-j2\pi fd}]_+ \right\} = \begin{cases} g(n), & n \geq -d, \\ 0, & n < -d. \end{cases}$$

Indeed, multiplying by $e^{-j2\pi fd}$ shifts the filter to the right by d , applying $[\cdot]_+$ keeps the causal part, and multiplying by $e^{j2\pi fd}$ shifts back. Thus the lookahead estimator keeps exactly the part of the filter that is allowed to use up to d future samples.

The MSE of the lookahead estimator is

$$\text{MSE}_d = \int_{-1/2}^{1/2} \left[S_X(f) - \left| \left[\frac{S_{XY}(f)e^{-j2\pi fd}}{S_Y^-(f)} \right]_+ \right|^2 \right] df.$$

For comparison, the non-causal Wiener filter MSE is

$$\text{MSE}_{\text{NC}} = \int_{-1/2}^{1/2} \left[S_X(f) - \frac{|S_{XY}(f)|^2}{S_Y(f)} \right] df = \int_{-1/2}^{1/2} \left[S_X(f) - \left| \frac{S_{XY}(f)}{S_Y^-(f)} \right|^2 \right] df.$$

As d increases, the estimator is allowed to keep more coefficients, so the MSE decreases toward the non-causal Wiener filter MSE. The causal case is $d = 0$.

Optimal Linear Prediction

We now specialize the causal Wiener filter to the problem of one-step linear prediction. The optimal linear predictor has the form

$$\hat{X}(n|n-1) = \sum_{k=1}^{\infty} a(k)X(n-k).$$

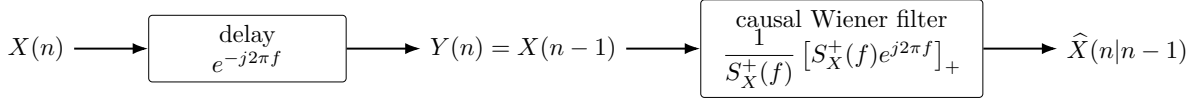


Figure 35: One-step linear prediction as a special case of causal Wiener filtering, with $Y(n) = X(n-1)$.

Equivalently, it is a strictly causal filter from X to $\hat{X}$:

$$X(n) \longrightarrow A(f) \longrightarrow \hat{X}(n|n-1).$$

To obtain it using the causal Wiener filter formula, set

$$Y(n) = X(n-1).$$

Then estimating $X(n)$ causally from $Y(n)$ is the same as estimating $X(n)$ from the past samples $X(n-1), X(n-2), \dots$. Since Y is a delayed version of X ,

$$S_Y(f) = S_X(f).$$

Also, using the convention $R_{XY}(k) = \mathbb{E}[X(n+k)Y(n)]$,

$$S_{XY}(f) = e^{j2\pi f} S_X(f).$$

Therefore,

$$\frac{S_{XY}(f)}{S_Y^-(f)} = \frac{e^{j2\pi f} S_X(f)}{S_X^-(f)} = e^{j2\pi f} S_X^+(f).$$

The causal Wiener filter from Y to $\hat{X}$ is

$$\frac{1}{S_X^+(f)} \left[e^{j2\pi f} S_X^+(f) \right]_+.$$

Since $Y(n) = X(n-1)$, the corresponding strictly causal predictor filter from X to $\hat{X}(n|n-1)$ is

$$A(f) = e^{-j2\pi f} \frac{1}{S_X^+(f)} \left[S_X^+(f) e^{j2\pi f} \right]_+.$$

This is strictly causal: the bracketed term is causal, $1/S_X^+$ is causal, and the outer factor $e^{-j2\pi f}$ adds one delay.

Prediction Error and the Kolmogorov-Szego Formula

The prediction error, or innovation, is

$$U(n) = X(n) - \hat{X}(n|n-1).$$

Recall from the previous lecture that the innovation process is white, and its variance σ_U^2 is exactly the MSE of the optimal linear predictor.

From the MSE formula for the causal Wiener filter in this special case,

$$\sigma_U^2 = \int_{-1/2}^{1/2} \left(|S_X^+(f)|^2 - \left| \left[S_X^+(f) e^{j2\pi f} \right]_+ \right|^2 \right) df.$$

Let

$$g(n) = \mathcal{F}^{-1}\{S_X^+(f)\}.$$

Since S_X^+ is causal, we can write

$$S_X^+(f) = \sum_{n=0}^{\infty} g(n)e^{-j2\pi fn}.$$

Then

$$S_X^+(f)e^{j2\pi f} = g(0)e^{j2\pi f} + g(1) + g(2)e^{-j2\pi f} + \dots.$$

Applying $[\cdot]_+$ removes the anti-causal term $g(0)e^{j2\pi f}$ and keeps the rest. Therefore, by Parseval,

$$\sigma_U^2 = |g(0)|^2.$$

Now recall the spectral-factor construction from the previous lecture. Let

$$\ell(n) = \mathcal{F}^{-1}\{\log S_X(f)\}.$$

Then

$$S_X^+(f) = \exp \left\{ \frac{\ell(0)}{2} + \sum_{n=1}^{\infty} \ell(n)e^{-j2\pi fn} \right\}.$$

The coefficient $g(0)$ is the constant Fourier coefficient of S_X^+ . Since all terms in $\sum_{n=1}^{\infty} \ell(n)e^{-j2\pi fn}$ have strictly positive delay, the constant coefficient comes only from $\exp\{\ell(0)/2\}$. Thus

$$g(0) = \exp \left\{ \frac{\ell(0)}{2} \right\}.$$

But

$$\ell(0) = \int_{-1/2}^{1/2} \log S_X(f) df.$$

Therefore,

$$\sigma_U^2 = \exp \left\{ \int_{-1/2}^{1/2} \log S_X(f) df \right\}.$$

This is the *Kolmogorov-Szego formula*. It says that the MSE of the optimal one-step linear predictor is the geometric mean of the PSD. Equivalently, it gives the innovation variance directly from the spectrum of the process.

Lecture 18: Course Wrap-Up and Prediction in Practice

Today's lecture roughly follows Lecture Notes 9 of Abbas El Gamal (AEG).

This final lecture briefly recaps the delayed causal Wiener filter, then steps back to discuss time-series prediction in practice, why linear estimation remains useful, the central themes of the course, and related follow-up directions.

Quick recap: delayed causal Wiener filtering

Let $X(n)$ and $Y(n)$ be zero-mean jointly WSS discrete-time random processes. We seek the optimal linear estimator of $X(n)$ based on the observations $\{Y(n-k) : k \geq -d\}$. Equivalently,

$$\hat{X}_d(n) = \sum_{k=-d}^{\infty} h_d(k)Y(n-k) = (h_d * Y)(n), \quad h_d(k) = 0 \text{ for all } k < -d.$$

Thus the estimator is allowed a lookahead of d samples. When $d = 0$, this reduces to the ordinary causal Wiener filter. When $d \rightarrow \infty$, this approaches the non-causal Wiener filter.

Assume $S_Y(f)$ satisfies the Paley–Wiener condition and admits the spectral factorization

$$S_Y(f) = S_Y^+(f)S_Y^-(f), \quad S_Y^-(f) = (S_Y^+(f))^*.$$

Then the frequency response of the optimal delayed causal Wiener filter is

$$H_d(f) = \frac{1}{S_Y^+(f)} e^{j2\pi fd} \left[\frac{S_{XY}(f)}{S_Y^-(f)} e^{-j2\pi fd} \right]_+.$$

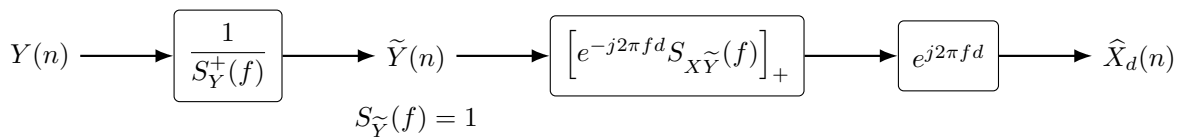
Here, for $G(f) = \sum_{n=-\infty}^{\infty} g(n)e^{-j2\pi fn}$, $[G(f)]_+$ denotes the Fourier transform of the causal part $g_+(n)$, where

$$g_+(n) = \begin{cases} g(n), & n \geq 0, \\ 0, & n < 0. \end{cases}$$

A useful interpretation of the formula is: first causally whiten Y by passing it through $1/S_Y^+(f)$, obtaining $\tilde{Y}$ with $S_{\tilde{Y}}(f) = 1$; next use the cross-spectrum

$$S_{X\tilde{Y}}(f) = \left(\frac{1}{S_Y^+(f)} \right)^* S_{XY}(f) = \frac{S_{XY}(f)}{S_Y^-(f)};$$

finally, the operation $e^{j2\pi fd}[\cdot e^{-j2\pi fd}]_+$ keeps exactly the coefficients allowed by the d -sample lookahead, annihilating the components of $S_{X\tilde{Y}}$ at times smaller than $-d$.



When $d = 0$, this becomes the causal Wiener filter. When $d > 0$, we may view the estimator as a causal filter with lookahead.

Time-series prediction in practice

We then shift from exact Wiener-filter formulas to a more practical question: how does time-series prediction typically get done in real applications? There are roughly two main modes, with a modern intermediate regime between them.

Mode I offline learning / training then separate inference / prediction	Mode I.5 in-context learning test-time training	Mode II online learning sequential decision making continual learning
---	---	---

In **Mode I**, one first trains or fits a model using training data, and then uses this fixed trained model to make predictions on the actual or testing data.

Approach A: use a neural network. A common modern approach is to fix a network size and architecture, run SGD or a related optimization method to find good weights/parameters on the training data, and deploy that trained network as a predictor on the actual or testing data. This is a flexible black-box approach and can perform well when enough data and compute are available.

Approach B: use a linear/stochastic-process model. A classical alternative is to posit a model such as ARMA(p, q), choose the model order (p, q), estimate the corresponding parameters from the training data, and use the optimal linear predictor associated with the fitted model.

Another possibility is to estimate the autocorrelation directly. Given observations $X(1), \dots, X(N)$, first define the empirical mean

$$\bar{X}_N = \frac{1}{N} \sum_{n=1}^N X(n).$$

For $0 \leq k \leq m(N) < N$, define

$$\hat{R}_X(k) = \frac{1}{N} \sum_{n=1}^{N-k} \left(X(n+k) - \bar{X}_N \right) \left(X(n) - \bar{X}_N \right).$$

For a real-valued process, extend this estimate to negative lags by setting

$$\hat{R}_X(-k) = \hat{R}_X(k),$$

and set $\hat{R}_X(k) = 0$ when $|k| > m(N)$.

The resulting sequence need not be a valid autocorrelation function: it may fail to be nonnegative definite, or equivalently its Fourier transform may take negative values. One approach is to repair the estimate in the spectral domain by enforcing nonnegativity. Alternatively, one may estimate the power spectral density directly:

$$\hat{S}(f) = \text{an estimate of the power spectral density at frequency } f.$$

The estimated autocorrelation or PSD can then be used to construct a corresponding linear predictor. Thus the linear prediction pipeline can be viewed either as *model fitting*, for example ARMA fitting, or as *second-order structure estimation*, autocorrelation / spectrum estimation, followed by Wiener-style prediction.

Figures of merit and tradeoffs

When comparing different prediction strategies, several figures of merit matter: amount of training data N , compute/time for training C_T , compute/time for inference/prediction C_I , prediction MSE, and interpretability of the model, i.e. “black box” versus “white box.”

Very roughly, neural networks tend to dominate in regimes where $N \gg 1$, $C_T \gg 1$, and $C_I \gg 1$, especially for substantially non-Gaussian data. Linear models tend to do better when N , C_T , and C_I are much smaller, and especially when the data are close to Gaussian. This is not an absolute rule, but it captures the main intuition behind why linear prediction remains highly relevant.

Why linear estimation?

The lecture emphasizes four reasons.

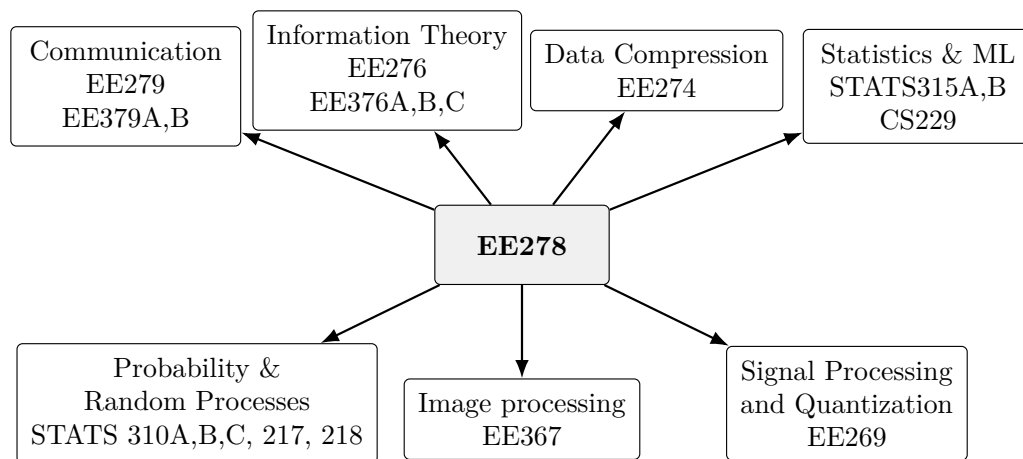
1. **It is simple to derive, compute, and implement.** Wiener filtering and related methods often reduce to explicit algebraic or spectral formulas.
2. **It needs only first and second moments.** One does not need a full distributional description of the data; means, autocorrelations, and spectra are enough.
3. **It is optimal under Gaussianity.** For jointly Gaussian variables or processes, the MMSE estimator is linear, so the optimal linear estimator is actually the true optimal estimator.
4. **It is conservative.** If one is unwilling to assume a complicated nonlinear or high-dimensional model, linear estimation is a robust middle ground. It uses the part of the statistical structure that is often most reliably estimable.

Central themes of the course

We end with a recap of the main topics that appeared throughout the course: hypothesis testing, detection, and related inference problems; random processes; stationarity, including strict-sense stationarity (SSS) and wide-sense stationarity (WSS); autocorrelation functions; process spectra/spectral densities; passage of WSS processes through linear systems; and linear estimation, including non-causal Wiener filtering, causal Wiener filtering, and Kalman filtering. These ideas form a coherent thread: second-order statistical structure, its transformation through systems, and the design of estimators that exploit that structure.

Related courses and directions

Here’s a rough map of how EE278 connects to neighboring areas.



The point of this picture is that EE278 sits at an intersection: probability and random processes feed into spectral analysis and estimation, which in turn connect to communication, information theory, signal processing, compression, image processing, and modern machine learning.