Machine Learning

A Bayesian and Optimization Perspective

Academic Press, 2015

Sergios Theodoridis[1]

[1]Dept. of Informatics and Telecommunications, National and Kapodistrian University of Athens, Athens, Greece.

Spring, 2015

Chapter 2
Probability and Stochastic Processes

Version I

Probability and Stochastic Precesses

The Notion of a Random Variable

- A random variable, x, is a variable whose variations are due to chance/randomness. A random variable can be considered as a function, which assigns a value to the outcome of an experiment. For example, in a coin tossing experiment, the corresponding random variable, x, can assume the values $x_1 = 0$ if the result of the experiment is "heads" and $x_2 = 1$ if the result is "tails."
- We will denote a random variable with a lower case roman, such as x, and the values it takes once an experiment has been performed, with mathmode italics, such as x.
- A random variable is described in terms of a set of probabilities if its values are of a discrete nature, or in terms of a probability density function (pdf) if its values lie anywhere within an interval of the real axis (non-countably infinite set).

Probability and Stochastic Precesses

The Notion of a Random Variable

- A random variable, x, is a variable whose variations are due to chance/randomness. A random variable can be considered as a function, which assigns a value to the outcome of an experiment. For example, in a coin tossing experiment, the corresponding random variable, x, can assume the values $x_1 = 0$ if the result of the experiment is "heads" and $x_2 = 1$ if the result is "tails."
- We will denote a random variable with a lower case roman, such as x, and the values it takes once an experiment has been performed, with mathmode italics, such as x.
- A random variable is described in terms of a set of probabilities if its values are of a discrete nature, or in terms of a probability density function (pdf) if its values lie anywhere within an interval of the real axis (non-countably infinite set).

Probability and Stochastic Precesses

The Notion of a Random Variable

- A random variable, x, is a variable whose variations are due to chance/randomness. A random variable can be considered as a function, which assigns a value to the outcome of an experiment. For example, in a coin tossing experiment, the corresponding random variable, x, can assume the values $x_1 = 0$ if the result of the experiment is "heads" and $x_2 = 1$ if the result is "tails."
- We will denote a random variable with a lower case roman, such as x, and the values it takes once an experiment has been performed, with mathmode italics, such as x.
- A random variable is described in terms of a set of probabilities if its values are of a discrete nature, or in terms of a probability density function (pdf) if its values lie anywhere within an interval of the real axis (non-countably infinite set).

Definitions of Probability

- Relative Frequency Definition: The probability, $P(A)$, of an event, A, is the limit

$$P(A) = \lim_{n \to \infty} \frac{n_A}{n},$$

where n is the total number of trials and n_A the number of times event A occurred.

- In practice, one can use

$$P(A) \approx \frac{n_A}{n},$$

for large enough values of n. However, care must be taken on how large n must be, especially when $P(A)$ is very small.

- From a physical reasoning point of view, probability can also be understood as a measure of our uncertainty concerning the corresponding event.

Definitions of Probability

- Relative Frequency Definition: The probability, $P(A)$, of an event, A, is the limit

$$P(A) = \lim_{n \to \infty} \frac{n_A}{n},$$

where n is the total number of trials and n_A the number of times event A occurred.

- In practice, one can use

$$P(A) \approx \frac{n_A}{n},$$

for large enough values of n. However, care must be taken on how large n must be, especially when $P(A)$ is very small.

- From a physical reasoning point of view, probability can also be understood as a measure of our uncertainty concerning the corresponding event.

Definitions of Probability

- Relative Frequency Definition: The probability, $P(A)$, of an event, A, is the limit

$$P(A) = \lim_{n \to \infty} \frac{n_A}{n},$$

where n is the total number of trials and n_A the number of times event A occurred.

- In practice, one can use

$$P(A) \approx \frac{n_A}{n},$$

for large enough values of n. However, care must be taken on how large n must be, especially when $P(A)$ is very small.

- From a physical reasoning point of view, probability can also be understood as a measure of our uncertainty concerning the corresponding event.

Definitions of Probability

- Axiomatic Definition: This definition of probability is traced back to 1933 to the work of Andrey Kolmogorov, who found a close connection between probability theory and the mathematical theory of sets and functions of a real variable, in the context of measure theory.
  1. The probability of an event A, $P(A)$ is a nonnegative number
  $$P(A) \geq 0.$$
  2. The probability of an event C, which is certain to occur, is equal to one,
  $$P(C) = 1.$$
  3. If two events, A and B, are mutually exclusive (they cannot occur simultaneously), then the probability of occurrence of either A or B (denoted as $A \cup B$) is given by
  $$P(A \cup B) = P(A) + P(B).$$
- These three defining properties (axioms), suffice to develop the rest of the theory.

Definitions of Probability

- Axiomatic Definition: This definition of probability is traced back to 1933 to the work of Andrey Kolmogorov, who found a close connection between probability theory and the mathematical theory of sets and functions of a real variable, in the context of measure theory.
  1. The probability of an event A, $P(A)$ is a nonnegative number
  $$P(A) \geq 0.$$
  2. The probability of an event C, which is certain to occur, is equal to one,
  $$P(C) = 1.$$
  3. If two events, A and B, are mutually exclusive (they cannot occur simultaneously), then the probability of occurrence of either A or B (denoted as $A \cup B$) is given by
  $$P(A \cup B) = P(A) + P(B).$$
- These three defining properties (axioms), suffice to develop the rest of the theory.

Definitions of Probability

- Axiomatic Definition: This definition of probability is traced back to 1933 to the work of Andrey Kolmogorov, who found a close connection between probability theory and the mathematical theory of sets and functions of a real variable, in the context of measure theory.
  1. The probability of an event A, $P(A)$ is a nonnegative number
  $$P(A) \geq 0.$$
  2. The probability of an event C, which is certain to occur, is equal to one,
  $$P(C) = 1.$$
  3. If two events, A and B, are mutually exclusive (they cannot occur simultaneously), then the probability of occurrence of either A or B (denoted as $A \cup B$) is given by
  $$P(A \cup B) = P(A) + P(B).$$
- These three defining properties (axioms), suffice to develop the rest of the theory.

Definitions of Probability

- Axiomatic Definition: This definition of probability is traced back to 1933 to the work of Andrey Kolmogorov, who found a close connection between probability theory and the mathematical theory of sets and functions of a real variable, in the context of measure theory.
  1. The probability of an event A, $P(A)$ is a nonnegative number
  $$P(A) \geq 0.$$
  2. The probability of an event C, which is certain to occur, is equal to one,
  $$P(C) = 1.$$
  3. If two events, A and B, are mutually exclusive (they cannot occur simultaneously), then the probability of occurrence of either A or B (denoted as $A \cup B$) is given by
  $$P(A \cup B) = P(A) + P(B).$$
- These three defining properties (axioms), suffice to develop the rest of the theory.

Discrete Random Variables

- A discrete random variable, x, can take any value from a finite or a countably infinite set, $\mathcal{X}$. The probability of an event "$\mathrm{x} = x$" is denoted as

$$P(\mathrm{x} = x) \text{ or simply } P(x).$$

- Assuming that no two values in $\mathcal{X}$ can occur simultaneously and that an experiment always returns a value, we have that

$$\sum_{x \in \mathcal{X}} P(x) = 1,$$

and $\mathcal{X}$ is known as the sample or state space.

- Joint probability: The joint probability of two events A and B to occur simultaneously is denoted as $P(A, B)$.
- Given two random variables $\mathrm{x} \in \mathcal{X}$ and $\mathrm{y} \in \mathcal{Y}$, the following sum rule is obtained

$$P(x) = \sum_{x \in \mathcal{Y}} P(x, y).$$

Discrete Random Variables

- A discrete random variable, x, can take any value from a finite or a countably infinite set, $\mathcal{X}$. The probability of an event "$\mathrm{x} = x$" is denoted as
$$P(\mathrm{x} = x) \text{ or simply } P(x).$$
- Assuming that no two values in $\mathcal{X}$ can occur simultaneously and that an experiment always returns a value, we have that
$$\sum_{x \in \mathcal{X}} P(x) = 1,$$
and $\mathcal{X}$ is known as the sample or state space.
- Joint probability: The joint probability of two events A and B to occur simultaneously is denoted as $P(A, B)$.
- Given two random variables $\mathrm{x} \in \mathcal{X}$ and $\mathrm{y} \in \mathcal{Y}$, the following sum rule is obtained
$$P(x) = \sum_{x \in \mathcal{Y}} P(x, y).$$

Discrete Random Variables

- A discrete random variable, x, can take any value from a finite or a countably infinite set, $\mathcal{X}$. The probability of an event "$\mathrm{x} = x$" is denoted as
$$P(\mathrm{x} = x) \text{ or simply } P(x).$$
- Assuming that no two values in $\mathcal{X}$ can occur simultaneously and that an experiment always returns a value, we have that
$$\sum_{x \in \mathcal{X}} P(x) = 1,$$
and $\mathcal{X}$ is known as the sample or state space.
- Joint probability: The joint probability of two events A and B to occur simultaneously is denoted as $P(A, B)$.
- Given two random variables $\mathrm{x} \in \mathcal{X}$ and $\mathrm{y} \in \mathcal{Y}$, the following sum rule is obtained
$$P(x) = \sum_{x \in \mathcal{Y}} P(x, y).$$

Discrete Random Variables

- A discrete random variable, x, can take any value from a finite or a countably infinite set, $\mathcal{X}$. The probability of an event "$\mathrm{x} = x$" is denoted as
$$P(\mathrm{x} = x) \text{ or simply } P(x).$$
- Assuming that no two values in $\mathcal{X}$ can occur simultaneously and that an experiment always returns a value, we have that
$$\sum_{x \in \mathcal{X}} P(x) = 1,$$
and $\mathcal{X}$ is known as the sample or state space.
- Joint probability: The joint probability of two events A and B to occur simultaneously is denoted as $P(A, B)$.
- Given two random variables $\mathrm{x} \in \mathcal{X}$ and $\mathrm{y} \in \mathcal{Y}$, the following sum rule is obtained
$$P(x) = \sum_{x \in \mathcal{Y}} P(x, y).$$

Discrete Random Variables

- Conditional probability: The conditional probability of an event A given another event B, is denoted as $P(A|B)$ and it is defined as

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

- The above definition gives rise to the following product rule

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

- Expressed in terms of two random variables, x and y, we have

$$P(x,y) = P(x|y)P(y).$$

- $P(x)$ and $P(y)$ are also known as the marginal probabilities to be distinguished from the joint and the conditional ones.
- Statistical independence: Two random variables, x and y, are said to be statistically independent if and only if

$$P(x,y) = P(x)P(y),\ \forall x \in \mathcal{X},\ y \in \mathcal{Y}.$$

Discrete Random Variables

- Conditional probability: The conditional probability of an event A given another event B, is denoted as $P(A|B)$ and it is defined as

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

- The above definition gives rise to the following product rule

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

- Expressed in terms of two random variables, x and y, we have

$$P(x,y) = P(x|y)P(y).$$

- $P(x)$ and $P(y)$ are also known as the marginal probabilities to be distinguished from the joint and the conditional ones.
- Statistical independence: Two random variables, x and y, are said to be statistically independent if and only if

$$P(x,y) = P(x)P(y),\ \forall x \in \mathcal{X},\ y \in \mathcal{Y}.$$

Discrete Random Variables

- Conditional probability: The conditional probability of an event A given another event B, is denoted as $P(A|B)$ and it is defined as

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

- The above definition gives rise to the following product rule

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

- Expressed in terms of two random variables, x and y, we have

$$P(x,y) = P(x|y)P(y).$$

- $P(x)$ and $P(y)$ are also known as the marginal probabilities to be distinguished from the joint and the conditional ones.
- Statistical independence: Two random variables, x and y, are said to be statistically independent if and only if

$$P(x,y) = P(x)P(y),\ \forall x \in \mathcal{X},\ y \in \mathcal{Y}.$$

Discrete Random Variables

- Conditional probability: The conditional probability of an event A given another event B, is denoted as $P(A|B)$ and it is defined as

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

- The above definition gives rise to the following product rule

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

- Expressed in terms of two random variables, x and y, we have

$$P(x,y) = P(x|y)P(y).$$

- $P(x)$ and $P(y)$ are also known as the marginal probabilities to be distinguished from the joint and the conditional ones.
- Statistical independence: Two random variables, x and y, are said to be statistically independent if and only if

$$P(x,y) = P(x)P(y),\ \forall x \in \mathcal{X},\ y \in \mathcal{Y}.$$

Discrete Random Variables

- Conditional probability: The conditional probability of an event A given another event B, is denoted as $P(A|B)$ and it is defined as

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

- The above definition gives rise to the following product rule

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

- Expressed in terms of two random variables, x and y, we have

$$P(x,y) = P(x|y)P(y).$$

- $P(x)$ and $P(y)$ are also known as the marginal probabilities to be distinguished from the joint and the conditional ones.
- Statistical independence: Two random variables, x and y, are said to be statistically independent if and only if

$$P(x,y) = P(x)P(y),\ \forall x \in \mathcal{X},\ y \in \mathcal{Y}.$$

Discrete Random Variables

- Bayes Theorem: This important and elegant theorem is a direct consequence of the product rule and the symmetry property of the joint probability, i.e., $P(x, y) = P(y, x)$, and it is given by the following two equations,

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)},$$
$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}.$$

  This theorem plays a very important role in Machine Learning.

- What this theorem says is that, our uncertainty as expressed by the conditional probability $P(y|x)$ of an output variable, say y, given the value of an input, x, can be expressed the other way round; that is, in terms of the (uncertainty) conditional, $P(x|y)$ and the two marginal probabilities, $P(x)$ and $P(y)$.

Discrete Random Variables

- Bayes Theorem: This important and elegant theorem is a direct consequence of the product rule and the symmetry property of the joint probability, i.e., $P(x, y) = P(y, x)$, and it is given by the following two equations,

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)},$$
$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}.$$

  This theorem plays a very important role in Machine Learning.

- What this theorem says is that, our uncertainty as expressed by the conditional probability $P(y|x)$ of an output variable, say y, given the value of an input, x, can be expressed the other way round; that is, in terms of the (uncertainty) conditional, $P(x|y)$ and the two marginal probabilities, $P(x)$ and $P(y)$.

Continuous Random Variables

- A continuous random variable, x, can take values anywhere in an in interval in the real axis $\mathbb{R}$.
- The starting point to develop tools for describing such variables is to build bridges with what we know from the discrete random variables case.
- The cumulative distribution function (cdf) is defined as

$$F_{\mathrm{x}}(x) := P(\mathrm{x} \leq x).$$

  That is, cdf is the probability of the discrete event: "x takes any value less or equal to x".
- Thus, we can write

$$P(x_1 < \mathrm{x} \leq x_2) = F_{\mathrm{x}}(x_2) - F_{\mathrm{x}}(x_1).$$

- Assuming $F_{\mathrm{x}}(x)$ to be differentiable, the probability density function (pdf), denoted with lower case p, is defined as

$$p_{\mathrm{x}}(x) := \frac{dF_{\mathrm{x}}(x)}{dx}.$$

Continuous Random Variables

- A continuous random variable, x, can take values anywhere in an in interval in the real axis $\mathbb{R}$.
- The starting point to develop tools for describing such variables is to build bridges with what we know from the discrete random variables case.
- The cumulative distribution function (cdf) is defined as

$$F_{\mathrm{x}}(x) := P(\mathrm{x} \leq x).$$

  That is, cdf is the probability of the discrete event: "x takes any value less or equal to x".
- Thus, we can write

$$P(x_1 < \mathrm{x} \leq x_2) = F_{\mathrm{x}}(x_2) - F_{\mathrm{x}}(x_1).$$

- Assuming $F_{\mathrm{x}}(x)$ to be differentiable, the probability density function (pdf), denoted with lower case p, is defined as

$$p_{\mathrm{x}}(x) := \frac{dF_{\mathrm{x}}(x)}{dx}.$$

Continuous Random Variables

- A continuous random variable, x, can take values anywhere in an in interval in the real axis $\mathbb{R}$.
- The starting point to develop tools for describing such variables is to build bridges with what we know from the discrete random variables case.
- The cumulative distribution function (cdf) is defined as

$$F_{\mathrm{x}}(x) := P(\mathrm{x} \leq x).$$

  That is, cdf is the probability of the discrete event: "x takes any value less or equal to x".

- Thus, we can write

$$P(x_1 < \mathrm{x} \leq x_2) = F_{\mathrm{x}}(x_2) - F_{\mathrm{x}}(x_1).$$

- Assuming $F_{\mathrm{x}}(x)$ to be differentiable, the probability density function (pdf), denoted with lower case p, is defined as

$$p_{\mathrm{x}}(x) := \frac{dF_{\mathrm{x}}(x)}{dx}.$$

Continuous Random Variables

- A continuous random variable, x, can take values anywhere in an in interval in the real axis $\mathbb{R}$.
- The starting point to develop tools for describing such variables is to build bridges with what we know from the discrete random variables case.
- The cumulative distribution function (cdf) is defined as

$$F_{\mathrm{x}}(x) := P(\mathrm{x} \leq x).$$

  That is, cdf is the probability of the discrete event: "x takes any value less or equal to x".
- Thus, we can write

$$P(x_1 < \mathrm{x} \leq x_2) = F_{\mathrm{x}}(x_2) - F_{\mathrm{x}}(x_1).$$

- Assuming $F_{\mathrm{x}}(x)$ to be differentiable, the probability density function (pdf), denoted with lower case p, is defined as

$$p_{\mathrm{x}}(x) := \frac{dF_{\mathrm{x}}(x)}{dx}.$$

Continuous Random Variables

- A continuous random variable, x, can take values anywhere in an in interval in the real axis $\mathbb{R}$.
- The starting point to develop tools for describing such variables is to build bridges with what we know from the discrete random variables case.
- The cumulative distribution function (cdf) is defined as

$$F_{\mathrm{x}}(x) := P(\mathrm{x} \leq x).$$

  That is, cdf is the probability of the discrete event: "x takes any value less or equal to x".
- Thus, we can write

$$P(x_1 < \mathrm{x} \leq x_2) = F_{\mathrm{x}}(x_2) - F_{\mathrm{x}}(x_1).$$

- Assuming $F_{\mathrm{x}}(x)$ to be differentiable, the probability density function (pdf), denoted with lower case p, is defined as

$$p_{\mathrm{x}}(x) := \frac{dF_{\mathrm{x}}(x)}{dx}.$$

Continuous Random Variables

- Then, it is readily seen that

$$P(x_1 < \mathrm{x} \leq x_2) = \int_{x_1}^{x_2} p_{\mathrm{x}}(x)dx,$$

and

$$F_{\mathrm{x}}(x) = \int_{-\infty}^{x} p_{\mathrm{x}}(z)dz.$$

- Since an event is certain to occur in $-\infty < \mathrm{x} < +\infty$, we have that

$$\int_{-\infty}^{+\infty} p_{\mathrm{x}}(x)dx = 1.$$

- The previously stated rules, for the discrete random variables case, are also valid for the continuous ones, i.e.,

$$p(x|y) = \frac{p(x,y)}{p(y)}, \quad p_{\mathrm{x}}(x) = \int_{-\infty}^{+\infty} p(x,y)dy.$$

Continuous Random Variables

- Then, it is readily seen that

$$P(x_1 < \mathrm{x} \leq x_2) = \int_{x_1}^{x_2} p_{\mathrm{x}}(x)dx,$$

and

$$F_{\mathrm{x}}(x) = \int_{-\infty}^{x} p_{\mathrm{x}}(z)dz.$$

- Since an event is certain to occur in $-\infty < \mathrm{x} < +\infty$, we have that

$$\int_{-\infty}^{+\infty} p_{\mathrm{x}}(x)dx = 1.$$

- The previously stated rules, for the discrete random variables case, are also valid for the continuous ones, i.e.,

$$p(x|y) = \frac{p(x,y)}{p(y)}, \quad p_{\mathrm{x}}(x) = \int_{-\infty}^{+\infty} p(x,y)dy.$$

Continuous Random Variables

- Then, it is readily seen that

$$P(x_1 < \mathrm{x} \leq x_2) = \int_{x_1}^{x_2} p_{\mathrm{x}}(x)dx,$$

and

$$F_{\mathrm{x}}(x) = \int_{-\infty}^{x} p_{\mathrm{x}}(z)dz.$$

- Since an event is certain to occur in $-\infty < \mathrm{x} < +\infty$, we have that

$$\int_{-\infty}^{+\infty} p_{\mathrm{x}}(x)dx = 1.$$

- The previously stated rules, for the discrete random variables case, are also valid for the continuous ones, i.e.,

$$p(x|y) = \frac{p(x,y)}{p(y)}, \quad p_{\mathrm{x}}(x) = \int_{-\infty}^{+\infty} p(x,y)dy.$$

Mean, Variance and Covariance

- Two of the most useful quantities associated with a random variable, x, are:
  - The mean value, which is defined as:

$$\mathbb{E}[\mathrm{x}] := \int_{-\infty}^{+\infty} x p(x) dx.$$

  - The variance, which is defined as:

$$\sigma_x^2 := \int_{-\infty}^{+\infty} \left(x - \mathbb{E}[\mathrm{x}]\right)^2 p(x) dx,$$

    with integrations substituted by summations for the case of discrete variables, e.g.,

$$\mathbb{E}[\mathrm{x}] := \sum_{x \in \mathcal{X}} x P(x).$$

- More general, when a function f is involved, we have,

$$\mathbb{E}[f(\mathrm{x})] := \int_{-\infty}^{+\infty} f(x) p(x) dx.$$

Mean, Variance and Covariance

- Two of the most useful quantities associated with a random variable, x, are:
  - The mean value, which is defined as:

$$\mathbb{E}[\mathrm{x}] := \int_{-\infty}^{+\infty} x p(x) dx.$$

  - The variance, which is defined as:

$$\sigma_x^2 := \int_{-\infty}^{+\infty} \left(x - \mathbb{E}[\mathrm{x}]\right)^2 p(x) dx,$$

    with integrations substituted by summations for the case of discrete variables, e.g.,

$$\mathbb{E}[\mathrm{x}] := \sum_{x \in \mathcal{X}} x P(x).$$

- More general, when a function f is involved, we have,

$$\mathbb{E}[f(\mathrm{x})] := \int_{-\infty}^{+\infty} f(x) p(x) dx.$$

Mean, Variance and Covariance

- Two of the most useful quantities associated with a random variable, x, are:
  - The mean value, which is defined as:

$$\mathbb{E}[\mathrm{x}] := \int_{-\infty}^{+\infty} xp(x)dx.$$

  - The variance, which is defined as:

$$\sigma_x^2 := \int_{-\infty}^{+\infty} \left(x - \mathbb{E}[\mathrm{x}]\right)^2 p(x)dx,$$

    with integrations substituted by summations for the case of discrete variables, e.g.,

$$\mathbb{E}[\mathrm{x}] := \sum_{x \in \mathcal{X}} xP(x).$$

- More general, when a function f is involved, we have,

$$\mathbb{E}[f(\mathrm{x})] := \int_{-\infty}^{+\infty} f(x)p(x)dx.$$

Mean, Variance and Covariance

- It can readily be deduced from the respective definitions that, the mean value with respect to two random variables can be written as:

$$\mathbb{E}[\mathrm{x},\mathrm{y}] := \mathbb{E}_{\mathrm{x}}\Big[\mathbb{E}_{\mathrm{y}|\mathrm{x}}[f(\mathrm{x},\mathrm{y})]\Big].$$

- Given two random variables, x, y, their covariance is defined as

$$\mathrm{cov}(\mathrm{x},\mathrm{y}) := \mathbb{E}\big[(\mathrm{x}-\mathbb{E}[\mathrm{x}])(\mathrm{y}-\mathbb{E}[\mathrm{y}])\big].$$

- Their correlation is defined as

$$r_{x,y} := \mathbb{E}[\mathrm{xy}] = \mathrm{cov}(\mathrm{x},\mathrm{y}) - \mathbb{E}[\mathrm{x}]\mathbb{E}[\mathrm{y}].$$

- A random vector is a collection of random variables, $\mathbf{x} := [\mathrm{x}_1,\ldots,\mathrm{x}_l]^T$ and their joint pdf is denoted as

$$p(\boldsymbol{x}) = p(x_1,\ldots,x_l),\ \boldsymbol{x} = [x_1,\ldots,x_l]^T.$$

Mean, Variance and Covariance

- It can readily be deduced from the respective definitions that, the mean value with respect to two random variables can be written as:

$$\mathbb{E}[\mathrm{x},\mathrm{y}] := \mathbb{E}_{\mathrm{x}}\Big[\mathbb{E}_{\mathrm{y}|\mathrm{x}}[f(\mathrm{x},\mathrm{y})]\Big].$$

- Given two random variables, $\mathrm{x},\ \mathrm{y}$, their covariance is defined as

$$\mathrm{cov}(\mathrm{x},\mathrm{y}) := \mathbb{E}\Big[(\mathrm{x}-\mathbb{E}[\mathrm{x}])(\mathrm{y}-\mathbb{E}[\mathrm{y}])\Big].$$

- Their correlation is defined as

$$r_{x,y} := \mathbb{E}[\mathrm{xy}] = \mathrm{cov}(\mathrm{x},\mathrm{y}) - \mathbb{E}[\mathrm{x}]\mathbb{E}[\mathrm{y}].$$

- A random vector is a collection of random variables, $\mathbf{x} := [\mathrm{x}_1,\ldots,\mathrm{x}_l]^T$ and their joint pdf is denoted as

$$p(\boldsymbol{x}) = p(x_1,\ldots,x_l),\ \boldsymbol{x} = [x_1,\ldots,x_l]^T.$$

Mean, Variance and Covariance

- It can readily be deduced from the respective definitions that, the mean value with respect to two random variables can be written as:

$$\mathbb{E}[\mathrm{x},\mathrm{y}] := \mathbb{E}_{\mathrm{x}}\Big[\mathbb{E}_{\mathrm{y}|\mathrm{x}}[f(\mathrm{x},\mathrm{y})]\Big].$$

- Given two random variables, $\mathrm{x},\ \mathrm{y}$, their covariance is defined as

$$\mathrm{cov}(\mathrm{x},\mathrm{y}) := \mathbb{E}\Big[(\mathrm{x}-\mathbb{E}[\mathrm{x}])(\mathrm{y}-\mathbb{E}[\mathrm{y}])\Big].$$

- Their correlation is defined as

$$r_{x,y} := \mathbb{E}[\mathrm{xy}] = \mathrm{cov}(\mathrm{x},\mathrm{y}) - \mathbb{E}[\mathrm{x}]\mathbb{E}[\mathrm{y}].$$

- A random vector is a collection of random variables, $\mathbf{x} := [\mathrm{x}_1,\ldots,\mathrm{x}_l]^T$ and their joint pdf is denoted as

$$p(\boldsymbol{x}) = p(x_1,\ldots,x_l),\ \boldsymbol{x} = [x_1,\ldots,x_l]^T.$$

Mean, Variance and Covariance

- The covariance matrix of a random vector, $\mathbf{x} \in \mathbb{R}^l$, is defined as

$$\text{Cov}(\mathbf{x}) := \mathbb{E}\big[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^T\big],$$

or

$$\text{Cov}(\mathbf{x}) = \begin{bmatrix} \text{cov}(\mathrm{x}_1, \mathrm{x}_1) & \dots & \text{cov}(\mathrm{x}_1, \mathrm{x}_l) \\ \vdots & \ddots & \vdots \\ \text{cov}(\mathrm{x}_l, \mathrm{x}_1) & \dots & \text{cov}(\mathrm{x}_l, \mathrm{x}_l) \end{bmatrix}.$$

- Similarly, the correlation matrix of a random vector, $\mathbf{x}$, is defined as

$$R_x := \mathbb{E}\left[\mathbf{x}\mathbf{x}^T\right],$$

or

$$\begin{aligned} R_x &= \begin{bmatrix} \mathbb{E}[\mathrm{x}_1, \mathrm{x}_1] & \dots & \mathbb{E}[\mathrm{x}_1, \mathrm{x}_l] \\ \vdots & \ddots & \vdots \\ \mathbb{E}[\mathrm{x}_l, \mathrm{x}_1] & \dots & \mathbb{E}[\mathrm{x}_l, \mathrm{x}_l] \end{bmatrix} \\ &= \text{Cov}(\mathbf{x}) + \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{x}^T]. \end{aligned}$$

Mean, Variance and Covariance

- The covariance matrix of a random vector, $\mathbf{x} \in \mathbb{R}^l$, is defined as

$$\text{Cov}(\mathbf{x}) := \mathbb{E}\big[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^T\big],$$

or

$$\text{Cov}(\mathbf{x}) = \begin{bmatrix} \text{cov}(\mathrm{x}_1, \mathrm{x}_1) & \dots & \text{cov}(\mathrm{x}_1, \mathrm{x}_l) \\ \vdots & \ddots & \vdots \\ \text{cov}(\mathrm{x}_l, \mathrm{x}_1) & \dots & \text{cov}(\mathrm{x}_l, \mathrm{x}_l) \end{bmatrix}.$$

- Similarly, the correlation matrix of a random vector, $\mathbf{x}$, is defined as

$$R_x := \mathbb{E}\left[\mathbf{x}\mathbf{x}^T\right],$$

or

$$\begin{aligned} R_x &= \begin{bmatrix} \mathbb{E}[\mathrm{x}_1, \mathrm{x}_1] & \dots & \mathbb{E}[\mathrm{x}_1, \mathrm{x}_l] \\ \vdots & \ddots & \vdots \\ \mathbb{E}[\mathrm{x}_l, \mathrm{x}_1] & \dots & \mathbb{E}[\mathrm{x}_l, \mathrm{x}_l] \end{bmatrix} \\ &= \text{Cov}(\mathbf{x}) + \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{x}^T]. \end{aligned}$$

Mean, Variance and Covariance

- Important Property: The covariance as well as the correlation matrices are positive semidefinite.
- A matrix A is called positive semidefinite, if

$$\boldsymbol{y}^T A \boldsymbol{y} \geq 0, \ \forall \boldsymbol{y} \in \mathbb{R}^l,$$

and it is called positive definite if the inequality is a strict one.

- **Proof:** For the covariance matrix, we have

$$\boldsymbol{y}^T \mathbb{E}\left[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^T\right] \boldsymbol{y} = \mathbb{E}\left[\left(\boldsymbol{y}^T (\mathbf{x} - \mathbb{E}[\mathbf{x}])\right)^2\right] \geq 0.$$

- Complex random variables: A complex random variable, $\mathrm{z} \in \mathbb{C}$, is defined as the sum

$$\mathrm{z} := \mathrm{x} + j\mathrm{y}, \ \mathrm{x}, \mathrm{y} \in \mathbb{R}, \ \text{where } j := \sqrt{-1}.$$

- The pdf $p(z)$ (probability $P(z)$) of a complex random variable is defined as the joint pdf of the respective real random variables,

$$p(z) := p(x, y), \ \text{or for discrete r.v.s, } \ P(z) := P(x, y).$$

Mean, Variance and Covariance

- Important Property: The covariance as well as the correlation matrices are positive semidefinite.
- A matrix A is called positive semidefinite, if

$$\boldsymbol{y}^T A \boldsymbol{y} \geq 0, \ \forall \boldsymbol{y} \in \mathbb{R}^l,$$

and it is called positive definite if the inequality is a strict one.

- **Proof:** For the covariance matrix, we have

$$\boldsymbol{y}^T \mathbb{E}\left[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^T\right] \boldsymbol{y} = \mathbb{E}\left[\left(\boldsymbol{y}^T (\mathbf{x} - \mathbb{E}[\mathbf{x}])\right)^2\right] \geq 0.$$

- Complex random variables: A complex random variable, $\mathrm{z} \in \mathbb{C}$, is defined as the sum

$$\mathrm{z} := \mathrm{x} + j\mathrm{y}, \ \mathrm{x}, \mathrm{y} \in \mathbb{R}, \ \text{where } j := \sqrt{-1}.$$

- The pdf $p(z)$ (probability $P(z)$) of a complex random variable is defined as the joint pdf of the respective real random variables,

$$p(z) := p(x, y), \ \text{or for discrete r.v.s, } P(z) := P(x, y).$$

Mean, Variance and Covariance

- Important Property: The covariance as well as the correlation matrices are positive semidefinite.
- A matrix A is called positive semidefinite, if

$$\boldsymbol{y}^T A \boldsymbol{y} \geq 0, \ \forall \boldsymbol{y} \in \mathbb{R}^l,$$

  and it is called positive definite if the inequality is a strict one.
- **Proof:** For the covariance matrix, we have

$$\boldsymbol{y}^T \mathbb{E}\left[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^T\right] \boldsymbol{y} = \mathbb{E}\left[\left(\boldsymbol{y}^T(\mathbf{x} - \mathbb{E}[\mathbf{x}])\right)^2\right] \geq 0.$$

- Complex random variables: A complex random variable, $\mathrm{z} \in \mathbb{C}$, is defined as the sum

$$\mathrm{z} := \mathrm{x} + j\mathrm{y}, \ \mathrm{x}, \mathrm{y} \in \mathbb{R}, \ \text{where } j := \sqrt{-1}.$$

- The pdf $p(z)$ (probability $P(z)$) of a complex random variable is defined as the joint pdf of the respective real random variables,

$$p(z) := p(x, y), \ \text{or for discrete r.v.s, } P(z) := P(x, y).$$

Mean, Variance and Covariance

- Important Property: The covariance as well as the correlation matrices are positive semidefinite.
- A matrix A is called positive semidefinite, if

$$\boldsymbol{y}^T A \boldsymbol{y} \geq 0, \ \forall \boldsymbol{y} \in \mathbb{R}^l,$$

  and it is called positive definite if the inequality is a strict one.
- **Proof:** For the covariance matrix, we have

$$\boldsymbol{y}^T \mathbb{E}\left[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^T\right] \boldsymbol{y} = \mathbb{E}\left[\left(\boldsymbol{y}^T(\mathbf{x} - \mathbb{E}[\mathbf{x}])\right)^2\right] \geq 0.$$

- Complex random variables: A complex random variable, $\mathrm{z} \in \mathbb{C}$, is defined as the sum

$$\mathrm{z} := \mathrm{x} + j\mathrm{y}, \ \mathrm{x}, \mathrm{y} \in \mathbb{R}, \text{ where } j := \sqrt{-1}.$$

- The pdf $p(z)$ (probability $P(z)$) of a complex random variable is defined as the joint pdf of the respective real random variables,

$$p(z) := p(x, y), \text{ or for discrete r.v.s, } P(z) := P(x, y).$$

Mean, Variance and Covariance

- Important Property: The covariance as well as the correlation matrices are positive semidefinite.
- A matrix A is called positive semidefinite, if

$$\boldsymbol{y}^T A \boldsymbol{y} \geq 0, \ \forall \boldsymbol{y} \in \mathbb{R}^l,$$

  and it is called positive definite if the inequality is a strict one.
- **Proof:** For the covariance matrix, we have

$$\boldsymbol{y}^T \mathbb{E}\left[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^T\right] \boldsymbol{y} = \mathbb{E}\left[\left(\boldsymbol{y}^T(\mathbf{x} - \mathbb{E}[\mathbf{x}])\right)^2\right] \geq 0.$$

- Complex random variables: A complex random variable, $\mathrm{z} \in \mathbb{C}$, is defined as the sum

$$\mathrm{z} := \mathrm{x} + j\mathrm{y}, \ \mathrm{x}, \mathrm{y} \in \mathbb{R}, \text{ where } j := \sqrt{-1}.$$

- The pdf $p(z)$ (probability $P(z)$) of a complex random variable is defined as the joint pdf of the respective real random variables,

$$p(z) := p(x, y), \text{ or for discrete r.v.s, } P(z) := P(x, y).$$

Complex Variables

- For complex random variables, the notions of mean and covariance are defined as,

$$\mathbb{E}[\mathrm{z}] := \mathbb{E}[\mathrm{x}] + j\mathbb{E}[\mathrm{y}], \text{ and}$$

$$\mathrm{cov}(\mathrm{z}_1, \mathrm{z}_2) := \mathbb{E}\left[\left(\mathrm{z}_1 - \mathbb{E}[\mathrm{z}_1]\right)\left(\mathrm{z}_2 - \mathbb{E}[\mathrm{z}_2]\right)^*\right],$$

where " $*$ " denotes complex conjugation.

- The latter definition leads to the variance of a complex variable,

$$\sigma_z^2 = \mathbb{E}\left[\left|\mathrm{z} - \mathbb{E}[\mathrm{z}]\right|^2\right] = \mathbb{E}\left[\left|\mathrm{z}\right|^2\right] - \left|\mathbb{E}\left[\mathrm{z}\right]\right|^2.$$

- Similarly, for complex random vectors, $\mathbf{z} = \mathbf{x} + j\mathbf{y} \in \mathbb{C}^l$, we have

$$p(\boldsymbol{z}) := p(x_1, \ldots, x_l, y_1, \ldots, y_l),$$

where $x_i, y_i,\ i = 1, 2, \ldots, l$, are the components of the involved real vectors, respectively.

- The covariance and correlation matrices are similarly defined, in terms of the Hermitian transposition,

$$\mathrm{Cov}(\mathbf{z}) := \mathbb{E}\left[\left(\mathbf{z} - \mathbb{E}[\mathbf{z}]\right)\left(\mathbf{z} - \mathbb{E}[\mathbf{z}]\right)^H\right],\ R_z := \mathbb{E}[\mathbf{z}\mathbf{z}^H].$$

Complex Variables

- For complex random variables, the notions of mean and covariance are defined as,

$$\mathbb{E}[\mathrm{z}] := \mathbb{E}[\mathrm{x}] + j\mathbb{E}[\mathrm{y}], \text{ and}$$

$$\mathrm{cov}(\mathrm{z}_1, \mathrm{z}_2) := \mathbb{E}\left[\big(\mathrm{z}_1 - \mathbb{E}[\mathrm{z}_1]\big)\big(\mathrm{z}_2 - \mathbb{E}[\mathrm{z}_2]\big)^*\right],$$

  where " $*$ " denotes complex conjugation.

- The latter definition leads to the variance of a complex variable,

$$\sigma_z^2 = \mathbb{E}\left[\big|\mathrm{z} - \mathbb{E}[\mathrm{z}]\big|^2\right] = \mathbb{E}\left[|\mathrm{z}|^2\right] - \big|\mathbb{E}[\mathrm{z}]\big|^2.$$

- Similarly, for complex random vectors, $\mathbf{z} = \mathbf{x} + j\mathbf{y} \in \mathbb{C}^l$, we have

$$p(\boldsymbol{z}) := p(x_1, \ldots, x_l, y_1, \ldots, y_l),$$

  where $x_i, y_i,\ i = 1, 2, \ldots, l$, are the components of the involved real vectors, respectively.

- The covariance and correlation matrices are similarly defined, in terms of the Hermitian transposition,

$$\mathrm{Cov}(\mathbf{z}) := \mathbb{E}\left[\big(\mathbf{z} - \mathbb{E}[\mathbf{z}]\big)\big(\mathbf{z} - \mathbb{E}[\mathbf{z}]\big)^H\right],\ R_z := \mathbb{E}[\mathbf{z}\mathbf{z}^H].$$

Complex Variables

- For complex random variables, the notions of mean and covariance are defined as,

$$\mathbb{E}[\mathrm{z}] := \mathbb{E}[\mathrm{x}] + j\mathbb{E}[\mathrm{y}], \text{ and}$$

$$\mathrm{cov}(\mathrm{z}_1, \mathrm{z}_2) := \mathbb{E}\left[\left(\mathrm{z}_1 - \mathbb{E}[\mathrm{z}_1]\right)\left(\mathrm{z}_2 - \mathbb{E}[\mathrm{z}_2]\right)^*\right],$$

  where " $*$ " denotes complex conjugation.

- The latter definition leads to the variance of a complex variable,

$$\sigma_z^2 = \mathbb{E}\left[\left|\mathrm{z} - \mathbb{E}[\mathrm{z}]\right|^2\right] = \mathbb{E}\left[\left|\mathrm{z}\right|^2\right] - \left|\mathbb{E}\left[\mathrm{z}\right]\right|^2.$$

- Similarly, for complex random vectors, $\mathbf{z} = \mathbf{x} + j\mathbf{y} \in \mathbb{C}^l$, we have

$$p(\boldsymbol{z}) := p(x_1, \ldots, x_l, y_1, \ldots, y_l),$$

  where $x_i, y_i,\ i = 1, 2, \ldots, l$, are the components of the involved real vectors, respectively.

- The covariance and correlation matrices are similarly defined, in terms of the Hermitian transposition,

$$\mathrm{Cov}(\mathbf{z}) := \mathbb{E}\left[\left(\mathbf{z} - \mathbb{E}[\mathbf{z}]\right)\left(\mathbf{z} - \mathbb{E}[\mathbf{z}]\right)^H\right],\ R_z := \mathbb{E}[\mathbf{z}\mathbf{z}^H].$$

Complex Variables

- For complex random variables, the notions of mean and covariance are defined as,

$$\mathbb{E}[\mathrm{z}] := \mathbb{E}[\mathrm{x}] + j\mathbb{E}[\mathrm{y}], \text{ and}$$

$$\mathrm{cov}(\mathrm{z}_1, \mathrm{z}_2) := \mathbb{E}\left[\left(\mathrm{z}_1 - \mathbb{E}[\mathrm{z}_1]\right)\left(\mathrm{z}_2 - \mathbb{E}[\mathrm{z}_2]\right)^*\right],$$

  where " $*$ " denotes complex conjugation.

- The latter definition leads to the variance of a complex variable,

$$\sigma_z^2 = \mathbb{E}\left[\left|\mathrm{z} - \mathbb{E}[\mathrm{z}]\right|^2\right] = \mathbb{E}\left[\left|\mathrm{z}\right|^2\right] - \left|\mathbb{E}\left[\mathrm{z}\right]\right|^2.$$

- Similarly, for complex random vectors, $\mathbf{z} = \mathbf{x} + j\mathbf{y} \in \mathbb{C}^l$, we have

$$p(\boldsymbol{z}) := p(x_1, \ldots, x_l, y_1, \ldots, y_l),$$

  where $x_i, y_i,\ i = 1, 2, \ldots, l$, are the components of the involved real vectors, respectively.

- The covariance and correlation matrices are similarly defined, in terms of the Hermitian transposition,

$$\mathrm{Cov}(\mathbf{z}) := \mathbb{E}\left[\left(\mathbf{z} - \mathbb{E}[\mathbf{z}]\right)\left(\mathbf{z} - \mathbb{E}[\mathbf{z}]\right)^H\right],\ R_z := \mathbb{E}[\mathbf{z}\mathbf{z}^H].$$

Complex Variables

- For complex random variables, the notions of mean and covariance are defined as,

$$\mathbb{E}[\mathrm{z}] := \mathbb{E}[\mathrm{x}] + j\mathbb{E}[\mathrm{y}], \text{ and}$$

$$\mathrm{cov}(\mathrm{z}_1, \mathrm{z}_2) := \mathbb{E}\left[\left(\mathrm{z}_1 - \mathbb{E}[\mathrm{z}_1]\right)\left(\mathrm{z}_2 - \mathbb{E}[\mathrm{z}_2]\right)^*\right],$$

  where " $*$ " denotes complex conjugation.

- The latter definition leads to the variance of a complex variable,

$$\sigma_z^2 = \mathbb{E}\left[\left|\mathrm{z} - \mathbb{E}[\mathrm{z}]\right|^2\right] = \mathbb{E}\left[\left|\mathrm{z}\right|^2\right] - \left|\mathbb{E}\left[\mathrm{z}\right]\right|^2.$$

- Similarly, for complex random vectors, $\mathbf{z} = \mathbf{x} + j\mathbf{y} \in \mathbb{C}^l$, we have

$$p(\boldsymbol{z}) := p(x_1, \ldots, x_l, y_1, \ldots, y_l),$$

  where $x_i, y_i,\ i = 1, 2, \ldots, l$, are the components of the involved real vectors, respectively.

- The covariance and correlation matrices are similarly defined, in terms of the Hermitian transposition,

$$\mathrm{Cov}(\mathbf{z}) := \mathbb{E}\left[\left(\mathbf{z} - \mathbb{E}[\mathbf{z}]\right)\left(\mathbf{z} - \mathbb{E}[\mathbf{z}]\right)^H\right],\ R_z := \mathbb{E}[\mathbf{z}\mathbf{z}^H].$$

Transformation of Random Variables

- Let $\mathbf{x}$, $\mathbf{y}$ be two random vectors, which are related via a transform,

$$\mathbf{y} = \boldsymbol{f}(\mathbf{x}).$$

- The vector function $\boldsymbol{f}$ is assumed to be invertible. That is, there is a uniquely defined vector function, denoted as $\boldsymbol{f}^{-1}$, so that,

$$\mathbf{x} = \boldsymbol{f}^{-1}(\mathbf{y}).$$

- Given the pdf, $p_{\mathbf{x}}(\boldsymbol{x})$, of $\mathbf{x}$, it can be shown that,

$$p_{\mathbf{y}}(\boldsymbol{y}) = \frac{p_{\mathbf{x}}(\boldsymbol{x})}{\left|\det(J(\mathbf{y};\mathbf{x}))\right|}\Bigg|_{\boldsymbol{x}=\boldsymbol{f}^{-1}(\boldsymbol{y})},$$

where the Jacobian matrix of the transformation is defined as

$$J(\mathbf{y};\mathbf{x}) := \frac{\partial(y_1, y_2, \ldots, y_l)}{\partial(x_1, x_2, \ldots, x_l)} := \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_l} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_l}{\partial x_1} & \cdots & \frac{\partial y_l}{\partial x_l} \end{bmatrix}.$$

Transformation of Random Variables

- Let $\mathbf{x}$, $\mathbf{y}$ be two random vectors, which are related via a transform,

$$\mathbf{y} = \boldsymbol{f}(\mathbf{x}).$$

- The vector function $\boldsymbol{f}$ is assumed to be invertible. That is, there is a uniquely defined vector function, denoted as $\boldsymbol{f}^{-1}$, so that,

$$\mathbf{x} = \boldsymbol{f}^{-1}(\mathbf{y}).$$

- Given the pdf, $p_{\mathbf{x}}(\boldsymbol{x})$, of $\mathbf{x}$, it can be shown that,

$$p_{\mathbf{y}}(\boldsymbol{y}) = \frac{p_{\mathbf{x}}(\boldsymbol{x})}{\left|\det(J(\mathbf{y};\mathbf{x}))\right|}\Bigg|_{\boldsymbol{x}=\boldsymbol{f}^{-1}(\boldsymbol{y})},$$

where the Jacobian matrix of the transformation is defined as

$$J(\mathbf{y};\mathbf{x}) := \frac{\partial(y_1, y_2, \ldots, y_l)}{\partial(x_1, x_2, \ldots, x_l)} := \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_l} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_l}{\partial x_1} & \cdots & \frac{\partial y_l}{\partial x_l} \end{bmatrix}.$$

Transformation of Random Variables

- Let $\mathbf{x}$, $\mathbf{y}$ be two random vectors, which are related via a transform,

$$\mathbf{y} = \boldsymbol{f}(\mathbf{x}).$$

- The vector function $\boldsymbol{f}$ is assumed to be invertible. That is, there is a uniquely defined vector function, denoted as $\boldsymbol{f}^{-1}$, so that,

$$\mathbf{x} = \boldsymbol{f}^{-1}(\mathbf{y}).$$

- Given the pdf, $p_{\mathbf{x}}(\boldsymbol{x})$, of $\mathbf{x}$, it can be shown that,

$$p_{\mathbf{y}}(\boldsymbol{y}) = \frac{p_{\mathbf{x}}(\boldsymbol{x})}{\left|\det(J(\mathbf{y};\mathbf{x}))\right|}\Bigg|_{\boldsymbol{x}=\boldsymbol{f}^{-1}(\boldsymbol{y})},$$

  where the Jacobian matrix of the transformation is defined as

$$J(\mathbf{y};\mathbf{x}) := \frac{\partial(y_1, y_2, \ldots, y_l)}{\partial(x_1, x_2, \ldots, x_l)} := \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_l} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_l}{\partial x_1} & \cdots & \frac{\partial y_l}{\partial x_l} \end{bmatrix}.$$

Transformation of Random Variables

- We have denoted as $\det(\cdot)$ the determinant of a matrix and $|\cdot|$ the absolute value.
- For the case of two random variables, the previous formula becomes

$$p_{\mathrm{y}}(y) = \left. \frac{p_{\mathrm{x}}(x)}{\left|\frac{dy}{dx}\right|} \right|_{x=f^{-1}(y)}.$$

- The proof of the previous formula can be justified by carefully looking at the following figure and noting that $p(x)|\Delta x| = p(y)|\Delta y|$.

Transformation of Random Variables

- We have denoted as $\det(\cdot)$ the determinant of a matrix and $|\cdot|$ the absolute value.
- For the case of two random variables, the previous formula becomes

$$p_{\mathrm{y}}(y) = \left. \frac{p_{\mathrm{x}}(x)}{\left| \frac{dy}{dx} \right|} \right|_{x=f^{-1}(y)}.$$

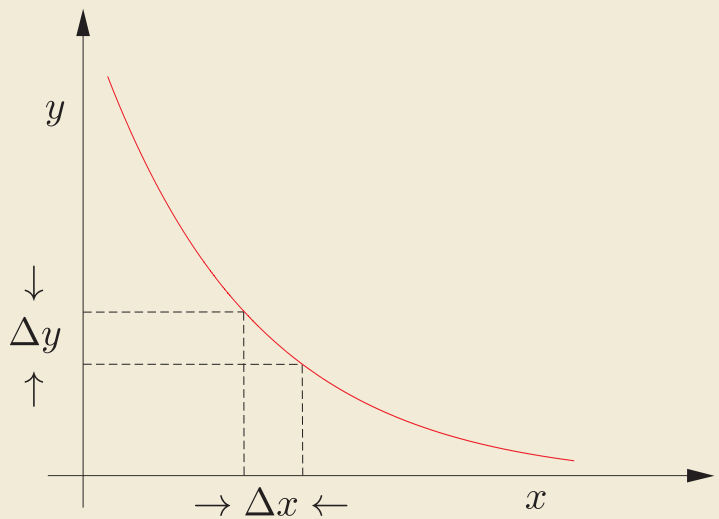
- The proof of the previous formula can be justified by carefully looking at the following figure and noting that $p(x)|\Delta x| = p(y)|\Delta y|$.

Transformation of Random Variables

- We have denoted as $\det(\cdot)$ the determinant of a matrix and $|\cdot|$ the absolute value.
- For the case of two random variables, the previous formula becomes

$$p_{\mathrm{y}}(y) = \left. \frac{p_{\mathrm{x}}(x)}{\left| \frac{dy}{dx} \right|} \right|_{x=f^{-1}(y)}.$$

- The proof of the previous formula can be justified by carefully looking at the following figure and noting that $p(x)|\Delta x| = p(y)|\Delta y|$.



Example

- Let the two random vectors $\mathbf{x}$ and $\mathbf{y}$ be related by a linear transform, via an invertible matrix A,

$$\mathbf{y} = A\mathbf{x}.$$

- Then, it is easily checked out that the Jacobian matrix is equal to the matrix A,

$$J(\mathbf{y};\mathbf{x}) = A.$$

- Thus, we readily obtain that,

$$p_{\mathbf{y}}(\boldsymbol{y}) = \frac{p_{\mathbf{x}}(A^{-1}\boldsymbol{x})}{|\det A|}.$$

Typical Distributions for Discrete Variables

- The Bernoulli distribution: A random variable is said to be distributed according to a Bernoulli distribution, if it is binary, $\mathcal{X} = \{0, 1\}$, with

$$P(\mathrm{x} = 1) = p, \quad P(\mathrm{x} = 0) = 1 - p.$$

- In a more compact way, we write that $\mathrm{x} \sim \mathrm{Bern}(x|p)$ where

$$P(x) = \mathrm{Bern}(x; p) := p^x (1 - p)^{1-x}.$$

- Its mean value is equal to:

$$\mathbb{E}[\mathrm{x}] = 1p + 0(1 - p) = p.$$

- Its variance is equal to:

$$\sigma_x^2 = (1 - p)^2 p + p^2 (1 - p) = p(1 - p).$$

Typical Distributions for Discrete Variables

- The Bernoulli distribution: A random variable is said to be distributed according to a Bernoulli distribution, if it is binary, $\mathcal{X} = \{0, 1\}$, with

$$P(\mathrm{x} = 1) = p, \quad P(\mathrm{x} = 0) = 1 - p.$$

- In a more compact way, we write that $\mathrm{x} \sim \mathrm{Bern}(x|p)$ where

$$P(x) = \mathrm{Bern}(x; p) := p^x (1 - p)^{1-x}.$$

- Its mean value is equal to:

$$\mathbb{E}[\mathrm{x}] = 1p + 0(1 - p) = p.$$

- Its variance is equal to:

$$\sigma_x^2 = (1 - p)^2 p + p^2 (1 - p) = p(1 - p).$$

Typical Distributions for Discrete Variables

- The Bernoulli distribution: A random variable is said to be distributed according to a Bernoulli distribution, if it is binary, $\mathcal{X} = \{0, 1\}$, with

$$P(\mathrm{x} = 1) = p, \quad P(\mathrm{x} = 0) = 1 - p.$$

- In a more compact way, we write that $\mathrm{x} \sim \mathrm{Bern}(x|p)$ where

$$P(x) = \mathrm{Bern}(x; p) := p^x (1 - p)^{1-x}.$$

- Its mean value is equal to:

$$\mathbb{E}[\mathrm{x}] = 1p + 0(1 - p) = p.$$

- Its variance is equal to:

$$\sigma_x^2 = (1 - p)^2 p + p^2 (1 - p) = p(1 - p).$$

Typical Distributions for Discrete Variables

- The Binomial Distribution: A random variable, x, is said to follow a binomial distribution, with parameters n, p and we write $\mathrm{x} \sim \mathsf{Bin}(x|n,p)$, if $\mathcal{X} = \{0, 1, \ldots, n\}$ and

$$P(\mathrm{x} = k) := \mathsf{Bin}(k|n,p) = \left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k}, \ k = 0, 1, \ldots, n.$$

- For example, this distribution models the times that head occurs in n successive trials, where $P(\mathsf{Head}) = p$.
- The binomial is a generalization of the Bernoulli distribution, which results if we set $n = 1$.
- The mean and variance of the binomial distribution are:

$$\mathbb{E}[\mathrm{x}] = np, \ \text{and} \ \sigma_x^2 = np(1-p).$$

Typical Distributions for Discrete Variables

- The Binomial Distribution: A random variable, x, is said to follow a binomial distribution, with parameters n, p and we write $\mathrm{x} \sim \mathsf{Bin}(x|n,p)$, if $\mathcal{X} = \{0, 1, \ldots, n\}$ and

$$P(\mathrm{x} = k) := \mathsf{Bin}(k|n,p) = \left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k}, \ k = 0, 1, \ldots, n.$$

- For example, this distribution models the times that head occurs in n successive trials, where $P(\mathsf{Head}) = p$.
- The binomial is a generalization of the Bernoulli distribution, which results if we set $n = 1$.
- The mean and variance of the binomial distribution are:

$$\mathbb{E}[\mathrm{x}] = np, \ \text{and} \ \sigma_x^2 = np(1-p).$$

Typical Distributions for Discrete Variables

- The Binomial Distribution: A random variable, x, is said to follow a binomial distribution, with parameters n, p and we write $\mathrm{x} \sim \mathrm{Bin}(x|n,p)$, if $\mathcal{X} = \{0, 1, \ldots, n\}$ and

$$P(\mathrm{x} = k) := \mathrm{Bin}(k|n,p) = \left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k}, \ k = 0, 1, \ldots, n.$$

- For example, this distribution models the times that head occurs in n successive trials, where $P(\mathrm{Head}) = p$.
- The binomial is a generalization of the Bernoulli distribution, which results if we set $n = 1$.
- The mean and variance of the binomial distribution are:

$$\mathbb{E}[\mathrm{x}] = np, \ \text{and} \ \sigma_x^2 = np(1-p).$$

Typical Distributions for Discrete Variables

- The Binomial Distribution: A random variable, x, is said to follow a binomial distribution, with parameters n, p and we write $\mathrm{x} \sim \mathsf{Bin}(x|n,p)$, if $\mathcal{X} = \{0, 1, \ldots, n\}$ and

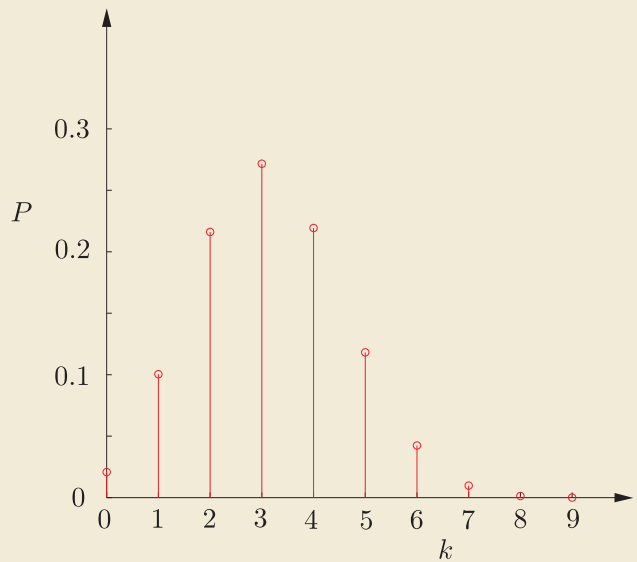
$$P(\mathrm{x} = k) := \mathsf{Bin}(k|n,p) = \left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k}, \ k = 0, 1, \ldots, n.$$

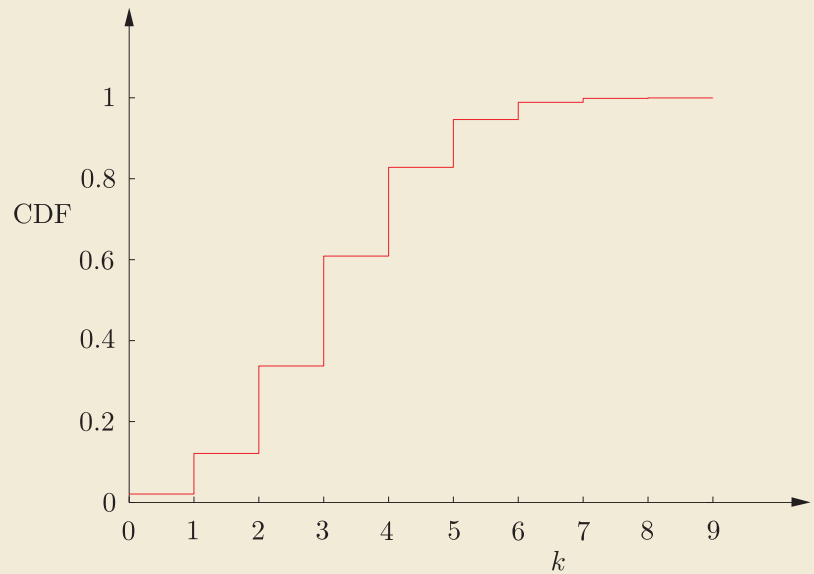
- For example, this distribution models the times that head occurs in n successive trials, where $P(\mathsf{Head}) = p$.
- The binomial is a generalization of the Bernoulli distribution, which results if we set $n = 1$.
- The mean and variance of the binomial distribution are:

$$\mathbb{E}[\mathrm{x}] = np, \ \text{and} \ \sigma_x^2 = np(1-p).$$

Typical Distributions for Discrete Variables

- The two figures below show he probability mass function and the corresponding CDF for the binomial distribution, for $p = 0.4$ and $n = 9$.

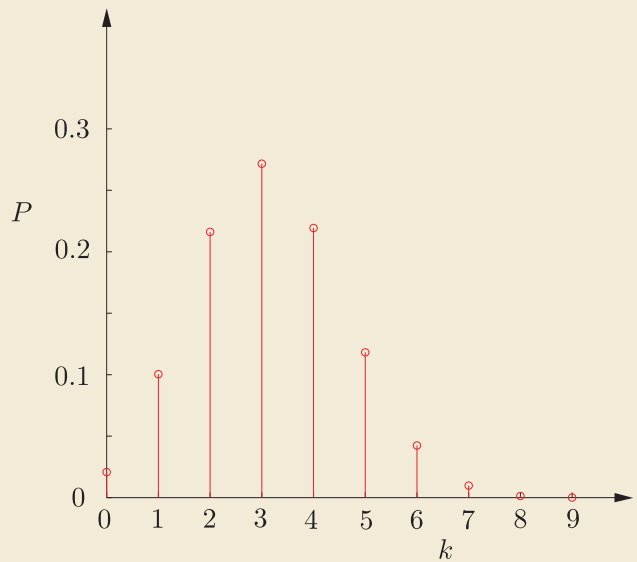


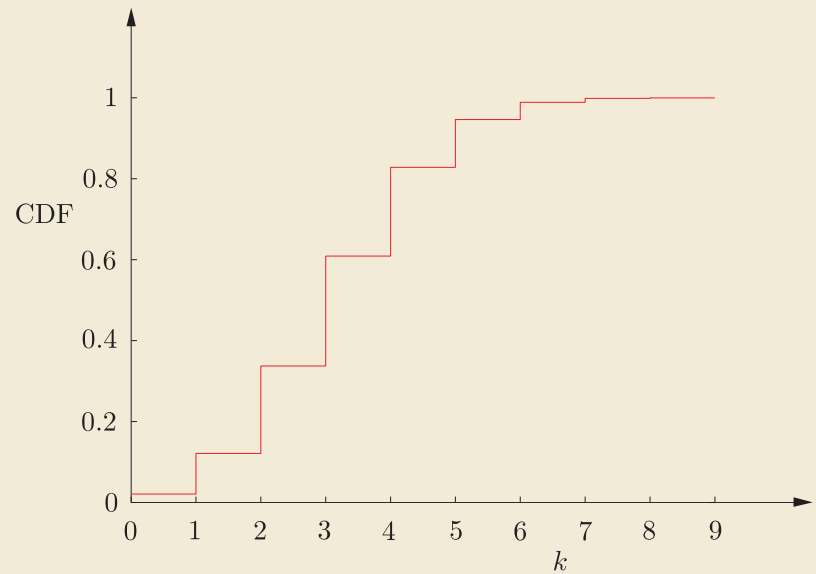


- Observe that in the case of discrete variables, the cdf function has a step-wise form.

Typical Distributions for Discrete Variables

- The two figures below show he probability mass function and the corresponding CDF for the binomial distribution, for $p = 0.4$ and $n = 9$.





- Observe that in the case of discrete variables, the cdf function has a step-wise form.

Typical Distributions for Discrete Variables

- The Multinomial Distribution: This is a generalization of the binomial distribution, if the outcome of each experiment is not binary, but it can take one out of K possible values. For example, instead of tossing a coin, a die with K sides is thrown.
- Each one of the possible K outcomes has probability $P_1, P_2, \ldots, P_K$ to occur, and we denote

$$\boldsymbol{P} = [P_1, P_2, \ldots, P_K]^T.$$

- After n experiments, assume that $x_1, x_2, \ldots, x_K$ times sides $\mathrm{x} = 1$, $\mathrm{x} = 2$,..., $\mathrm{x} = K$ occurred, respectively.

Typical Distributions for Discrete Variables

- The Multinomial Distribution: This is a generalization of the binomial distribution, if the outcome of each experiment is not binary, but it can take one out of K possible values. For example, instead of tossing a coin, a die with K sides is thrown.
- Each one of the possible K outcomes has probability $P_1, P_2, \ldots, P_K$ to occur, and we denote

$$\boldsymbol{P} = [P_1, P_2, \ldots, P_K]^T.$$

- After n experiments, assume that $x_1, x_2, \ldots, x_K$ times sides $\mathrm{x} = 1$, $\mathrm{x} = 2$,..., $\mathrm{x} = K$ occurred, respectively.

Typical Distributions for Discrete Variables

- The Multinomial Distribution: This is a generalization of the binomial distribution, if the outcome of each experiment is not binary, but it can take one out of K possible values. For example, instead of tossing a coin, a die with K sides is thrown.
- Each one of the possible K outcomes has probability $P_1, P_2, \ldots, P_K$ to occur, and we denote

$$\boldsymbol{P} = [P_1, P_2, \ldots, P_K]^T.$$

- After n experiments, assume that $x_1, x_2, \ldots, x_K$ times sides $\mathrm{x} = 1$, $\mathrm{x} = 2$,..., $\mathrm{x} = K$ occurred, respectively.

Typical Distributions for Discrete Variables

- The random (discrete) vector,

$$\mathbf{x} = [\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_K]^T,$$

follows a multinomial distribution, $\mathbf{x} \sim \text{Mult}(\boldsymbol{x}|n, \boldsymbol{P})$, if

$$P(\boldsymbol{x}) = \text{Mult}(\boldsymbol{x}|n, \boldsymbol{P}) := \binom{n}{x_1, x_2, \ldots, x_K} \prod_{i=1}^{K} P_k^{x_k},$$

where

$$\binom{n}{x_1, x_2, \ldots, x_K} = \frac{n!}{x_1! x_2! \ldots x_K!}.$$

- Note that the variables, $\mathrm{x}_1, \ldots, \mathrm{x}_K$, are subject to the constraints

$$\sum_{k=1}^{K} \mathrm{x}_k = n, \; \sum_{k=1}^{K} P_K = 1.$$

Typical Distributions for Discrete Variables

- The random (discrete) vector,

$$\mathbf{x} = [\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_K]^T,$$

follows a multinomial distribution, $\mathbf{x} \sim \text{Mult}(\boldsymbol{x}|n, \boldsymbol{P})$, if

$$P(\boldsymbol{x}) = \text{Mult}(\boldsymbol{x}|n, \boldsymbol{P}) := \left(\begin{array}{c} n \\ x_1, x_2, \ldots, x_K \end{array} \right) \prod_{i=1}^{K} P_k^{x_k},$$

where

$$\left(\begin{array}{c} n \\ x_1, x_2, \ldots, x_K \end{array} \right) = \frac{n!}{x_1! x_2! \ldots x_K!}.$$

- Note that the variables, $\mathrm{x}_1, \ldots, \mathrm{x}_K$, are subject to the constraints

$$\sum_{k=1}^{K} \mathrm{x}_k = n, \quad \sum_{k=1}^{K} P_K = 1.$$

Typical Distributions for Discrete Variables

- The random (discrete) vector,

$$\mathbf{x} = [\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_K]^T,$$

follows a multinomial distribution, $\mathbf{x} \sim \text{Mult}(\boldsymbol{x}|n, \boldsymbol{P})$, if

$$P(\boldsymbol{x}) = \text{Mult}(\boldsymbol{x}|n, \boldsymbol{P}) := \begin{pmatrix} n \\ x_1, x_2, \ldots, x_K \end{pmatrix} \prod_{i=1}^{K} P_k^{x_k},$$

where

$$\begin{pmatrix} n \\ x_1, x_2, \ldots, x_K \end{pmatrix} = \frac{n!}{x_1! x_2! \ldots x_K!}.$$

- Note that the variables, $\mathrm{x}_1, \ldots, \mathrm{x}_K$, are subject to the constraints

$$\sum_{k=1}^{K} \mathrm{x}_k = n, \quad \sum_{k=1}^{K} P_K = 1.$$

Typical Distributions for Discrete Variables

- For the multinomial distribution:
  - the mean values is given by,

$$\mathbb{E}[\mathbf{x}] = n\boldsymbol{P},$$

  - the variances by

$$\sigma_k^2 = nP_k(1 - P_k),\ k = 1, 2, \ldots, K,$$

  - and the covariances by

$$\text{cov}(\mathrm{x}_i, \mathrm{x}_j) = -nP_iP_j,\ i \neq j.$$

Typical Distributions for Continuous Variables

- The Uniform Distribution: A random variable, x, is said to follow a uniform distribution in an interval $[a, b]$ and we write $\mathrm{x} \sim \mathcal{U}(a, b)$, with $a > -\infty$ and $b < +\infty$, if

$$p(x) = \begin{cases} \frac{1}{b-a}, & \text{if } a \leq x \leq b, \\ 0 & \text{otherwise.} \end{cases}$$

- The distribution is shown in the figure below:

  - The mean value and the variance are equal to

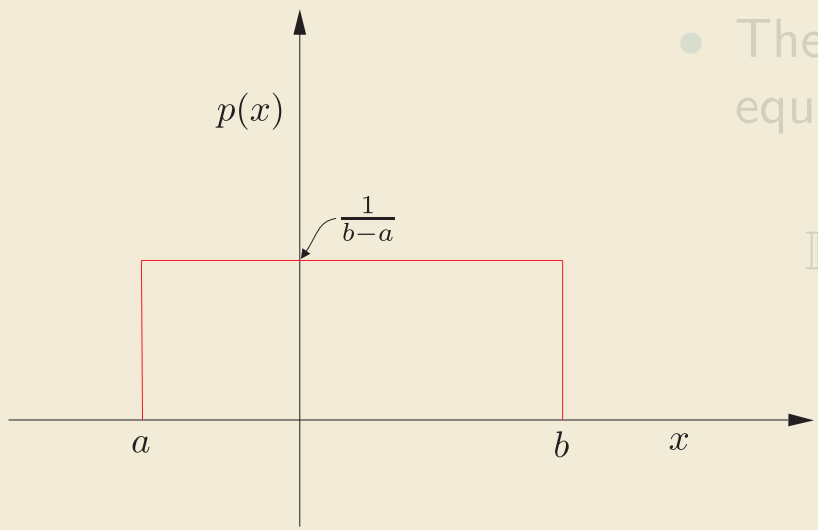
$$\mathbb{E}[\mathrm{x}] = \frac{a+b}{2}, \ \sigma_x^2 = \frac{1}{12}(b-a)^2.$$

Typical Distributions for Continuous Variables

- The Uniform Distribution: A random variable, x, is said to follow a uniform distribution in an interval $[a, b]$ and we write $\mathrm{x} \sim \mathcal{U}(a, b)$, with $a > -\infty$ and $b < +\infty$, if

$$p(x) = \begin{cases} \frac{1}{b-a}, & \text{if } a \leq x \leq b, \\ 0 & \text{otherwise.} \end{cases}$$

- The distribution is shown in the figure below:



- The mean value and the variance are equal to

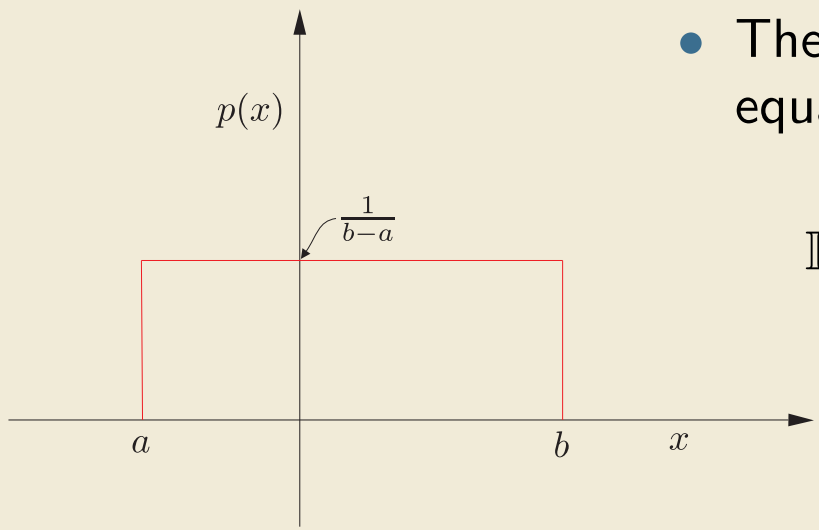
$$\mathbb{E}[\mathrm{x}] = \frac{a+b}{2}, \ \sigma_x^2 = \frac{1}{12}(b-a)^2.$$

Typical Distributions for Continuous Variables

- The Uniform Distribution: A random variable, x, is said to follow a uniform distribution in an interval $[a, b]$ and we write $\mathrm{x} \sim \mathcal{U}(a, b)$, with $a > -\infty$ and $b < +\infty$, if

$$p(x) = \begin{cases} \frac{1}{b-a}, & \text{if } a \leq x \leq b, \\ 0 & \text{otherwise.} \end{cases}$$

- The distribution is shown in the figure below:



- The mean value and the variance are equal to

$$\mathbb{E}[\mathrm{x}] = \frac{a+b}{2}, \ \sigma_x^2 = \frac{1}{12}(b-a)^2.$$

Typical Distributions for Continuous Variables

- The Gaussian Distribution: The Gaussian or normal distribution is one among the most widely used distributions in all scientific disciplines. We say that a random variable, x, is Gaussian or normal with parameters μ and σ^2, and we write $x \sim \mathcal{N}(\mu, \sigma^2)$ or $\mathcal{N}(x|\mu, \sigma^2)$, if

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(- \frac{(x-\mu)^2}{2\sigma^2} \right).$$

- The distribution is shown in the figure below:

  - The mean value and the variance are equal to:

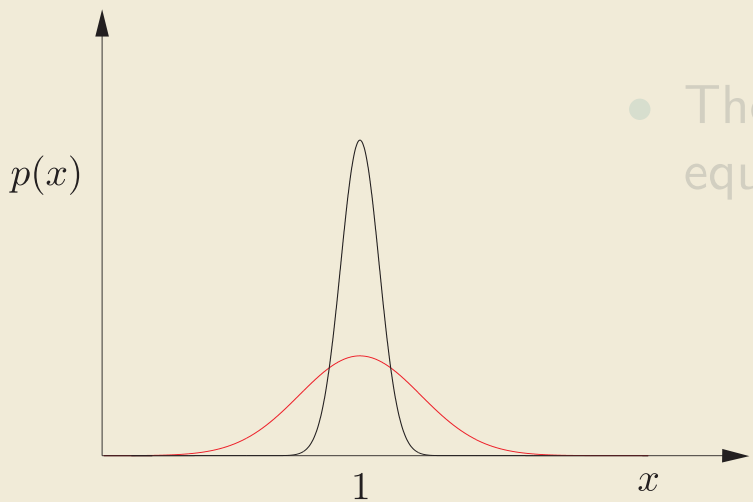
$$\mathbb{E}[\mathrm{x}] = \mu, \ \sigma_x^2 = \sigma^2.$$

Typical Distributions for Continuous Variables

- The Gaussian Distribution: The Gaussian or normal distribution is one among the most widely used distributions in all scientific disciplines. We say that a random variable, x, is Gaussian or normal with parameters μ and σ^2, and we write $x \sim \mathcal{N}(\mu, \sigma^2)$ or $\mathcal{N}(x|\mu, \sigma^2)$, if

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(- \frac{(x-\mu)^2}{2\sigma^2} \right).$$

- The distribution is shown in the figure below:



- The mean value and the variance are equal to:

$$\mathbb{E}[\mathrm{x}] = \mu, \ \sigma_x^2 = \sigma^2.$$

Typical Distributions for Continuous Variables

- The Gaussian Distribution: The Gaussian or normal distribution is one among the most widely used distributions in all scientific disciplines. We say that a random variable, x, is Gaussian or normal with parameters μ and σ^2, and we write $x \sim \mathcal{N}(\mu, \sigma^2)$ or $\mathcal{N}(x|\mu, \sigma^2)$, if

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(- \frac{(x-\mu)^2}{2\sigma^2} \right).$$

- The distribution is shown in the figure below:

- The mean value and the variance are equal to:

$$\mathbb{E}[\mathrm{x}] = \mu, \ \sigma_x^2 = \sigma^2.$$

Typical Distributions for Continuous Variables: The Gaussian

- **Proof of the mean value**: By the definition of the mean value, we have that,

$$\begin{aligned}\mathbb{E}[\mathrm{x}] &= \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{+\infty} x\exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big)dx \\ &= \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{+\infty} (y+\mu)\exp\Big(-\frac{y^2}{2\sigma^2}\Big)dy.\end{aligned}$$

Due to the symmetry of the exponential function, performing the integration involving y gives zero and the only surviving term is due to μ. Taking into account that a pdf integrates to one, we obtain the result.

Typical Distributions for Continuous Variables: The Gaussian

- **Proof of the variance**: For the variance, we have that,

$$\int_{-\infty}^{+\infty} \exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big)dx = \sqrt{2\pi}\sigma.$$

- Taking the derivative of both sides with respect to σ, we obtain

$$\int_{-\infty}^{+\infty} \frac{(x-\mu)^2}{\sigma^3} \exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big)dx = \sqrt{2\pi},$$

or

$$\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{+\infty} (x-\mu)^2 \exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big)dx = \sigma^2,$$

which proves the claim.

Typical Distributions for Continuous Variables: The Gaussian

- **Proof of the variance**: For the variance, we have that,

$$\int_{-\infty}^{+\infty} \exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big)dx = \sqrt{2\pi}\sigma.$$

- Taking the derivative of both sides with respect to σ, we obtain

$$\int_{-\infty}^{+\infty} \frac{(x-\mu)^2}{\sigma^3} \exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big)dx = \sqrt{2\pi},$$

or

$$\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{+\infty} (x-\mu)^2 \exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big)dx = \sigma^2,$$

which proves the claim.

Typical Distributions for Continuous Variables: The Gaussian

- Multivariate Gaussian: This is the generalization of the Gaussian to vector variables, $\mathbf{x} \in \mathbb{R}^l$. We write $\mathbf{x} \sim \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \Sigma)$, with parameters μ and Σ, and it is defined as

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{l/2}|\Sigma|^{1/2}} \exp\Big(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\Big),$$

where $|\cdot|$ denotes the determinant of a matrix. It can be shown that,

$$\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu} \quad \text{and Cov}(\mathbf{x}) = \Sigma.$$

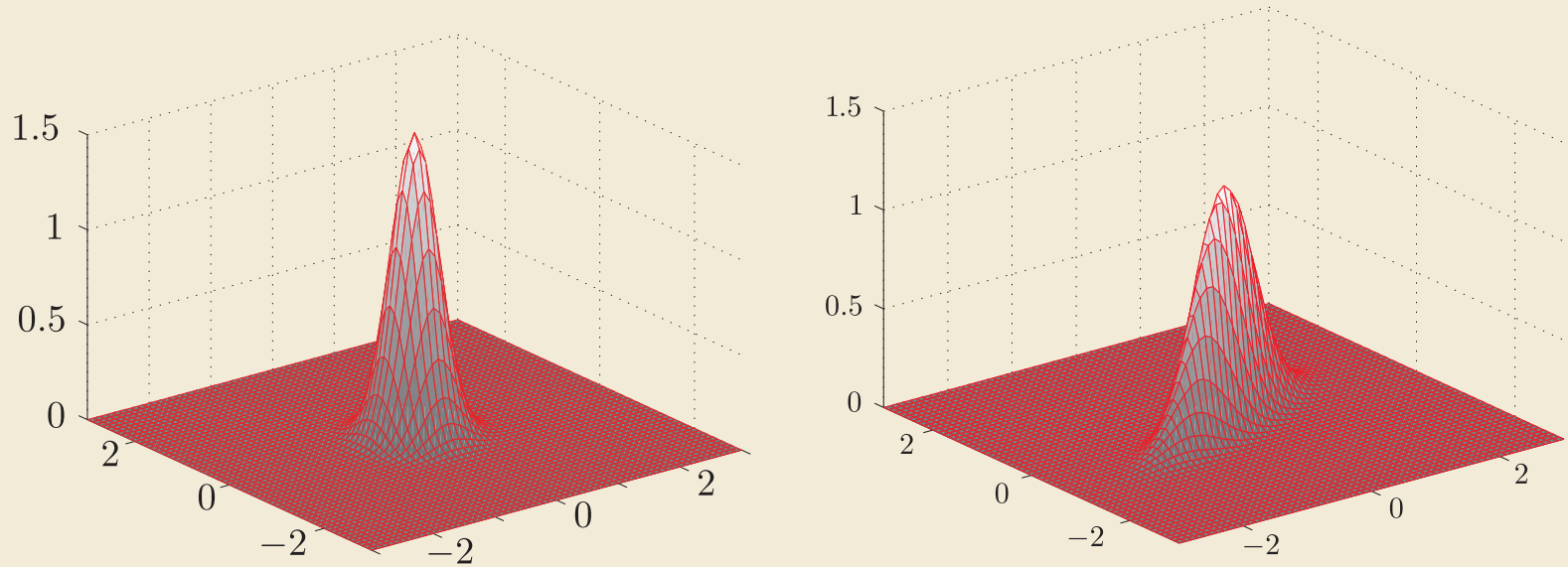
Typical Distributions for Continuous Variables: The Gaussian

- Multivariate Gaussian: This is the generalization of the Gaussian to vector variables, $\mathbf{x} \in \mathbb{R}^l$. We write $\mathbf{x} \sim \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \Sigma)$, with parameters μ and Σ, and it is defined as

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{l/2}|\Sigma|^{1/2}} \exp\Big(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\Big),$$

where $|\cdot|$ denotes the determinant of a matrix. It can be shown that,

$$\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu} \quad \text{and} \quad \text{Cov}(\mathbf{x}) = \Sigma.$$



Typical Distributions for Continuous Variables: The Gaussian

- Isovalue curves of multivariate Gaussians: The isovalue curves are formed by all the points which correspond to the same value of the pdf, i.e., $p(\boldsymbol{x}) = c$,

$$(\boldsymbol{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) = \text{constant} = c.$$

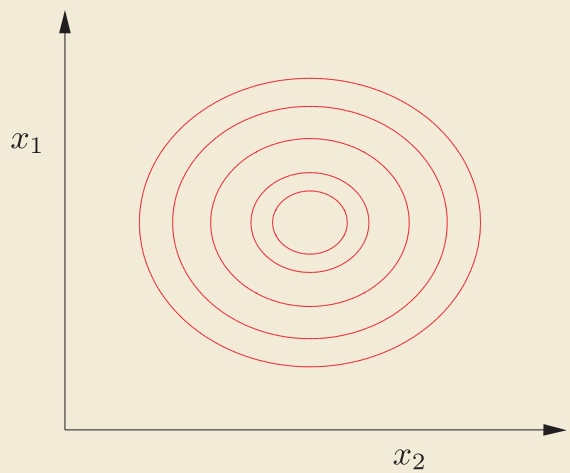
- The isovalue curves are of a quadric nature: circles (hyperspheres) or ellipses (hyperellipsoids) centered at the mean value. The minor/major axes are determined by the eigenstructure of the corresponding covariance matrix Σ.

Typical Distributions for Continuous Variables: The Gaussian

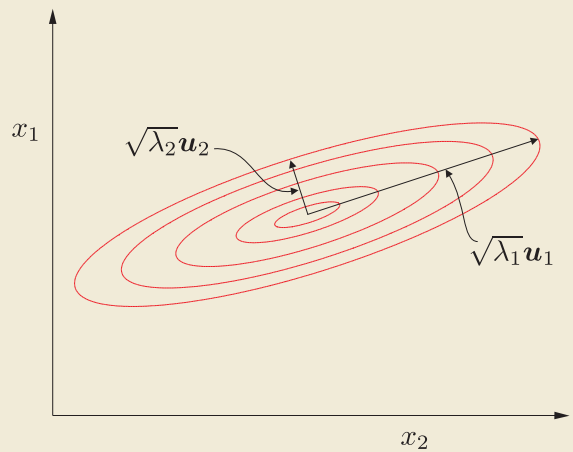
- Isovalue curves of multivariate Gaussians: The isovalue curves are formed by all the points which correspond to the same value of the pdf, i.e., $p(\boldsymbol{x}) = c$,

$$(\boldsymbol{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) = \text{constant} = c.$$

- The isovalue curves are of a quadric nature: circles (hyperspheres) or ellipses (hyperellipsoids) centered at the mean value. The minor/major axes are determined by the eigenstructure of the corresponding covariance matrix Σ.



$$\Sigma = \sigma^2 I$$



$$\Sigma \neq \sigma^2 I$$

Typical Distributions for Continuous Variables: The Gaussian

- **Proof for the shape of the contours**: All points $\boldsymbol{x} \in \mathbb{R}^l$, lying on a isovalue contour satisfy

$$(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = \text{constant} = c.$$

- The covariance matrix is symmetric, $\Sigma = \Sigma^T$. Thus, its eigenvalues are real and the corresponding eigenvectors can be chosen to form an orthonormal basis, which leads to its diagonalization, i.e.,

$$\Sigma = U^T \Lambda U, \text{ with } U := [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_l],$$

where $\boldsymbol{u}_i,\ i = 1, 2, \ldots, l$, are the corresponding orthonormal eigenvectors, and

$$\Lambda := \text{diag}\{\lambda_1, \ldots, \lambda_l\},$$

is the diagonal matrix comprising the respective eigenvalues.

Typical Distributions for Continuous Variables: The Gaussian

- **Proof for the shape of the contours**: All points $\boldsymbol{x} \in \mathbb{R}^l$, lying on a isovalue contour satisfy

$$(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = \text{constant} = c.$$

- The covariance matrix is symmetric, $\Sigma = \Sigma^T$. Thus, its eigenvalues are real and the corresponding eigenvectors can be chosen to form an orthonormal basis, which leads to its diagonalization, i.e.,

$$\Sigma = U^T \Lambda U, \text{ with } U := [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_l],$$

where $\boldsymbol{u}_i,\ i = 1, 2, \ldots, l$, are the corresponding orthonormal eigenvectors, and

$$\Lambda := \text{diag}\{\lambda_1, \ldots, \lambda_l\},$$

is the diagonal matrix comprising the respective eigenvalues.

Typical Distributions for Continuous Variables: The Gaussian

- Assuming Σ to be invertible, all eigenvalues are positive (being a positive definite matrix, it has positive eigenvalues). Due to the orthonormality of the eigenvectors, matrix U is unitary, i.e., $UU^T = U^TU = I$. Thus, we can now write

$$\boldsymbol{y}^T\Lambda^{-1}\boldsymbol{y} = c, \text{ where } \boldsymbol{y} := U(\boldsymbol{x} - \boldsymbol{\mu}), \tag{1}$$

  which corresponds to a rotation of the axes by U and a translation of the origin to $\boldsymbol{\mu}$.

- Equation (1) can be written as

$$\frac{y_1^2}{\lambda_1} + \ldots + \frac{y_l^2}{\lambda_l} = c.$$

- The last equation is describing a (hyper)ellipsoid in the $\mathbb{R}^l$. It is centered at $\boldsymbol{\mu}$ and the major axes of the ellipsoid are parallel to $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_l$. The size of the respective axes are controlled by the values of the corresponding eigenvalues.

Typical Distributions for Continuous Variables: The Gaussian

- Assuming Σ to be invertible, all eigenvalues are positive (being a positive definite matrix, it has positive eigenvalues). Due to the orthonormality of the eigenvectors, matrix U is unitary, i.e., $UU^T = U^TU = I$. Thus, we can now write

$$\boldsymbol{y}^T\Lambda^{-1}\boldsymbol{y} = c, \text{ where } \boldsymbol{y} := U(\boldsymbol{x} - \boldsymbol{\mu}), \tag{1}$$

  which corresponds to a rotation of the axes by U and a translation of the origin to $\boldsymbol{\mu}$.

- Equation (1) can be written as

$$\frac{y_1^2}{\lambda_1} + \ldots + \frac{y_l^2}{\lambda_l} = c.$$

- The last equation is describing a (hyper)ellipsoid in the $\mathbb{R}^l$. It is centered at $\boldsymbol{\mu}$ and the major axes of the ellipsoid are parallel to $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_l$. The size of the respective axes are controlled by the values of the corresponding eigenvalues.

Typical Distributions for Continuous Variables: The Gaussian

- Assuming Σ to be invertible, all eigenvalues are positive (being a positive definite matrix, it has positive eigenvalues). Due to the orthonormality of the eigenvectors, matrix U is unitary, i.e., $UU^T = U^TU = I$. Thus, we can now write

$$\boldsymbol{y}^T\Lambda^{-1}\boldsymbol{y} = c, \text{ where } \boldsymbol{y} := U(\boldsymbol{x} - \boldsymbol{\mu}), \tag{1}$$

  which corresponds to a rotation of the axes by U and a translation of the origin to $\boldsymbol{\mu}$.

- Equation (1) can be written as

$$\frac{y_1^2}{\lambda_1} + \ldots + \frac{y_l^2}{\lambda_l} = c.$$

- The last equation is describing a (hyper)ellipsoid in the $\mathbb{R}^l$. It is centered at $\boldsymbol{\mu}$ and the major axes of the ellipsoid are parallel to $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_l$. The size of the respective axes are controlled by the values of the corresponding eigenvalues.

Typical Distributions for Continuous Variables: The Gaussian

- Properties of the Gaussian distribution: If the covariance matrix is diagonal,

$$\Sigma = \text{diag}\{\sigma_1^2, \ldots, \sigma_l^2\},$$

that is, when the covariance of all the elements $\text{cov}(\mathrm{x}_i, \mathrm{x}_j) = 0,\ i, j = 1, 2, \ldots, l$, then the random variables comprising $\mathbf{x}$ are statistically independent. This is not true in general. Uncorrelated variables do not necessarily mean that they are independent. Independence is a much stronger condition.

- Indeed, if the covariance matrix is diagonal, then the multivariate Gaussian is written as,

$$p(\boldsymbol{x}) = \prod_{i=1}^{l} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\Big(-\frac{(x_i - \mu_i)^2}{2\sigma_i^2}\Big).$$

In other words,

$$p(\boldsymbol{x}) = \prod_{i=1}^{l} p(x_i),$$

which is the condition for statistical independence.

Typical Distributions for Continuous Variables: The Gaussian

- Properties of the Gaussian distribution: If the covariance matrix is diagonal,

$$\Sigma = \text{diag}\{\sigma_1^2, \ldots, \sigma_l^2\},$$

that is, when the covariance of all the elements $\text{cov}(\text{x}_i, \text{x}_j) = 0, \ i, j = 1, 2, \ldots, l$, then the random variables comprising $\mathbf{x}$ are statistically independent. This is not true in general. Uncorrelated variables do not necessarily mean that they are independent. Independence is a much stronger condition.

- Indeed, if the covariance matrix is diagonal, then the multivariate Gaussian is written as,

$$p(\boldsymbol{x}) = \prod_{i=1}^{l} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\Big(- \frac{(x_i - \mu_i)^2}{2\sigma_i^2} \Big).$$

In other words,

$$p(\boldsymbol{x}) = \prod_{i=1}^{l} p(x_i),$$

which is the condition for statistical independence.

Typical Distributions for Continuous Variables: The Gaussian

- The Central Limit Theorem: Consider N mutually independent random variables, each following its own distribution with mean values μ_i and variances $\sigma_i^2,\ i = 1, 2, \ldots, N$. Define a new random variable as their sum, i.e.,

$$\mathrm{x} = \sum_{i=1}^{N} \mathrm{x}_i.$$

Then, the mean and variance of the new variable are given by,

$$\mu = \sum_{i=1}^{N} \mu_i, \quad \text{and} \quad \sigma_x^2 = \sum_{i=1}^{N} \sigma_i^2.$$

- It can be shown that, as $N \longrightarrow \infty$ the distribution of the normalized variable

$$\mathrm{z} = \frac{\mathrm{x} - \mu}{\sigma},$$

tends to the standard normal distribution, $\mathcal{N}(z|0, 1)$

Typical Distributions for Continuous Variables: The Gaussian

- The Central Limit Theorem: Consider N mutually independent random variables, each following its own distribution with mean values μ_i and variances $\sigma_i^2,\ i = 1, 2, \ldots, N$. Define a new random variable as their sum, i.e.,

$$\mathrm{x} = \sum_{i=1}^{N} \mathrm{x}_i.$$

  Then, the mean and variance of the new variable are given by,

$$\mu = \sum_{i=1}^{N} \mu_i, \quad \text{and} \quad \sigma_x^2 = \sum_{i=1}^{N} \sigma_i^2.$$

- It can be shown that, as $N \longrightarrow \infty$ the distribution of the normalized variable

$$\mathrm{z} = \frac{\mathrm{x} - \mu}{\sigma},$$

  tends to the standard normal distribution, $\mathcal{N}(z|0, 1)$

Typical Distributions for Continuous Variables: The Gaussian

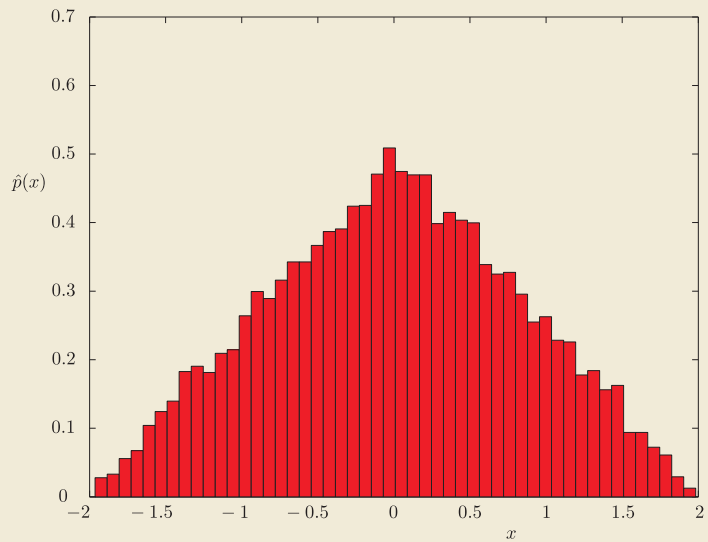
- The Central Limit Theorem is one of the most important theorems in probability and statistics and it partly explains the popularity of the Gaussian distribution.
- In practice, even summing up a relatively small number of random variables, one can obtain a good approximation to a Gaussian. For example, if the individual pdfs are smooth enough and the random variables are identically and independently distributed (iid), a number between 5 to 10 may be sufficient.

Typical Distributions for Continuous Variables: The Gaussian

- The Central Limit Theorem is one of the most important theorems in probability and statistics and it partly explains the popularity of the Gaussian distribution.
- In practice, even summing up a relatively small number of random variables, one can obtain a good approximation to a Gaussian. For example, if the individual pdfs are smooth enough and the random variables are identically and independently distributed (iid), a number between 5 to 10 may be sufficient.

Typical Distributions for Continuous Variables: The Gaussian

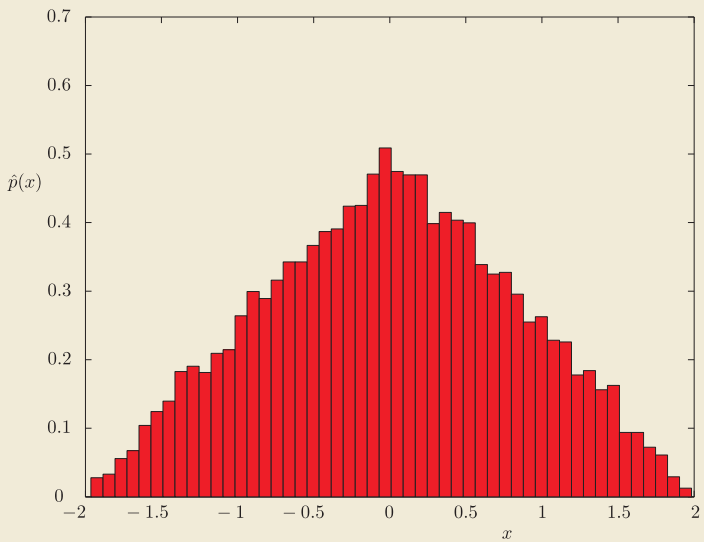
- The Central Limit Theorem is one of the most important theorems in probability and statistics and it partly explains the popularity of the Gaussian distribution.
- In practice, even summing up a relatively small number of random variables, one can obtain a good approximation to a Gaussian. For example, if the individual pdfs are smooth enough and the random variables are identically and independently distributed (iid), a number between 5 to 10 may be sufficient.



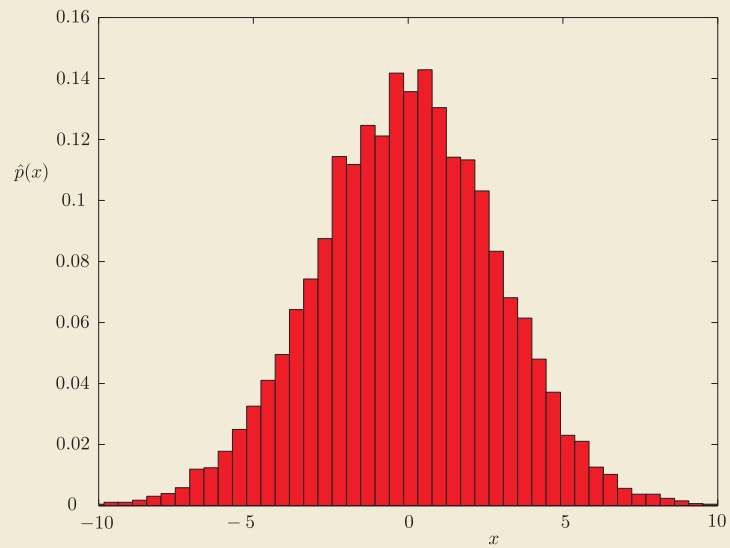
Sum of two i.i.d variables from a uniform in [-1,1]

Typical Distributions for Continuous Variables: The Gaussian

- The Central Limit Theorem is one of the most important theorems in probability and statistics and it partly explains the popularity of the Gaussian distribution.
- In practice, even summing up a relatively small number of random variables, one can obtain a good approximation to a Gaussian. For example, if the individual pdfs are smooth enough and the random variables are identically and independently distributed (iid), a number between 5 to 10 may be sufficient.



Sum of two i.i.d variables from a uniform in [-1,1]



Sum of twenty five i.i.d r.v from a uniform in [-1,1]

Typical Distributions for Continuous Variables

- The Exponential Distribution: We say that a random variable follows an exponential distribution with parameter $\lambda > 0$, if

$$p(x) = \begin{cases} \lambda \exp\left(-\lambda x\right), & \text{if } x \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

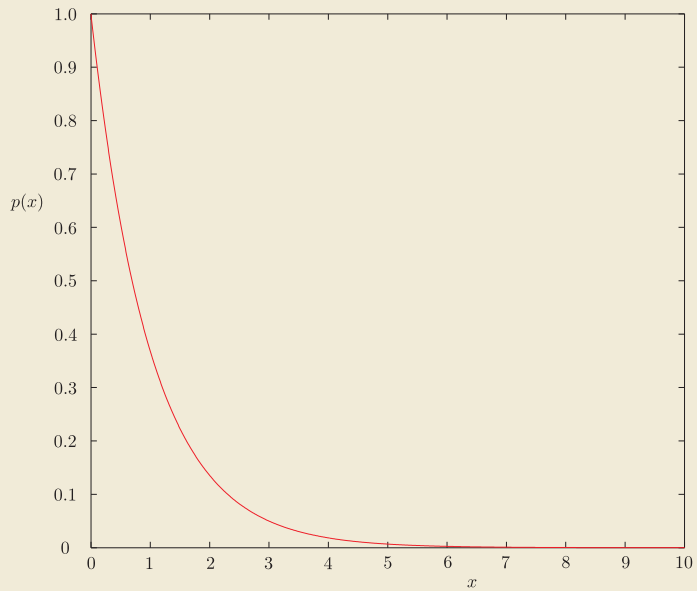
- The distribution has been used, for example, to model the time between arrivals of telephone calls or of a bus at a bus stop.

Typical Distributions for Continuous Variables

- The Exponential Distribution: We say that a random variable follows an exponential distribution with parameter $\lambda > 0$, if

$$p(x) = \begin{cases} \lambda \exp(-\lambda x), & \text{if } x \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

- The distribution has been used, for example, to model the time between arrivals of telephone calls or of a bus at a bus stop.



- The mean value and the variance are equal to:

$$\mathbb{E}[\mathrm{x}] = \frac{1}{\lambda}, \quad \sigma_x^2 = \frac{1}{\lambda^2}.$$

Typical Distributions for Continuous Variables

- The Beta Distribution: We say that a random variable, $\mathrm{x} \in [0, 1]$, follows a beta distribution with positive parameters, a, b, and we write, $\mathrm{x} \sim \text{Beta}(x|a, b,)$, if

$$p(x) = \begin{cases} \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}, & \text{if } 0 \leq x \leq 1, \\ 0 & \text{otherwise,} \end{cases}$$

where $B(a, b)$ is the beta function, defined as,

$$B(a, b) := \int_0^1 x^{a-1}(1-x)^{b-1} dx, \text{ and } B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)},$$

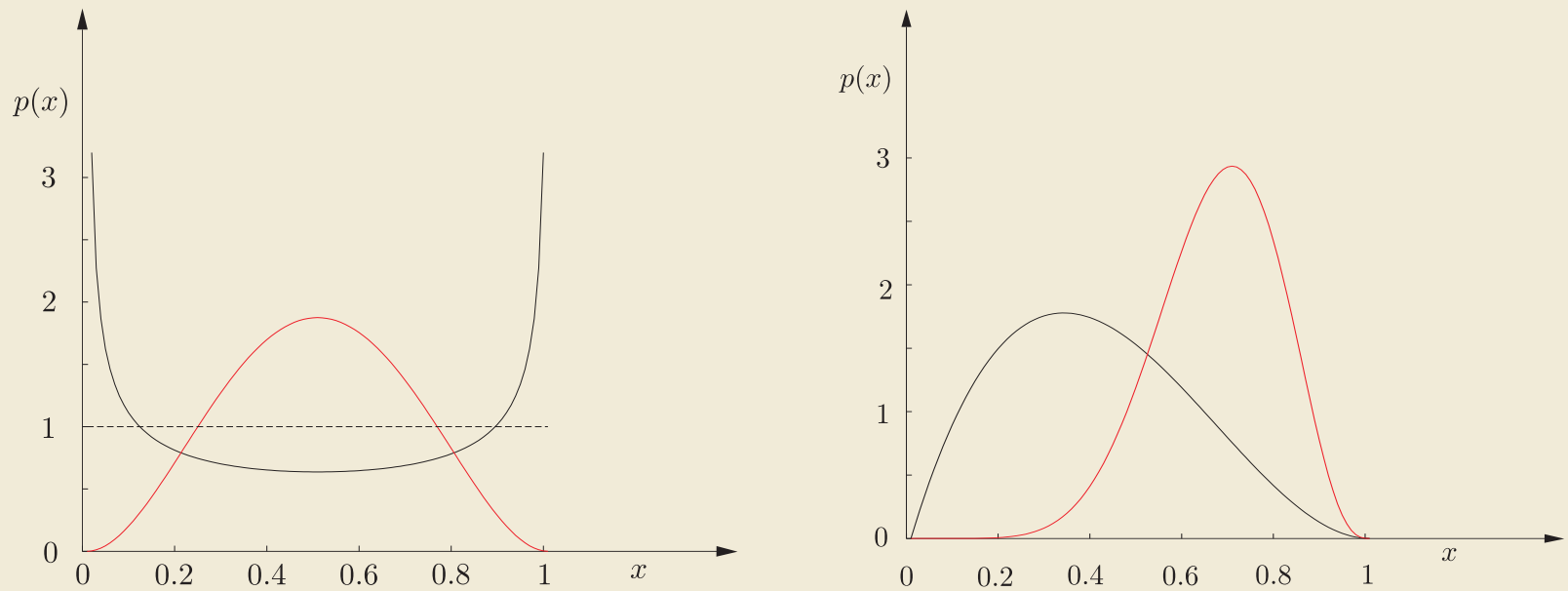
where $\Gamma(\cdot)$ is the gamma function defined as,

$$\Gamma(a) = \int_0^\infty x^{a-1} e^{-x} dx.$$

Typical Distributions for Continuous Variables-Beta

- The mean value and the variance are equal to:

$$\mathbb{E}[\mathrm{x}] = \frac{a}{a+b}, \quad \sigma_x^2 = \frac{ab}{(a+b)^2(a+b+1)}.$$

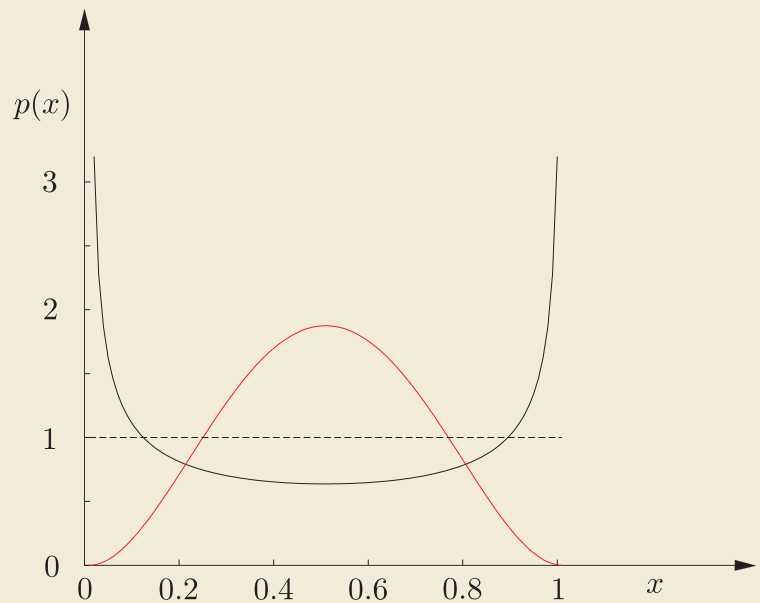


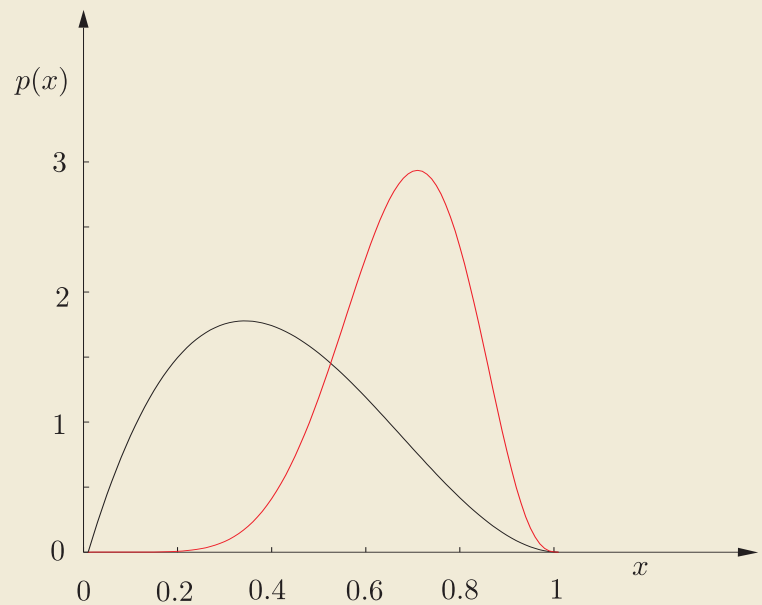
- The graphs of the pdfs of the Beta distribution for different values of the parameters. a) The dotted line corresponds to $a = 1,\ b = 1$, the gray line to $a = 0.5,\ b = 0.5$ and the red one to $a = 3,\ b = 3$. b) The gray line corresponds to $a = 2,\ b = 3$ and the red one to $a = 8,\ b = 4$. For values $a = b$, the shape is symmetric around $1/2$. For $a < 1,\ b < 1$ it is convex. For $a > 1,\ b > 1$, it is zero at $x = 0$ and $x = 1$. For $a = 1 = b$, it becomes the uniform distribution. If $a < 1,\ p(x) \longrightarrow \infty,\ x \longrightarrow 0$ and if $b < 1,\ p(x) \longrightarrow \infty,\ x \longrightarrow 1$.

Typical Distributions for Continuous Variables-Beta

- The mean value and the variance are equal to:

$$\mathbb{E}[\mathrm{x}] = \frac{a}{a+b}, \quad \sigma_x^2 = \frac{ab}{(a+b)^2(a+b+1)}.$$



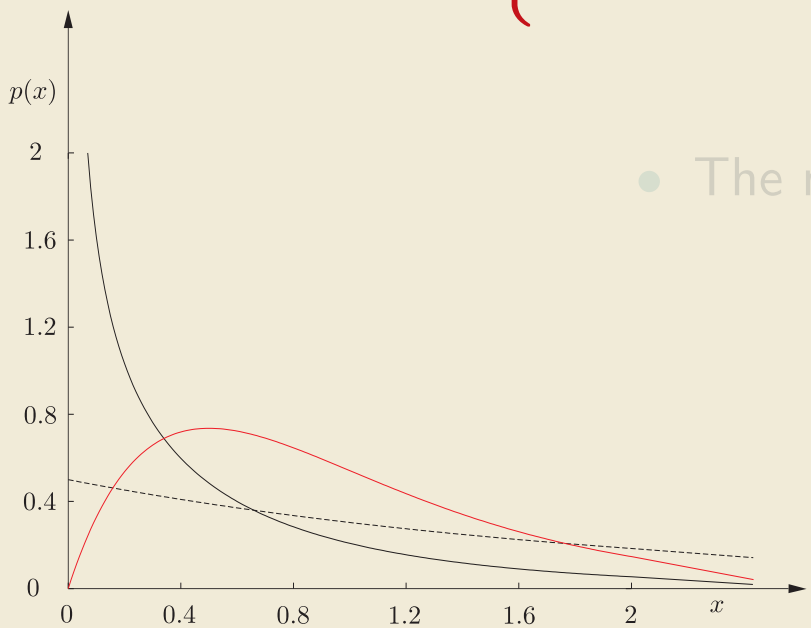


- The graphs of the pdfs of the Beta distribution for different values of the parameters. a) The dotted line corresponds to $a = 1,\ b = 1$, the gray line to $a = 0.5,\ b = 0.5$ and the red one to $a = 3,\ b = 3$. b) The gray line corresponds to $a = 2,\ b = 3$ and the red one to $a = 8,\ b = 4$. For values $a = b$, the shape is symmetric around $1/2$. For $a < 1,\ b < 1$ it is convex. For $a > 1,\ b > 1$, it is zero at $x = 0$ and $x = 1$. For $a = 1 = b$, it becomes the uniform distribution. If $a < 1,\ p(x) \longrightarrow \infty,\ x \longrightarrow 0$ and if $b < 1,\ p(x) \longrightarrow \infty,\ x \longrightarrow 1$.

Typical Distributions for Continuous Variables-Beta

- The Gamma Distribution: A random variable follows the gamma distribution with positive parameters a, b, and we write $\mathrm{x} \sim \text{Gamma}(x|a,b)$, if

$$p(x) = \begin{cases} \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}, & x > 0, \\ 0 & \text{otherwise.} \end{cases}$$



- The mean and variance are given by

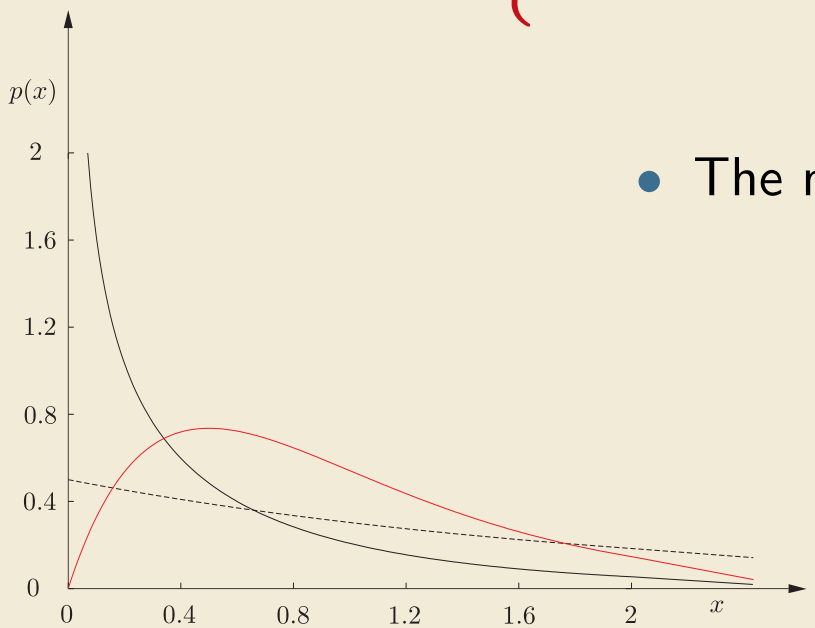
$$\mathbb{E}[\mathrm{x}] = \frac{a}{b}, \quad \sigma_x^2 = \frac{a}{b^2}.$$

- The gamma distribution also takes various shapes by varying the parameters. For $a < 1$, it is strictly decreasing and $p(x) \longrightarrow \infty$ as $x \longrightarrow 0$ and $p(x) \longrightarrow 0$ as $x \longrightarrow \infty$.

Typical Distributions for Continuous Variables-Beta

- The Gamma Distribution: A random variable follows the gamma distribution with positive parameters a, b, and we write $\mathrm{x} \sim \text{Gamma}(x|a,b)$, if

$$p(x) = \begin{cases} \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}, & x > 0, \\ 0 & \text{otherwise.} \end{cases}$$



- The mean and variance are given by

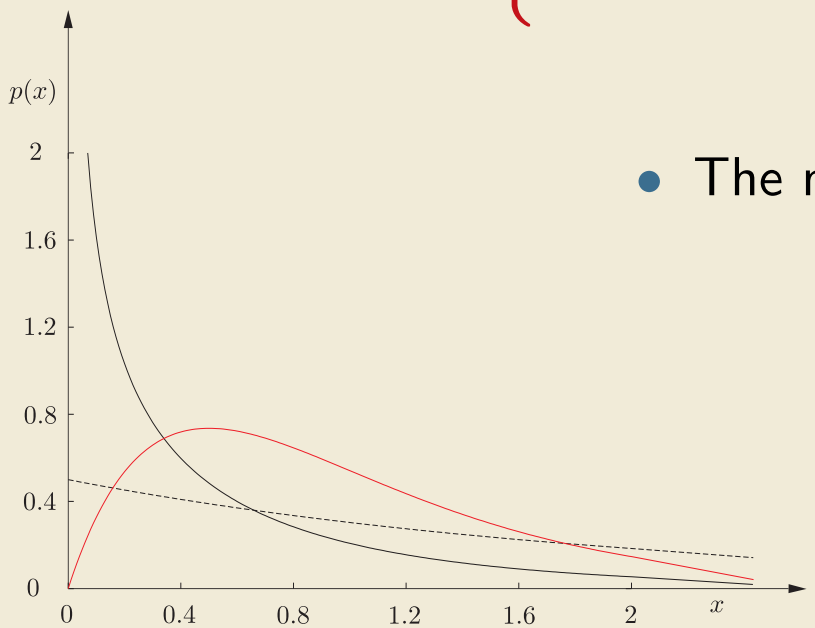
$$\mathbb{E}[\mathrm{x}] = \frac{a}{b}, \quad \sigma_x^2 = \frac{a}{b^2}.$$

- The gamma distribution also takes various shapes by varying the parameters. For $a < 1$, it is strictly decreasing and $p(x) \longrightarrow \infty$ as $x \longrightarrow 0$ and $p(x) \longrightarrow 0$ as $x \longrightarrow \infty$.

Typical Distributions for Continuous Variables-Beta

- The Gamma Distribution: A random variable follows the gamma distribution with positive parameters a, b, and we write $\mathrm{x} \sim \text{Gamma}(x|a,b)$, if

$$p(x) = \begin{cases} \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}, & x > 0, \\ 0 & \text{otherwise.} \end{cases}$$



- The mean and variance are given by

$$\mathbb{E}[\mathrm{x}] = \frac{a}{b}, \quad \sigma_x^2 = \frac{a}{b^2}.$$

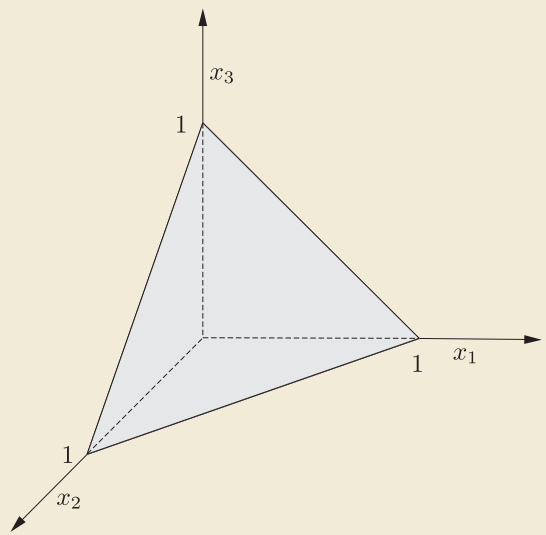
- The gamma distribution also takes various shapes by varying the parameters. For $a < 1$, it is strictly decreasing and $p(x) \longrightarrow \infty$ as $x \longrightarrow 0$ and $p(x) \longrightarrow 0$ as $x \longrightarrow \infty$.

Typical Distributions for Continuous Variables

- The Dirichlet Distribution: The Dirichlet distribution can be considered as the multivariate generalization of the beta distribution. Let $\mathbf{x} = [\mathrm{x}_1, \ldots, \mathrm{x}_K]^T$ be a random vector, with components such as

$$0 \leq \mathrm{x}_k \leq 1, \ k = 1, 2, \ldots, K, \quad \text{and} \quad \sum_{k=1}^{K} \mathrm{x}_k = 1.$$

  In other words, the random variables lie on $(K-1)$-dimensional simplex, as shown below



Typical Distributions for Continuous Variables-Dirichlet

- We say that the random vector, $\mathbf{x}$, follows a Dirichlet distribution with parameters $\boldsymbol{a} = [a_1, \ldots, a_K]^T$, and we write $\mathbf{x} \sim \text{Dir}(\boldsymbol{x}|\boldsymbol{a})$, if

$$p(\boldsymbol{x}) = \text{Dir}(\boldsymbol{x}|\boldsymbol{a}) := \frac{\Gamma(\bar{a})}{\Gamma(a_1)\ldots\Gamma(a_K)} \prod_{k=1}^{K} x_k^{a_k - 1}, \quad \bar{a} := \sum_{k=1}^{K} a_k.$$

- The mean, variance and covariances of the involved random variables are given by,

$$\mathbb{E}[\mathbf{x}] = \frac{1}{\bar{a}}\boldsymbol{a}, \quad \sigma_k^2 = \frac{a_k(\bar{a} - a_k)}{\bar{a}^2(\bar{a} + 1)}, \quad \text{cov}(\mathrm{x}_i, \mathrm{x}_j) = -\frac{a_i a_j}{\bar{a}^2(\bar{a} + 1)}.$$

Typical Distributions for Continuous Variables-Dirichlet

- We say that the random vector, $\mathbf{x}$, follows a Dirichlet distribution with parameters $\boldsymbol{a} = [a_1, \ldots, a_K]^T$, and we write $\mathbf{x} \sim \text{Dir}(\boldsymbol{x}|\boldsymbol{a})$, if

$$p(\boldsymbol{x}) = \text{Dir}(\boldsymbol{x}|\boldsymbol{a}) := \frac{\Gamma(\bar{a})}{\Gamma(a_1)\ldots\Gamma(a_K)} \prod_{k=1}^{K} x_k^{a_k - 1}, \quad \bar{a} := \sum_{k=1}^{K} a_k.$$

- The mean, variance and covariances of the involved random variables are given by,

$$\mathbb{E}[\mathbf{x}] = \frac{1}{\bar{a}}\boldsymbol{a}, \quad \sigma_k^2 = \frac{a_k(\bar{a} - a_k)}{\bar{a}^2(\bar{a} + 1)}, \quad \text{cov}(\mathrm{x}_i, \mathrm{x}_j) = -\frac{a_i a_j}{\bar{a}^2(\bar{a} + 1)}.$$

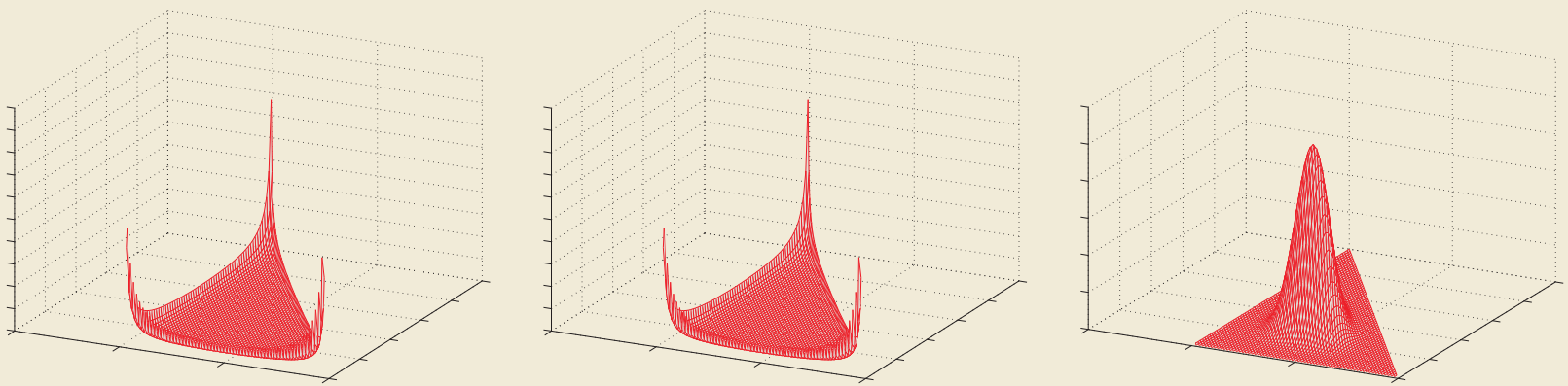
Typical Distributions for Continuous Variables-Dirichlet

- We say that the random vector, $\mathbf{x}$, follows a Dirichlet distribution with parameters $\boldsymbol{a} = [a_1, \ldots, a_K]^T$, and we write $\mathbf{x} \sim \text{Dir}(\boldsymbol{x}|\boldsymbol{a})$, if

$$p(\boldsymbol{x}) = \text{Dir}(\boldsymbol{x}|\boldsymbol{a}) := \frac{\Gamma(\bar{a})}{\Gamma(a_1)\ldots\Gamma(a_K)} \prod_{k=1}^{K} x_k^{a_k - 1}, \quad \bar{a} := \sum_{k=1}^{K} a_k.$$

- The mean, variance and covariances of the involved random variables are given by,

$$\mathbb{E}[\mathbf{x}] = \frac{1}{\bar{a}}\boldsymbol{a}, \quad \sigma_k^2 = \frac{a_k(\bar{a} - a_k)}{\bar{a}^2(\bar{a}+1)}, \quad \text{cov}(\mathrm{x}_i, \mathrm{x}_j) = -\frac{a_i a_j}{\bar{a}^2(\bar{a}+1)}.$$



- The Dirichlet distribution over the 2D-simplex for a) (0.1,0.1,0.1), b) (1,1,1) and c) (10,10,10).

Stochastic Processes

- The notion of a stochastic process is used to describe random experiments where the outcome of each experiment is a function or a sequence; in other words, the outcome of each experiment is an infinite number of values. Our focus will be on sequences. Thus, the result of a random experiment is a sequence, u_n (or sometimes denoted as $u(n)$), $n \in \mathbb{Z}$, where $\mathbb{Z}$ is the set of integers. Usually, n is interpreted as a time index, and u_n is called a time series or in the signal processing jargon a discrete-time signal. In contrast, if the outcome is a function, $u(t)$, it is called a continuous-time signal.

- We are going to use u_n to denote the specific sequence resulting from a single experiment and the roman font, u_n, to denote the corresponding discrete-time random process; that is, the rule that assigns a specific sequence as the outcome of an experiment.

Stochastic Processes

- The notion of a stochastic process is used to describe random experiments where the outcome of each experiment is a function or a sequence; in other words, the outcome of each experiment is an infinite number of values. Our focus will be on sequences. Thus, the result of a random experiment is a sequence, u_n (or sometimes denoted as $u(n)$), $n \in \mathbb{Z}$, where $\mathbb{Z}$ is the set of integers. Usually, n is interpreted as a time index, and u_n is called a time series or in the signal processing jargon a discrete-time signal. In contrast, if the outcome is a function, $u(t)$, it is called a continuous-time signal.
- We are going to use u_n to denote the specific sequence resulting from a single experiment and the roman font, u_n, to denote the corresponding discrete-time random process; that is, the rule that assigns a specific sequence as the outcome of an experiment.

Stochastic Processes

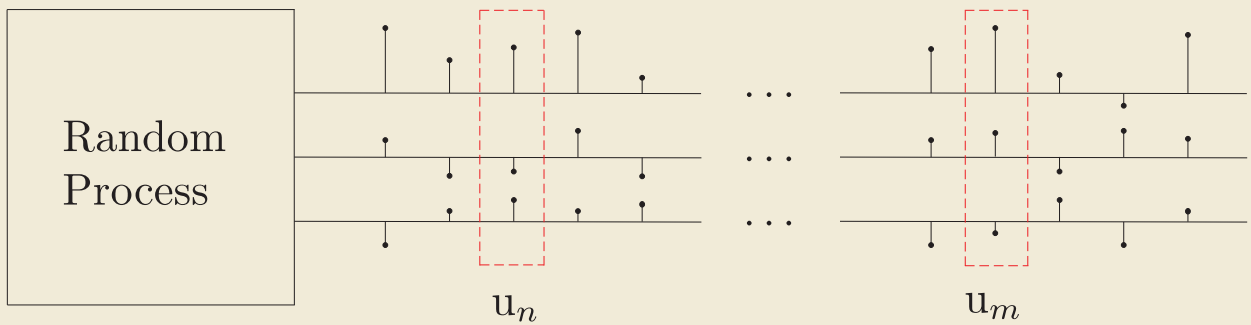
- A stochastic process can be considered as a family or ensemble of sequences. The individual sequences are known as sample sequences or simply as realizations.
- Note that fixing the time to a specific value, e.g., $n = n_0$, then u_{n_0} is a random variable. Indeed, for each random experiment, which we perform, a single value results at time instant n_0. From this perspective, a random process can be considered as the collection of infinite many random variables, i.e., $\{\mathrm{u}_n,\ n \in \mathbb{Z}\}$.

Stochastic Processes

- A stochastic process can be considered as a family or ensemble of sequences. The individual sequences are known as sample sequences or simply as realizations.
- Note that fixing the time to a specific value, e.g., $n = n_0$, then u_{n_0} is a random variable. Indeed, for each random experiment, which we perform, a single value results at time instant n_0. From this perspective, a random process can be considered as the collection of infinite many random variables, i.e., $\{\mathrm{u}_n,\ n \in \mathbb{Z}\}$.

Stochastic Processes

- A stochastic process can be considered as a family or ensemble of sequences. The individual sequences are known as sample sequences or simply as realizations.
- Note that fixing the time to a specific value, e.g., $n = n_0$, then u_{n_0} is a random variable. Indeed, for each random experiment, which we perform, a single value results at time instant n_0. From this perspective, a random process can be considered as the collection of infinite many random variables, i.e., $\{\mathrm{u}_n,\ n \in \mathbb{Z}\}$.



- The outcome of each experiment, associated with a *discrete-time* stochastic process, is a *sequence* of values. For each one of the realizations, the corresponding values obtained at any instant, e.g., n or m, comprise the outcomes of a corresponding random variable, u_n or u_m respectively.

Stochastic Processes

- First and Second Order Statistics: For a stochastic process to be fully described, one must know the joint pdfs (pmfs for discrete-valued random variables)

$$p(u_n, u_m, \ldots, u_r),$$

for all possible combinations of random variables, $\mathrm{u}_n, \mathrm{u}_m, \ldots, \mathrm{u}_r$. However, in practice, the emphasis is on computing first and second order statistics only, based on $p(u_n)$ and $p(u_n, u_m)$.

- Mean at Time n:

$$\mu_n := \mathbb{E}[\mathrm{u}_n] = \int_{-\infty}^{+\infty} u_n p(u_n) du_n.$$

- Autocovariance at Time Instants, n, m:

$$\mathrm{cov}(n, m) := \mathbb{E}\left[(\mathrm{u}_n - E[\mathrm{u}_n])(\mathrm{u}_m - E[\mathrm{u}_m])\right].$$

Stochastic Processes

- First and Second Order Statistics: For a stochastic process to be fully described, one must know the joint pdfs (pmfs for discrete-valued random variables)

$$p(u_n, u_m, \ldots, u_r),$$

for all possible combinations of random variables, $\mathrm{u}_n, \mathrm{u}_m, \ldots, \mathrm{u}_r$. However, in practice, the emphasis is on computing first and second order statistics only, based on $p(u_n)$ and $p(u_n, u_m)$.

- Mean at Time n:

$$\mu_n := \mathbb{E}[\mathrm{u}_n] = \int_{-\infty}^{+\infty} u_n p(u_n) du_n.$$

- Autocovariance at Time Instants, n, m:

$$\mathrm{cov}(n, m) := \mathbb{E}\left[(\mathrm{u}_n - E[\mathrm{u}_n])(\mathrm{u}_m - E[\mathrm{u}_m])\right].$$

Stochastic Processes

- First and Second Order Statistics: For a stochastic process to be fully described, one must know the joint pdfs (pmfs for discrete-valued random variables)

$$p(u_n, u_m, \ldots, u_r),$$

for all possible combinations of random variables, $\mathrm{u}_n, \mathrm{u}_m, \ldots, \mathrm{u}_r$. However, in practice, the emphasis is on computing first and second order statistics only, based on $p(u_n)$ and $p(u_n, u_m)$.

- Mean at Time n:

$$\mu_n := \mathbb{E}[\mathrm{u}_n] = \int_{-\infty}^{+\infty} u_n p(u_n) du_n.$$

- Autocovariance at Time Instants, n, m:

$$\mathrm{cov}(n, m) := \mathbb{E}\left[\left(\mathrm{u}_n - E[\mathrm{u}_n]\right)\left(\mathrm{u}_m - E[\mathrm{u}_m]\right)\right].$$

Stochastic Processes

- Autocorrelation at Time Instants, n, m:

$$r(n,m) := \mathbb{E}\left[\mathrm{u}_n \mathrm{u}_m\right].$$

- We refer to these mean values as ensemble averages, to stress out that they convey statistical information over the ensemble of sequences, that comprise the process.
- The respective definitions for complex stochastic processes are:

$$\mathrm{cov}(n,m) = \mathbb{E}\left[\left(\mathrm{u}_n - E[\mathrm{u}_n]\right)\left(\mathrm{u}_m - E[\mathrm{u}_m]\right)^*\right]$$

and

$$r(n,m) = \mathbb{E}\left[\mathrm{u}_n \mathrm{u}_m^*\right].$$

Stochastic Processes

- Autocorrelation at Time Instants, n, m:

$$r(n,m) := \mathbb{E}\left[\mathrm{u}_n \mathrm{u}_m\right].$$

- We refer to these mean values as ensemble averages, to stress out that they convey statistical information over the ensemble of sequences, that comprise the process.
- The respective definitions for complex stochastic processes are:

$$\mathrm{cov}(n,m) = \mathbb{E}\left[\left(\mathrm{u}_n - E[\mathrm{u}_n]\right)\left(\mathrm{u}_m - E[\mathrm{u}_m]\right)^*\right]$$

and

$$r(n,m) = \mathbb{E}\left[\mathrm{u}_n \mathrm{u}_m^*\right].$$

Stochastic Processes

- Autocorrelation at Time Instants, n, m:

$$r(n,m) := \mathbb{E}\left[\mathrm{u}_n \mathrm{u}_m\right].$$

- We refer to these mean values as ensemble averages, to stress out that they convey statistical information over the ensemble of sequences, that comprise the process.
- The respective definitions for complex stochastic processes are:

$$\mathrm{cov}(n,m) = \mathbb{E}\left[\left(\mathrm{u}_n - E[\mathrm{u}_n]\right)\left(\mathrm{u}_m - E[\mathrm{u}_m]\right)^*\right]$$

and

$$r(n,m) = \mathbb{E}\left[\mathrm{u}_n \mathrm{u}_m^*\right].$$

Stationarity and Ergodicity

- Strict Sense Stationarity: A stochastic process, u_n, is said to be strict-sense stationary (SSS) if its statistical properties are invariant to a shift of the origin, i.e., if $\forall k \in \mathbb{Z}$

$$p(u_n, u_m, \ldots, u_r) = p(u_{n-k}, u_{m-k}, \ldots, u_{r-k}),$$

  and for *any* possible combination of time instants, $n, m, \ldots, r$. In other words, the stochastic processes u_n and u_{n-k} are described by the same joint pdfs of all orders.

- A weaker version of stationarity is that of the mth order stationarity, where joint pdfs involving up to m variables, are invariant to the choice of the origin. For example, for a second order ($m = 2$) stationary process, we have that $p(u_n) = p(u_{n-k})$ and $p(u_n, u_r) = p(u_{n-k}, u_{r-k}),\ \forall n, r, k \in \mathbb{Z}$.

Stationarity and Ergodicity

- Strict Sense Stationarity: A stochastic process, u_n, is said to be strict-sense stationary (SSS) if its statistical properties are invariant to a shift of the origin, i.e., if $\forall k \in \mathbb{Z}$

$$p(u_n, u_m, \ldots, u_r) = p(u_{n-k}, u_{m-k}, \ldots, u_{r-k}),$$

  and for *any* possible combination of time instants, $n, m, \ldots, r$. In other words, the stochastic processes u_n and u_{n-k} are described by the same joint pdfs of all orders.

- A weaker version of stationarity is that of the mth order stationarity, where joint pdfs involving up to m variables, are invariant to the choice of the origin. For example, for a second order ($m = 2$) stationary process, we have that $p(u_n) = p(u_{n-k})$ and $p(u_n, u_r) = p(u_{n-k}, u_{r-k}),\ \forall n, r, k \in \mathbb{Z}$.

Stationarity and Ergodicity

- Wide Sense Stationarity: A stochastic process, u_n, is said to be wide-sense stationary (WSS) if the mean value is constant over all time instants and the autocorrelation/autocovariance sequences depend on the difference of the involved time indices, i.e.,

$$\mu_n = \mu, \quad \text{and} \quad r(n, n-k) = r(k).$$

  A WSS is a weaker version of the second order stationarity; in the latter, all possible second order statistics are independent of the origin. In the former, this is only required for the autocorrelation (autocovariance).

- A strict-sense stationary precess is also wide-sense stationary but, in general, not the other way round.
- For stationary processes, the autocorrelation becomes a sequence with a single time index as the free parameter; thus, its value, that measures a relation of the variables at two time instants, depends solely on how much these time instants differ, and not on their specific values.

Stationarity and Ergodicity

- Wide Sense Stationarity: A stochastic process, u_n, is said to be wide-sense stationary (WSS) if the mean value is constant over all time instants and the autocorrelation/autocovariance sequences depend on the difference of the involved time indices, i.e.,

$$\mu_n = \mu, \quad \text{and} \quad r(n, n-k) = r(k).$$

  A WSS is a weaker version of the second order stationarity; in the latter, all possible second order statistics are independent of the origin. In the former, this is only required for the autocorrelation (autocovariance).
- A strict-sense stationary precess is also wide-sense stationary but, in general, not the other way round.
- For stationary processes, the autocorrelation becomes a sequence with a single time index as the free parameter; thus, its value, that measures a relation of the variables at two time instants, depends solely on how much these time instants differ, and not on their specific values.

Stationarity and Ergodicity

- Wide Sense Stationarity: A stochastic process, u_n, is said to be wide-sense stationary (WSS) if the mean value is constant over all time instants and the autocorrelation/autocovariance sequences depend on the difference of the involved time indices, i.e.,

$$\mu_n = \mu, \quad \text{and} \quad r(n, n-k) = r(k).$$

  A WSS is a weaker version of the second order stationarity; in the latter, all possible second order statistics are independent of the origin. In the former, this is only required for the autocorrelation (autocovariance).
- A strict-sense stationary precess is also wide-sense stationary but, in general, not the other way round.
- For stationary processes, the autocorrelation becomes a sequence with a single time index as the free parameter; thus, its value, that measures a relation of the variables at two time instants, depends solely on how much these time instants differ, and not on their specific values.

Stationarity and Ergodicity

- Ergodicity: A stochastic process is said to be ergodic, if the complete statistics can be determined by any one of the realizations.
- In other words, if a process is ergodic, every single realization carries an identical statistical information and it can describe the entire random process. Since from a single sequence only one set of pdfs can be obtained, we conclude that every ergodic process is necessarily stationary.
- A special type of ergodicity is that of the second order ergodicity. This means that only statistics up to a second order can be obtained from a single realization. Second order ergodic processes are necessarily wide-sense stationary.
- For second order ergodic processes, the following are true:

$$\mathbb{E}[\mathrm{u}_n] = \mu = \lim_{N\to\infty} \hat{\mu}_n, \text{ where } \hat{\mu}_n := \frac{1}{2N+1} \sum_{n=-N}^{N} u_n.$$

Stationarity and Ergodicity

- Ergodicity: A stochastic process is said to be ergodic, if the complete statistics can be determined by any one of the realizations.
- In other words, if a process is ergodic, every single realization carries an identical statistical information and it can describe the entire random process. Since from a single sequence only one set of pdfs can be obtained, we conclude that every ergodic process is necessarily stationary.
- A special type of ergodicity is that of the second order ergodicity. This means that only statistics up to a second order can be obtained from a single realization. Second order ergodic processes are necessarily wide-sense stationary.
- For second order ergodic processes, the following are true:

$$\mathbb{E}[\mathrm{u}_n] = \mu = \lim_{N\to\infty} \hat{\mu}_n, \text{ where } \hat{\mu}_n := \frac{1}{2N+1} \sum_{n=-N}^{N} u_n.$$

Stationarity and Ergodicity

- Ergodicity: A stochastic process is said to be ergodic, if the complete statistics can be determined by any one of the realizations.
- In other words, if a process is ergodic, every single realization carries an identical statistical information and it can describe the entire random process. Since from a single sequence only one set of pdfs can be obtained, we conclude that every ergodic process is necessarily stationary.
- A special type of ergodicity is that of the second order ergodicity. This means that only statistics up to a second order can be obtained from a single realization. Second order ergodic processes are necessarily wide-sense stationary.
- For second order ergodic processes, the following are true:

$$\mathbb{E}[\mathrm{u}_n] = \mu = \lim_{N\to\infty} \hat{\mu}_n, \text{ where } \hat{\mu}_n := \frac{1}{2N+1} \sum_{n=-N}^{N} u_n.$$

Stationarity and Ergodicity

- Ergodicity: A stochastic process is said to be ergodic, if the complete statistics can be determined by any one of the realizations.
- In other words, if a process is ergodic, every single realization carries an identical statistical information and it can describe the entire random process. Since from a single sequence only one set of pdfs can be obtained, we conclude that every ergodic process is necessarily stationary.
- A special type of ergodicity is that of the second order ergodicity. This means that only statistics up to a second order can be obtained from a single realization. Second order ergodic processes are necessarily wide-sense stationary.
- For second order ergodic processes, the following are true:

$$\mathbb{E}[\mathrm{u}_n] = \mu = \lim_{N\to\infty} \hat{\mu}_n, \text{ where } \hat{\mu}_n := \frac{1}{2N+1} \sum_{n=-N}^{N} u_n.$$

Stationarity and Ergodicity

- Also,

$$\text{cov}(k) = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} (u_n - \mu)(u_{n-k} - \mu),$$

where the limits are in the mean square sense; that is,

$$\lim_{N\to\infty} \mathbb{E}\left[|\hat{\mu}_N - \mu|^2\right] = 0,$$

and similarly for the autocovariance.

- Note that often, ergodicity is only required to be assumed for the computation of the mean and covariance and not for all possible second order statistics. In this case, we talk about mean-ergodic and covariance-ergodic processes.

Stationarity and Ergodicity

- Also,

$$\text{cov}(k) = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} (u_n - \mu)(u_{n-k} - \mu),$$

  where the limits are in the mean square sense; that is,

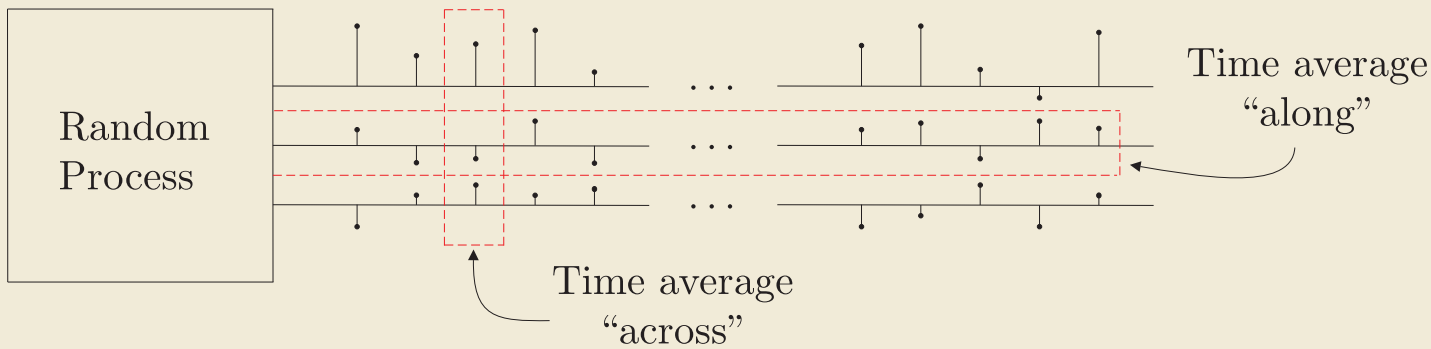
$$\lim_{N\to\infty} \mathbb{E}\left[|\hat{\mu}_N - \mu|^2\right] = 0,$$

  and similarly for the autocovariance.

- Note that often, ergodicity is only required to be assumed for the computation of the mean and covariance and not for all possible second order statistics. In this case, we talk about mean-ergodic and covariance-ergodic processes.

Stationarity and Ergodicity

- In summary, when ergodic processes are involved, ensemble averages "**across the process**" can be obtained as time averages "**along the process**".



Example

- The goal of this example is to construct a process which is WSS, yet it is not ergodic. Let a WSS process, x_n, i.e.,

$$\mathbb{E}[\mathrm{x}_n] = \mu, \text{ and } \mathbb{E}[\mathrm{x}_n \mathrm{x}_{n-k}] = r_x(k).$$

- Define the process,

$$\mathrm{z}_n := \mathrm{a}\mathrm{x}_n,$$

where a is a random variable taking values in $\{0, 1\}$, with probabilities $P(0) = P(1) = 0.5$. Moreover, a and x_n are statistically independent. Then, we have that

$$\mathbb{E}[\mathrm{z}_n] = \mathbb{E}[\mathrm{a}\mathrm{x}_n] = \mathbb{E}[\mathrm{a}]\mathbb{E}[\mathrm{x}_n] = 0.5\mu,$$

and

$$\mathbb{E}[\mathrm{z}_n \mathrm{z}_{n-k}] = \mathbb{E}[\mathrm{a}^2]\mathbb{E}[\mathrm{x}_n \mathrm{x}_{n-k}] = 0.5 r_x(k).$$

- Thus, z_n is WSS. However, it is not covariance-ergodic. Indeed, some of the realizations will be equal to zero (when $a = 0$), and the mean value and autocorrelation will be zero, which is different from the ensemble average.

Example

- The goal of this example is to construct a process which is WSS, yet it is not ergodic. Let a WSS process, x_n, i.e.,

$$\mathbb{E}[\mathrm{x}_n] = \mu, \text{ and } \mathbb{E}[\mathrm{x}_n \mathrm{x}_{n-k}] = r_x(k).$$

- Define the process,

$$\mathrm{z}_n := \mathrm{a}\mathrm{x}_n,$$

where a is a random variable taking values in $\{0, 1\}$, with probabilities $P(0) = P(1) = 0.5$. Moreover, a and x_n are statistically independent. Then, we have that

$$\mathbb{E}[\mathrm{z}_n] = \mathbb{E}[\mathrm{a}\mathrm{x}_n] = \mathbb{E}[\mathrm{a}]\mathbb{E}[\mathrm{x}_n] = 0.5\mu,$$

and

$$\mathbb{E}[\mathrm{z}_n \mathrm{z}_{n-k}] = \mathbb{E}[\mathrm{a}^2]\mathbb{E}[\mathrm{x}_n \mathrm{x}_{n-k}] = 0.5 r_x(k).$$

- Thus, z_n is WSS. However, it is not covariance-ergodic. Indeed, some of the realizations will be equal to zero (when $a = 0$), and the mean value and autocorrelation will be zero, which is different from the ensemble average.

Example

- The goal of this example is to construct a process which is WSS, yet it is not ergodic. Let a WSS process, x_n, i.e.,

$$\mathbb{E}[\mathrm{x}_n] = \mu, \text{ and } \mathbb{E}[\mathrm{x}_n \mathrm{x}_{n-k}] = r_x(k).$$

- Define the process,

$$\mathrm{z}_n := \mathrm{a}\mathrm{x}_n,$$

where a is a random variable taking values in $\{0, 1\}$, with probabilities $P(0) = P(1) = 0.5$. Moreover, a and x_n are statistically independent. Then, we have that

$$\mathbb{E}[\mathrm{z}_n] = \mathbb{E}[\mathrm{a}\mathrm{x}_n] = \mathbb{E}[\mathrm{a}]\mathbb{E}[\mathrm{x}_n] = 0.5\mu,$$

and

$$\mathbb{E}[\mathrm{z}_n \mathrm{z}_{n-k}] = \mathbb{E}[\mathrm{a}^2]\mathbb{E}[\mathrm{x}_n \mathrm{x}_{n-k}] = 0.5 r_x(k).$$

- Thus, z_n is WSS. However, it is not covariance-ergodic. Indeed, some of the realizations will be equal to zero (when $a = 0$), and the mean value and autocorrelation will be zero, which is different from the ensemble average.

Autocorrelation Sequence: Properties

- Let u_n be a wide-sense stationary process. Its autocorrelation sequence has the following properties:

1. $$r(k) = r^*(-k), \ \forall k \in \mathbb{Z}.$$

   Proof: This property is a direct consequence of the invariance with respect to the choice of the origin. Indeed,

   $$r(k) = \mathbb{E}[\mathrm{u}_n \mathrm{u}_{n-k}^*] = \mathbb{E}[\mathrm{u}_{n+k} \mathrm{u}_n^*] = r^*(-k).$$

2. $$r(0) = \mathbb{E}\left[|\mathrm{u}_n|^2\right].$$

   That is, the value of the autocorrelation at $k = 0$ is equal to the mean square value of the process. Interpreting the square of a variable as its energy, then $r(0)$ can be interpreted as the corresponding (average) power.

Autocorrelation Sequence: Properties

- Let u_n be a wide-sense stationary process. Its autocorrelation sequence has the following properties:

1. $$r(k) = r^*(-k), \ \forall k \in \mathbb{Z}.$$

   Proof: This property is a direct consequence of the invariance with respect to the choice of the origin. Indeed,

   $$r(k) = \mathbb{E}[\mathrm{u}_n \mathrm{u}_{n-k}^*] = \mathbb{E}[\mathrm{u}_{n+k} \mathrm{u}_n^*] = r^*(-k).$$

2. $$r(0) = \mathbb{E}\left[|\mathrm{u}_n|^2\right].$$

   That is, the value of the autocorrelation at $k = 0$ is equal to the mean square value of the process. Interpreting the square of a variable as its energy, then $r(0)$ can be interpreted as the corresponding (average) power.

Autocorrelation Sequence: Properties

- Let u_n be a wide-sense stationary process. Its autocorrelation sequence has the following properties:

1. $$r(k) = r^*(-k), \ \forall k \in \mathbb{Z}.$$

   Proof: This property is a direct consequence of the invariance with respect to the choice of the origin. Indeed,

   $$r(k) = \mathbb{E}[\mathrm{u}_n \mathrm{u}_{n-k}^*] = \mathbb{E}[\mathrm{u}_{n+k} \mathrm{u}_n^*] = r^*(-k).$$

2. $$r(0) = \mathbb{E}\left[|\mathrm{u}_n|^2\right].$$

   That is, the value of the autocorrelation at $k = 0$ is equal to the mean square value of the process. Interpreting the square of a variable as its energy, then $r(0)$ can be interpreted as the corresponding (average) power.

Autocorrelation Sequence: Properties

- (Properties continued)

3. $$r(0) \geq |r(k)|, \ \forall k \neq 0.$$
   In other words, the correlation of the variables, corresponding to two different time instants, cannot be larger (in magnitude) than $r(0)$. This property is essentially the Cauchy-Schwartz inequality for the inner products.
4. The autocorrelation of a stochastic process is a positive definite sequence. That is,
   $$\sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m^* r(n,m) \geq 0, \ \forall a_n \in \mathbb{C}, \ n = 1, 2, \ldots, N, \ \forall N \in \mathbb{Z}.$$
   Proof: The proof is easily obtained by the definition of the autocorrelation,
   $$0 \leq \mathbb{E}\Big[\Big|\sum_{n=1}^{N} a_n \mathrm{x}_n\Big|^2\Big] = \sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m^* \mathbb{E}\left[\mathrm{x}_n \mathrm{x}_m\right],$$
   which proves the claim.

Autocorrelation Sequence: Properties

- (Properties continued)

3. $$r(0) \geq |r(k)|, \ \forall k \neq 0.$$
   In other words, the correlation of the variables, corresponding to two different time instants, cannot be larger (in magnitude) than $r(0)$. This property is essentially the Cauchy-Schwartz inequality for the inner products.
4. The autocorrelation of a stochastic process is a positive definite sequence. That is,
   $$\sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m^* r(n,m) \geq 0, \ \forall a_n \in \mathbb{C}, \ n = 1, 2, \ldots, N, \ \forall N \in \mathbb{Z}.$$
   Proof: The proof is easily obtained by the definition of the autocorrelation,
   $$0 \leq \mathbb{E}\Big[\Big|\sum_{n=1}^{N} a_n \mathrm{x}_n\Big|^2\Big] = \sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m^* \mathbb{E}\left[\mathrm{x}_n \mathrm{x}_m\right],$$
   which proves the claim.

Autocorrelation Sequence: Properties

- (Properties continued)

3. $$r(0) \geq |r(k)|, \ \forall k \neq 0.$$
   In other words, the correlation of the variables, corresponding to two different time instants, cannot be larger (in magnitude) than $r(0)$. This property is essentially the Cauchy-Schwartz inequality for the inner products.
4. The autocorrelation of a stochastic process is a positive definite sequence. That is,
   $$\sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m^* r(n,m) \geq 0, \ \forall a_n \in \mathbb{C}, \ n = 1, 2, \ldots, N, \ \forall N \in \mathbb{Z}.$$
   Proof: The proof is easily obtained by the definition of the autocorrelation,
   $$0 \leq \mathbb{E}\left[\left|\sum_{n=1}^{N} a_n \mathrm{x}_n\right|^2\right] = \sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m^* \mathbb{E}\left[\mathrm{x}_n \mathrm{x}_m\right],$$
   which proves the claim.

Autocorrelation Sequence: Properties

- (Properties continued)

5. Let u_n and v_n be two WSS processes. Define the new process

$$\mathrm{z}_n = \mathrm{u}_n + \mathrm{v}_n.$$

Then,

$$r_z(k) = r_u(k) + r_v(k) + r_{uv}(k) + r_{vu}(k),$$

where the cross-correlation between two jointly WSS stationary stochastic processes is defined as

$$r_{uv}(k) := \mathbb{E}[\mathrm{u}_n \mathrm{v}^*_{n-k}], \ k \in \mathbb{Z}.$$

The proof is a direct consequence of the definition. Note that if the two processes are *uncorrelated*, i.e., $r_{uv}(k) = r_{vu}(k) = 0$, then

$$r_z(k) = r_u(k) + r_v(k).$$

Obviously, this is also true if the processes u_n and v_n are independent and of zero mean value, since then $\mathbb{E}[\mathrm{u}_n \mathrm{v}_{n-k}] = \mathbb{E}[\mathrm{u}_n]\mathbb{E}[\mathrm{v}_{n-k}] = 0$. Note that, uncorelateness is a weaker condition and it does not necessarily imply independence; the opposite is true, for zero mean values.

- (Properties continued)

5. Let u_n and v_n be two WSS processes. Define the new process

$$\mathrm{z}_n = \mathrm{u}_n + \mathrm{v}_n.$$

Then,

$$r_z(k) = r_u(k) + r_v(k) + r_{uv}(k) + r_{vu}(k),$$

where the cross-correlation between two jointly WSS stationary stochastic processes is defined as

$$r_{uv}(k) := \mathbb{E}[\mathrm{u}_n \mathrm{v}^*_{n-k}], \ k \in \mathbb{Z}.$$

The proof is a direct consequence of the definition. Note that if the two processes are *uncorrelated*, i.e., $r_{uv}(k) = r_{vu}(k) = 0$, then

$$r_z(k) = r_u(k) + r_v(k).$$

Obviously, this is also true if the processes u_n and v_n are independent and of zero mean value, since then $\mathbb{E}[\mathrm{u}_n \mathrm{v}_{n-k}] = \mathbb{E}[\mathrm{u}_n]\mathbb{E}[\mathrm{v}_{n-k}] = 0$. Note that, uncorelateness is a weaker condition and it does not necessarily imply independence; the opposite is true, for zero mean values.

Autocorrelation Sequence: Properties

- (Properties continued)

6. $$r_{uv}(k) = r_{vu}^*(-k)$$

7. $$r_u(0)r_v(0) \geq |r_{uv}(k)|, \ \forall k \in \mathbb{Z}.$$

Power Spectral Density

- Power Spectral Density: Given a WSS stochastic process, u_n, its power spectral density (PSD) (or simply the power spectrum) is defined as the Fourier transform of its autocorrelation sequence, i.e.,

$$S(\omega) := \sum_{k=-\infty}^{\infty} r(k) \exp\left(-j\omega k\right).$$

The autocorrelation sequence is obtained via the inverse Fourier transform, i.e.,

$$r(k) = \frac{1}{2\pi} \int_{-\pi}^{+\pi} S(\omega) \exp\left(j\omega k\right) d\omega. \tag{2}$$

Autocorrelation Sequence: Properties

- (Properties continued)

6. $$r_{uv}(k) = r_{vu}^*(-k)$$

7. $$r_u(0) r_v(0) \geq |r_{uv}(k)|, \ \forall k \in \mathbb{Z}.$$

Power Spectral Density

- Power Spectral Density: Given a WSS stochastic process, u_n, its power spectral density (PSD) (or simply the power spectrum) is defined as the Fourier transform of its autocorrelation sequence, i.e.,

$$S(\omega) := \sum_{k=-\infty}^{\infty} r(k) \exp\left(-j\omega k\right).$$

The autocorrelation sequence is obtained via the inverse Fourier transform, i.e.,

$$r(k) = \frac{1}{2\pi} \int_{-\pi}^{+\pi} S(\omega) \exp\left(j\omega k\right) d\omega. \tag{2}$$

Autocorrelation Sequence: Properties

- (Properties continued)
  6. $$r_{uv}(k) = r^*_{vu}(-k)$$
  7. $$r_u(0) r_v(0) \geq |r_{uv}(k)|, \ \forall k \in \mathbb{Z}.$$

Power Spectral Density

- Power Spectral Density: Given a WSS stochastic process, u_n, its power spectral density (PSD) (or simply the power spectrum) is defined as the Fourier transform of its autocorrelation sequence, i.e.,
$$S(\omega) := \sum_{k=-\infty}^{\infty} r(k) \exp\left(-j\omega k\right).$$
The autocorrelation sequence is obtained via the inverse Fourier transform, i.e.,
$$r(k) = \frac{1}{2\pi} \int_{-\pi}^{+\pi} S(\omega) \exp\left(j\omega k\right) d\omega. \tag{2}$$

Properties of PSD

- The PSD of a WSS stochastic process is a real and non-negative function of ω.

  Proof: Indeed, we have that,

$$\begin{aligned} S(\omega) &= \sum_{k=-\infty}^{+\infty} r(k)\exp\left(-j\omega k\right) \\ &= r(0) + \sum_{k=-\infty}^{-1} r(k)\exp\left(-j\omega k\right) + \sum_{k=1}^{\infty} r(k)\exp\left(-j\omega k\right) \\ &= r(0) + \sum_{k=1}^{+\infty} r^*(k)\exp\left(j\omega k\right) + \sum_{k=1}^{\infty} r(k)\exp\left(-j\omega k\right) \\ &= r(0) + 2\sum_{k=1}^{+\infty} \text{Real}\big(r(k)\exp\left(-j\omega k\right)\big), \end{aligned}$$

  which proves the claim that PSD is a real number. In the proof, Property 1 of the autocorrelation sequence has been used. We defer the proof for the non-negative part for later on.

Properties of PSD

- The PSD of a WSS stochastic process is a real and non-negative function of ω.

  Proof: Indeed, we have that,

$$\begin{aligned} S(\omega) &= \sum_{k=-\infty}^{+\infty} r(k)\exp\left(-j\omega k\right) \\ &= r(0) + \sum_{k=-\infty}^{-1} r(k)\exp\left(-j\omega k\right) + \sum_{k=1}^{\infty} r(k)\exp\left(-j\omega k\right) \\ &= r(0) + \sum_{k=1}^{+\infty} r^*(k)\exp\left(j\omega k\right) + \sum_{k=1}^{\infty} r(k)\exp\left(-j\omega k\right) \\ &= r(0) + 2\sum_{k=1}^{+\infty} \text{Real}\big(r(k)\exp\left(-j\omega k\right)\big), \end{aligned}$$

  which proves the claim that PSD is a real number. In the proof, Property 1 of the autocorrelation sequence has been used. We defer the proof for the non-negative part for later on.

Properties of PSD

- The area under the graph of $S(\omega)$ is equal to the power of the stochastic process, i.e.,

$$\mathbb{E}[|\mathrm{u}_n|^2] = r(0) = \frac{1}{2\pi}\int_{-\pi}^{+\pi} S(\omega)d\omega,$$

  which is obtained from (2) if we set $k = 0$. We will come to the physical meaning of this property very soon.

- Transmission through a linear system: We will now derive the relation between the PSDs of the input and output in a linear filtering operation, expressed via the convolution sum,

$$\mathrm{d}_n = w_n * \mathrm{u}_n := \sum_{k=-\infty}^{+\infty} w_k^* \mathrm{u}_{n-k}$$

  where $\ldots, w_0, w_1, w_2, \ldots$ are the parameters comprising the impulse response describing the filter.

Properties of PSD

- The area under the graph of $S(\omega)$ is equal to the power of the stochastic process, i.e.,

$$\mathbb{E}[|\mathrm{u}_n|^2] = r(0) = \frac{1}{2\pi}\int_{-\pi}^{+\pi} S(\omega)d\omega,$$

  which is obtained from (2) if we set $k = 0$. We will come to the physical meaning of this property very soon.

- Transmission through a linear system: We will now derive the relation between the PSDs of the input and output in a linear filtering operation, expressed via the convolution sum,

$$\mathrm{d}_n = w_n * \mathrm{u}_n := \sum_{k=-\infty}^{+\infty} w_k^* \mathrm{u}_{n-k}$$

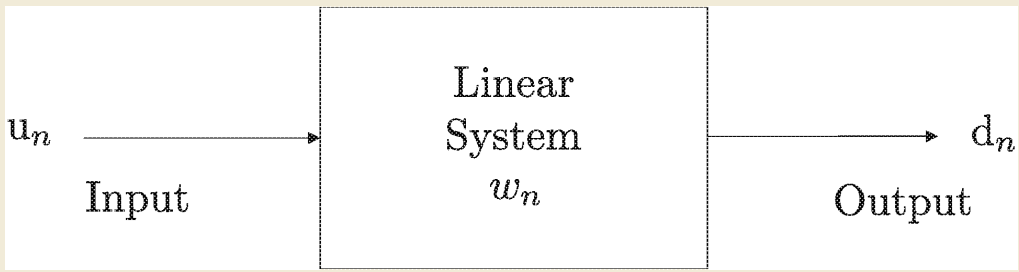
  where $\ldots, w_0, w_1, w_2, \ldots$ are the parameters comprising the impulse response describing the filter.

Properties of PSD

- In case the impulse response is of finite duration, for example, $w_0, w_1, \ldots, w_{l-1}$, then the convolution can be written as

$$\mathrm{d}_n = \sum_{k=0}^{l-1} w_k^* \mathrm{u}_{n-k} = \boldsymbol{w}^H \mathbf{u}_n,$$

$$\boldsymbol{w} := [w_0, w_1, \ldots, w_{l-1}]^T, \ \mathbf{u}_n := [\mathrm{u}_n, \mathrm{u}_{n-1}, \ldots, \mathrm{u}_{n-l+1}]^T \in \mathbb{R}^l.$$



Properties of PSD

- The random vector at the input

$$\mathbf{u}_n := [\mathrm{u}_n, \mathrm{u}_{n-1}, \ldots, \mathrm{u}_{n-l+1}]^T \in \mathbb{R}^l.$$

is known as the input vector of order l and at time n. Note that its elements are part of the stochastic process at successive time instants. This imposes on the respective autocorrelation matrix a rich structure, which can be exploited to develop efficient computational algorithms for its inversion.

Moreover, observe that, if the impulse response of the system is zero for negative values of the time index, n, this guarantees causality. That is, the output depends only on the values of the input at the current and previous time instants only, and there is no dependence on future values.

Properties of PSD

- The random vector at the input

$$\mathbf{u}_n := [\mathrm{u}_n, \mathrm{u}_{n-1}, \ldots, \mathrm{u}_{n-l+1}]^T \in \mathbb{R}^l.$$

is known as the input vector of order l and at time n. Note that its elements are part of the stochastic process at successive time instants. This imposes on the respective autocorrelation matrix a rich structure, which can be exploited to develop efficient computational algorithms for its inversion.

Moreover, observe that, if the impulse response of the system is zero for negative values of the time index, n, this guarantees causality. That is, the output depends only on the values of the input at the current and previous time instants only, and there is no dependence on future values.

Properties of PSD

- Theorem: The power spectral density of the output, d_n, of a linear time invariant system, when it is excited by a WSS stochastic process, u_n, is given by,

$$S_d(\omega) = \left|W(\omega)\right|^2 S_u(\omega),$$

where

$$W(\omega) := \sum_{n=-\infty}^{+\infty} w_n \exp\left(-j\omega n\right).$$

- **Proof**: First, it is shown that,

$$r_d(k) = r_u(k) * w_k * w^*_{-k}.$$

Then the claim is proved by taking the Fourier transform of both sides. Two well known properties of the Fourier transform have been used, i.e.,

$$r_u(k) * w_k \longmapsto S_u(\omega)W(\omega), \quad \text{and} \quad w^*_{-k} \longmapsto W^*(\omega).$$

Properties of PSD

- Theorem: The power spectral density of the output, d_n, of a linear time invariant system, when it is excited by a WSS stochastic process, u_n, is given by,

$$S_d(\omega) = \left|W(\omega)\right|^2 S_u(\omega),$$

where

$$W(\omega) := \sum_{n=-\infty}^{+\infty} w_n \exp\left(-j\omega n\right).$$

- **Proof**: First, it is shown that,

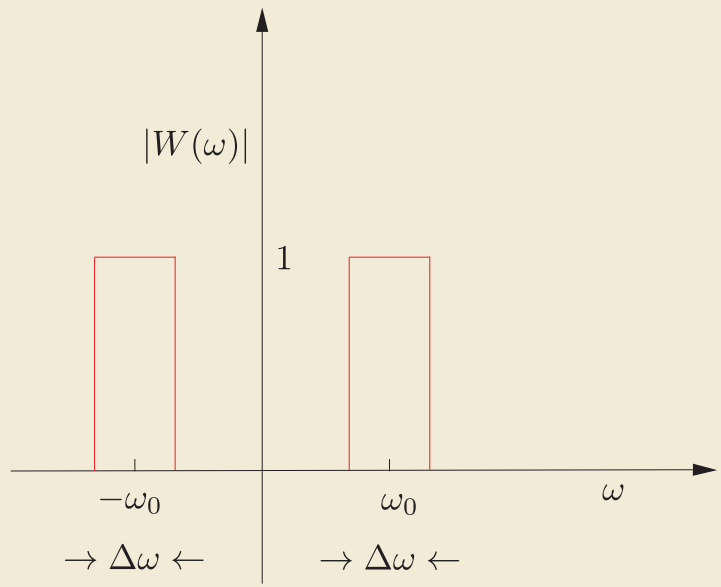
$$r_d(k) = r_u(k) * w_k * w^*_{-k}.$$

Then the claim is proved by taking the Fourier transform of both sides. Two well known properties of the Fourier transform have been used, i.e.,

$$r_u(k) * w_k \longmapsto S_u(\omega)W(\omega), \quad \text{and} \quad w^*_{-k} \longmapsto W^*(\omega).$$

Power Spectral Density

- Physical Interpretation of the PSD: The following figure shows the Fourier transform of the impulse response of a very narrow bandpass filter.



An ideal bandpass filter. The output contains frequencies only in the range of $|\omega - \omega_0| < \Delta\omega/2$.

Power Spectral Density

- We assume that $\Delta\omega$ is very small. Then, for this special case, the input-output PSD relation can be written as

$$\Delta P := \mathbb{E}\left[|\mathrm{d}_n|^2\right] = r_d(0) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} S_d(\omega)d\omega \approx S_u(\omega_o)\frac{\Delta\omega}{\pi}.$$

where real data have been assumed, which guarantees the symmetry of the (magnitude) of the Fourier transform $(S_u(\omega) = S_u(-\omega))$.

Hence,

$$\frac{1}{\pi}S_u(\omega_o) = \frac{\Delta P}{\Delta\omega}.$$

In other words, the value $S_u(\omega_o)$ can be interpreted as the power density (power per frequency interval) in the frequency (spectrum) domain.

- This also establishes what was said before: the PSD is a non-negative real function, for any value of $\omega \in [-\pi, +\pi]$ (The PSD, being the Fourier transform of a sequence, is periodic with period 2π).

Power Spectral Density

- We assume that $\Delta\omega$ is very small. Then, for this special case, the input-output PSD relation can be written as

$$\Delta P := \mathbb{E}\left[|\mathrm{d}_n|^2\right] = r_d(0) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} S_d(\omega)d\omega \approx S_u(\omega_o)\frac{\Delta\omega}{\pi}.$$

where real data have been assumed, which guarantees the symmetry of the (magnitude) of the Fourier transform $(S_u(\omega) = S_u(-\omega))$.
Hence,

$$\frac{1}{\pi}S_u(\omega_o) = \frac{\Delta P}{\Delta\omega}.$$

In other words, the value $S_u(\omega_o)$ can be interpreted as the power density (power per frequency interval) in the frequency (spectrum) domain.

- This also establishes what was said before: the PSD is a non-negative real function, for any value of $\omega \in [-\pi, +\pi]$ (The PSD, being the Fourier transform of a sequence, is periodic with period 2π).

Power Spectral Density

- We assume that $\Delta\omega$ is very small. Then, for this special case, the input-output PSD relation can be written as

$$\Delta P := \mathbb{E}\left[|\mathrm{d}_n|^2\right] = r_d(0) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} S_d(\omega)d\omega \approx S_u(\omega_o)\frac{\Delta\omega}{\pi}.$$

where real data have been assumed, which guarantees the symmetry of the (magnitude) of the Fourier transform $(S_u(\omega) = S_u(-\omega))$.
Hence,

$$\frac{1}{\pi}S_u(\omega_o) = \frac{\Delta P}{\Delta\omega}.$$

In other words, the value $S_u(\omega_o)$ can be interpreted as the power density (power per frequency interval) in the frequency (spectrum) domain.

- This also establishes what was said before: the PSD is a non-negative real function, for any value of $\omega \in [-\pi, +\pi]$ (The PSD, being the Fourier transform of a sequence, is periodic with period 2π).

Example: White Noise Sequence

- A stochastic process, η_n, is said to be white noise if the mean and its autocorrelation sequence satisfy the following:

$$\mathbb{E}[\eta_n] = 0 \text{ and } r(k) = \begin{cases} \sigma_\eta^2 & \text{if } k = 0, \\ 0, & \text{if } k \neq 0, \end{cases}$$

  where σ_η^2 is its variance.

- In other words, all variables at different time instants are uncorrelated. If, in addition, they are independent, we say that it is strictly white noise.

- It is readily seen that its PSD is given by

$$S_\eta(\omega) = \sigma_\eta^2.$$

  That is, it is constant and this is the reason it is called white noise, in analogy to the white light whose spectrum is equally spread over all the wavelengths.

Example: White Noise Sequence

- A stochastic process, η_n, is said to be white noise if the mean and its autocorrelation sequence satisfy the following:

$$\mathbb{E}[\eta_n] = 0 \text{ and } r(k) = \begin{cases} \sigma_\eta^2 & \text{if } k = 0, \\ 0, & \text{if } k \neq 0, \end{cases}$$

  where σ_η^2 is its variance.
- In other words, all variables at different time instants are uncorrelated. If, in addition, they are independent, we say that it is strictly white noise.
- It is readily seen that its PSD is given by

$$S_\eta(\omega) = \sigma_\eta^2.$$

  That is, it is constant and this is the reason it is called white noise, in analogy to the white light whose spectrum is equally spread over all the wavelengths.

Example: White Noise Sequence

- A stochastic process, η_n, is said to be white noise if the mean and its autocorrelation sequence satisfy the following:

$$\mathbb{E}[\eta_n] = 0 \text{ and } r(k) = \begin{cases} \sigma_\eta^2 & \text{if } k = 0, \\ 0, & \text{if } k \neq 0, \end{cases}$$

  where σ_η^2 is its variance.
- In other words, all variables at different time instants are uncorrelated. If, in addition, they are independent, we say that it is strictly white noise.
- It is readily seen that its PSD is given by

$$S_\eta(\omega) = \sigma_\eta^2.$$

  That is, it is constant and this is the reason it is called white noise, in analogy to the white light whose spectrum is equally spread over all the wavelengths.

Autoregressive Processes

- Autoregressive Models: Autoregressive processes are one among the most popular and widely used models. An autoregressive process of order l, denoted as AR(l), is defined via the following recursion

$$\mathrm{u}_n + a_1 \mathrm{u}_{n-1} + \ldots + a_l \mathrm{u}_{n-l} = \eta_n,$$

  where η_n is a white noise process with variance σ_η^2.
- To generate samples, one starts form some initial conditions. The input samples here correspond to the white noise sequence and the initial conditions are set equal to zero, $u_{-1} = \ldots u_{-l} = 0$.
- This is not a stationary process. Indeed, time instant $n = 0$ is distinctly different form all the rest, since it is the time that initial conditions are applied in.
- The effects of the initial conditions tend asymptotically to zero, if all the roots of the corresponding characteristic polynomial,

$$z^l + a_1 z^{l-1} + \ldots + a_l = 0,$$

  have magnitude less that unity (the solution of the corresponding homogeneous equation, without input, tends to zero). Then, it can be shown that AR(l) models become asymptotically WSS.

Autoregressive Processes

- Autoregressive Models: Autoregressive processes are one among the most popular and widely used models. An autoregressive process of order l, denoted as AR(l), is defined via the following recursion

$$\mathrm{u}_n + a_1 \mathrm{u}_{n-1} + \ldots + a_l \mathrm{u}_{n-l} = \eta_n,$$

  where η_n is a white noise process with variance σ_η^2.
- To generate samples, one starts form some initial conditions. The input samples here correspond to the white noise sequence and the initial conditions are set equal to zero, $u_{-1} = \ldots u_{-l} = 0$.
- This is not a stationary process. Indeed, time instant $n = 0$ is distinctly different form all the rest, since it is the time that initial conditions are applied in.
- The effects of the initial conditions tend asymptotically to zero, if all the roots of the corresponding characteristic polynomial,

$$z^l + a_1 z^{l-1} + \ldots + a_l = 0,$$

  have magnitude less that unity (the solution of the corresponding homogeneous equation, without input, tends to zero). Then, it can be shown that AR(l) models become asymptotically WSS.

Autoregressive Processes

- Autoregressive Models: Autoregressive processes are one among the most popular and widely used models. An autoregressive process of order l, denoted as AR(l), is defined via the following recursion

$$\mathrm{u}_n + a_1 \mathrm{u}_{n-1} + \ldots + a_l \mathrm{u}_{n-l} = \eta_n,$$

  where η_n is a white noise process with variance σ_η^2.
- To generate samples, one starts form some initial conditions. The input samples here correspond to the white noise sequence and the initial conditions are set equal to zero, $u_{-1} = \ldots u_{-l} = 0$.
- This is not a stationary process. Indeed, time instant $n = 0$ is distinctly different form all the rest, since it is the time that initial conditions are applied in.
- The effects of the initial conditions tend asymptotically to zero, if all the roots of the corresponding characteristic polynomial,

$$z^l + a_1 z^{l-1} + \ldots + a_l = 0,$$

  have magnitude less that unity (the solution of the corresponding homogeneous equation, without input, tends to zero). Then, it can be shown that AR(l) models become asymptotically WSS.

Autoregressive Processes

- Autoregressive Models: Autoregressive processes are one among the most popular and widely used models. An autoregressive process of order l, denoted as AR(l), is defined via the following recursion

$$\mathrm{u}_n + a_1 \mathrm{u}_{n-1} + \ldots + a_l \mathrm{u}_{n-l} = \eta_n,$$

  where η_n is a white noise process with variance σ_η^2.
- To generate samples, one starts form some initial conditions. The input samples here correspond to the white noise sequence and the initial conditions are set equal to zero, $u_{-1} = \ldots u_{-l} = 0$.
- This is not a stationary process. Indeed, time instant $n = 0$ is distinctly different form all the rest, since it is the time that initial conditions are applied in.
- The effects of the initial conditions tend asymptotically to zero, if all the roots of the corresponding characteristic polynomial,

$$z^l + a_1 z^{l-1} + \ldots + a_l = 0,$$

  have magnitude less that unity (the solution of the corresponding homogeneous equation, without input, tends to zero). Then, it can be shown that AR(l) models become asymptotically WSS.

Autoregressive Processes

- Autocorrelation sequence of an AR process: Multiplying both sides of the defining equation with $\mathrm{u}_{n-k},\ k>0$, and taking the expectation, we obtain

$$\sum_{i=0}^{l} a_i \mathbb{E}[\mathrm{u}_{n-i}\mathrm{u}_{n-k}] = \mathbb{E}[\eta_n \mathrm{u}_{n-k}],\ k>0,\ \text{where } a_0 := 1,\ \text{or}$$

$$\sum_{i=1}^{l} a_i r(k-i) = 0.$$

We have used the fact that $\mathbb{E}[\eta_n \mathrm{u}_{n-k}]$, $k>0$ is zero. Indeed, u_{n-k} depends recursively on $\eta_{n-k}, \eta_{n-k-1}\ldots$, which are all uncorrelated to η_n, since this is a white noise process.

- Note that the above is a difference equation and it can be solved, provided we have the initial conditions. To this end, we again multiply the defining equation by u_n and take expectations, which results in

$$\sum_{i=0}^{l} a_i r(i) = \sigma_\eta^2,$$

since u_n recursively depends on η_n, which contributes the σ_η^2 term, and $\eta_{n-1}, \ldots$, which result to zeros.

Autoregressive Processes

- Autocorrelation sequence of an AR process: Multiplying both sides of the defining equation with $\mathrm{u}_{n-k},\ k>0$, and taking the expectation, we obtain

$$\sum_{i=0}^{l} a_i \mathbb{E}[\mathrm{u}_{n-i}\mathrm{u}_{n-k}] = \mathbb{E}[\eta_n \mathrm{u}_{n-k}],\ k>0,\ \text{where } a_0 := 1,\ \text{or}$$

$$\sum_{i=1}^{l} a_i r(k-i) = 0.$$

We have used the fact that $\mathbb{E}[\eta_n \mathrm{u}_{n-k}],\ k>0$ is zero. Indeed, u_{n-k} depends recursively on $\eta_{n-k}, \eta_{n-k-1}\ldots$, which are all uncorrelated to η_n, since this is a white noise process.

- Note that the above is a difference equation and it can be solved, provided we have the initial conditions. To this end, we again multiply the defining equation by u_n and take expectations, which results in

$$\sum_{i=0}^{l} a_i r(i) = \sigma_\eta^2,$$

since u_n recursively depends on η_n, which contributes the σ_η^2 term, and $\eta_{n-1}, \ldots$, which result to zeros.

Autoregressive Processes

- Autocorrelation sequence of an AR process: Multiplying both sides of the defining equation with $\mathrm{u}_{n-k},\ k>0$, and taking the expectation, we obtain

$$\sum_{i=0}^{l} a_i \mathbb{E}[\mathrm{u}_{n-i}\mathrm{u}_{n-k}] = \mathbb{E}[\eta_n \mathrm{u}_{n-k}],\ k>0,\ \text{where } a_0 := 1,\ \text{or}$$

$$\sum_{i=1}^{l} a_i r(k-i) = 0.$$

We have used the fact that $\mathbb{E}[\eta_n \mathrm{u}_{n-k}]$, $k>0$ is zero. Indeed, u_{n-k} depends recursively on $\eta_{n-k}, \eta_{n-k-1}\ldots$, which are all uncorrelated to η_n, since this is a white noise process.

- Note that the above is a difference equation and it can be solved, provided we have the initial conditions. To this end, we again multiply the defining equation by u_n and take expectations, which results in

$$\sum_{i=0}^{l} a_i r(i) = \sigma_\eta^2,$$

since u_n recursively depends on η_n, which contributes the σ_η^2 term, and $\eta_{n-1}, \ldots$, which result to zeros.

Autoregressive Processes

- Yule-Walker equations: Combining the previous two equations, we end up with the elegant linear system of equations:

$$\begin{bmatrix} r(0) & r(1) & \dots & r(l) \\ r(1) & r(0) & \dots & r(l-1) \\ \vdots & \vdots & \vdots & \vdots \\ r(l) & r(l-1) & \dots & r(0) \end{bmatrix} \begin{bmatrix} 1 \\ a_1 \\ \vdots \\ a_l \end{bmatrix} = \begin{bmatrix} \sigma_\eta^2 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

  These are known as the Yule-Walker equations, whose solution results in the values, $r(0), \dots, r(l)$, which are then used as the initial conditions to solve the corresponding difference equation and obtain $r(k)$, $\forall k \in \mathbb{Z}$.

- Observe the special structure of the matrix in the linear system. This type of matrices are known as Toeplitz. All the elements along any diagonal are equal.

- Moving Average (MA) models: These are defined by the recursion,

$$\mathrm{u}_n = b_1 \eta_n + \dots + b_m \eta_{n-m}.$$

- Autoregressive-Moving Average (ARMA) models: These are defined as,

$$\mathrm{u}_n + a_1 \mathrm{u}_{n-1} + \dots \mathrm{u}_{n-l} = b_1 \eta_n + \dots + b_m \eta_{n-m}.$$

Autoregressive Processes

- Yule-Walker equations: Combining the previous two equations, we end up with the elegant linear system of equations:

$$\begin{bmatrix} r(0) & r(1) & \dots & r(l) \\ r(1) & r(0) & \dots & r(l-1) \\ \vdots & \vdots & \vdots & \vdots \\ r(l) & r(l-1) & \dots & r(0) \end{bmatrix} \begin{bmatrix} 1 \\ a_1 \\ \vdots \\ a_l \end{bmatrix} = \begin{bmatrix} \sigma_\eta^2 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

  These are known as the Yule-Walker equations, whose solution results in the values, $r(0), \dots, r(l)$, which are then used as the initial conditions to solve the corresponding difference equation and obtain $r(k),\ \forall k \in \mathbb{Z}$.

- Observe the special structure of the matrix in the linear system. This type of matrices are known as Toeplitz. All the elements along any diagonal are equal.

- Moving Average (MA) models: These are defined by the recursion,

$$\mathrm{u}_n = b_1 \eta_n + \dots + b_m \eta_{n-m}.$$

- Autoregressive-Moving Average (ARMA) models: These are defined as,

$$\mathrm{u}_n + a_1 \mathrm{u}_{n-1} + \dots \mathrm{u}_{n-l} = b_1 \eta_n + \dots + b_m \eta_{n-m}.$$

Autoregressive Processes

- Yule-Walker equations: Combining the previous two equations, we end up with the elegant linear system of equations:

$$\begin{bmatrix} r(0) & r(1) & \dots & r(l) \\ r(1) & r(0) & \dots & r(l-1) \\ \vdots & \vdots & \vdots & \vdots \\ r(l) & r(l-1) & \dots & r(0) \end{bmatrix} \begin{bmatrix} 1 \\ a_1 \\ \vdots \\ a_l \end{bmatrix} = \begin{bmatrix} \sigma_\eta^2 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

  These are known as the Yule-Walker equations, whose solution results in the values, $r(0), \dots, r(l)$, which are then used as the initial conditions to solve the corresponding difference equation and obtain $r(k),\ \forall k \in \mathbb{Z}$.

- Observe the special structure of the matrix in the linear system. This type of matrices are known as Toeplitz. All the elements along any diagonal are equal.

- Moving Average (MA) models: These are defined by the recursion,

$$\mathrm{u}_n = b_1 \eta_n + \dots + b_m \eta_{n-m}.$$

- Autoregressive-Moving Average (ARMA) models: These are defined as,

$$\mathrm{u}_n + a_1 \mathrm{u}_{n-1} + \dots \mathrm{u}_{n-l} = b_1 \eta_n + \dots + b_m \eta_{n-m}.$$

- Yule-Walker equations: Combining the previous two equations, we end up with the elegant linear system of equations:

$$\begin{bmatrix} r(0) & r(1) & \dots & r(l) \\ r(1) & r(0) & \dots & r(l-1) \\ \vdots & \vdots & \vdots & \vdots \\ r(l) & r(l-1) & \dots & r(0) \end{bmatrix} \begin{bmatrix} 1 \\ a_1 \\ \vdots \\ a_l \end{bmatrix} = \begin{bmatrix} \sigma_\eta^2 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

  These are known as the Yule-Walker equations, whose solution results in the values, $r(0), \dots, r(l)$, which are then used as the initial conditions to solve the corresponding difference equation and obtain $r(k),\ \forall k \in \mathbb{Z}$.

- Observe the special structure of the matrix in the linear system. This type of matrices are known as Toeplitz. All the elements along any diagonal are equal.

- Moving Average (MA) models: These are defined by the recursion,

$$\mathrm{u}_n = b_1 \eta_n + \dots + b_m \eta_{n-m}.$$

- Autoregressive-Moving Average (ARMA) models: These are defined as,

$$\mathrm{u}_n + a_1 \mathrm{u}_{n-1} + \dots \mathrm{u}_{n-l} = b_1 \eta_n + \dots + b_m \eta_{n-m}.$$

Autoregressive Processes

- Yule-Walker equations: Combining the previous two equations, we end up with the elegant linear system of equations:

$$\begin{bmatrix} r(0) & r(1) & \dots & r(l) \\ r(1) & r(0) & \dots & r(l-1) \\ \vdots & \vdots & \vdots & \vdots \\ r(l) & r(l-1) & \dots & r(0) \end{bmatrix} \begin{bmatrix} 1 \\ a_1 \\ \vdots \\ a_l \end{bmatrix} = \begin{bmatrix} \sigma_\eta^2 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

  These are known as the Yule-Walker equations, whose solution results in the values, $r(0), \dots, r(l)$, which are then used as the initial conditions to solve the corresponding difference equation and obtain $r(k),\ \forall k \in \mathbb{Z}$.

- Observe the special structure of the matrix in the linear system. This type of matrices are known as Toeplitz. All the elements along any diagonal are equal.

- Moving Average (MA) models: These are defined by the recursion,

$$\mathrm{u}_n = b_1 \eta_n + \dots + b_m \eta_{n-m}.$$

- Autoregressive-Moving Average (ARMA) models: These are defined as,

$$\mathrm{u}_n + a_1 \mathrm{u}_{n-1} + \dots \mathrm{u}_{n-l} = b_1 \eta_n + \dots + b_m \eta_{n-m}.$$

Example: AR processes

- Consider the AR(1) process. The goal is to compute the corresponding autocorrelation sequence. To this end, we have

$$\mathrm{u}_n + a\mathrm{u}_{n-1} = \eta_n.$$

- Following the general methodology explained before, we have

$$\begin{aligned} r(k) + ar(k-1) &= 0, \ k = 1, 2, \dots \\ r(0) + ar(1) &= \sigma_\eta^2. \end{aligned}$$

  Considering the first equation for $k = 1$ together with the second one readily results in

$$r(0) = \frac{\sigma_\eta^2}{1 - a^2}.$$

  Plugging this value in the difference equation, we recursively obtain

$$r(k) = (-a)^{|k|} \frac{\sigma_\eta^2}{1 - a^2}, \ k = 0, \pm 1, \pm 2, \dots,$$

  where we used the property, $r(k) = r(-k)$.

- Remark: Observe that if $|a| > 1$, $r(0) < 0$ which is meaningless. Also, $|a| < 1$ guarantees that the magnitude of the root of the characteristic polynomial ($z_* = -a$) is smaller than one. Moreover, $|a| < 1$ guarantees that $r(k) \longrightarrow 0$ as $k \longrightarrow \infty$. This is in line with common sense, since random variables which are far away in time must be uncorrelated.

Example: AR processes

- Consider the AR(1) process. The goal is to compute the corresponding autocorrelation sequence. To this end, we have

$$\mathrm{u}_n + a\mathrm{u}_{n-1} = \eta_n.$$

- Following the general methodology explained before, we have

$$\begin{aligned} r(k) + ar(k-1) &= 0, \ k = 1, 2, \ldots \\ r(0) + ar(1) &= \sigma_\eta^2. \end{aligned}$$

Considering the first equation for $k = 1$ together with the second one readily results in

$$r(0) = \frac{\sigma_\eta^2}{1 - a^2}.$$

Plugging this value in the difference equation, we recursively obtain

$$r(k) = (-a)^{|k|} \frac{\sigma_\eta^2}{1 - a^2}, \ k = 0, \pm 1, \pm 2, \ldots,$$

where we used the property, $r(k) = r(-k)$.

- Remark: Observe that if $|a| > 1$, $r(0) < 0$ which is meaningless. Also, $|a| < 1$ guarantees that the magnitude of the root of the characteristic polynomial ($z_* = -a$) is smaller than one. Moreover, $|a| < 1$ guarantees that $r(k) \longrightarrow 0$ as $k \longrightarrow \infty$. This is in line with common sense, since random variables which are far away in time must be uncorrelated.

Example: AR processes

- Consider the AR(1) process. The goal is to compute the corresponding autocorrelation sequence. To this end, we have

$$\mathrm{u}_n + a\mathrm{u}_{n-1} = \eta_n.$$

- Following the general methodology explained before, we have

$$\begin{aligned} r(k) + ar(k-1) &= 0, \ k = 1, 2, \ldots \\ r(0) + ar(1) &= \sigma_\eta^2. \end{aligned}$$

Considering the first equation for $k = 1$ together with the second one readily results in

$$r(0) = \frac{\sigma_\eta^2}{1 - a^2}.$$

Plugging this value in the difference equation, we recursively obtain

$$r(k) = (-a)^{|k|} \frac{\sigma_\eta^2}{1 - a^2}, \ k = 0, \pm 1, \pm 2, \ldots,$$

where we used the property, $r(k) = r(-k)$.

- Remark: Observe that if $|a| > 1$, $r(0) < 0$ which is meaningless. Also, $|a| < 1$ guarantees that the magnitude of the root of the characteristic polynomial ($z_* = -a$) is smaller than one. Moreover, $|a| < 1$ guarantees that $r(k) \longrightarrow 0$ as $k \longrightarrow \infty$. This is in line with common sense, since random variables which are far away in time must be uncorrelated.

Example: AR processes

- Consider the AR(1) process. The goal is to compute the corresponding autocorrelation sequence. To this end, we have

$$\mathrm{u}_n + a\mathrm{u}_{n-1} = \eta_n.$$

- Following the general methodology explained before, we have

$$\begin{aligned} r(k) + ar(k-1) &= 0, \ k = 1, 2, \ldots \\ r(0) + ar(1) &= \sigma_\eta^2. \end{aligned}$$

Considering the first equation for $k = 1$ together with the second one readily results in

$$r(0) = \frac{\sigma_\eta^2}{1 - a^2}.$$

Plugging this value in the difference equation, we recursively obtain

$$r(k) = (-a)^{|k|} \frac{\sigma_\eta^2}{1 - a^2}, \ k = 0, \pm 1, \pm 2, \ldots,$$

where we used the property, $r(k) = r(-k)$.

- Remark: Observe that if $|a| > 1$, $r(0) < 0$ which is meaningless. Also, $|a| < 1$ guarantees that the magnitude of the root of the characteristic polynomial ($z_* = -a$) is smaller than one. Moreover, $|a| < 1$ guarantees that $r(k) \longrightarrow 0$ as $k \longrightarrow \infty$. This is in line with common sense, since random variables which are far away in time must be uncorrelated.

Example: AR processes

- Consider the AR(1) process. The goal is to compute the corresponding autocorrelation sequence. To this end, we have

$$\mathrm{u}_n + a\mathrm{u}_{n-1} = \eta_n.$$

- Following the general methodology explained before, we have

$$\begin{aligned} r(k) + ar(k-1) &= 0, \ k = 1, 2, \dots \\ r(0) + ar(1) &= \sigma_\eta^2. \end{aligned}$$

Considering the first equation for $k = 1$ together with the second one readily results in

$$r(0) = \frac{\sigma_\eta^2}{1 - a^2}.$$

Plugging this value in the difference equation, we recursively obtain

$$r(k) = (-a)^{|k|} \frac{\sigma_\eta^2}{1 - a^2}, \ k = 0, \pm 1, \pm 2, \dots,$$

where we used the property, $r(k) = r(-k)$.

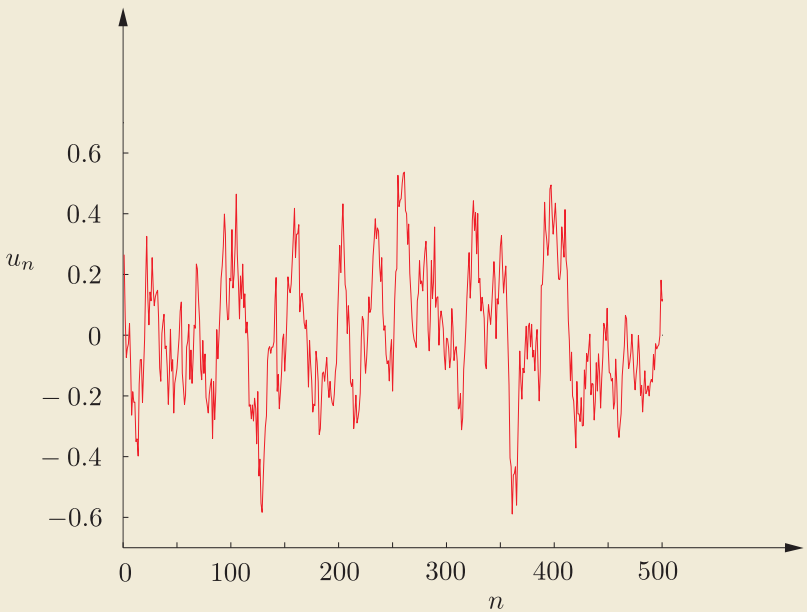
- Remark: Observe that if $|a| > 1$, $r(0) < 0$ which is meaningless. Also, $|a| < 1$ guarantees that the magnitude of the root of the characteristic polynomial ($z_* = -a$) is smaller than one. Moreover, $|a| < 1$ guarantees that $r(k) \longrightarrow 0$ as $k \longrightarrow \infty$. This is in line with common sense, since random variables which are far away in time must be uncorrelated.

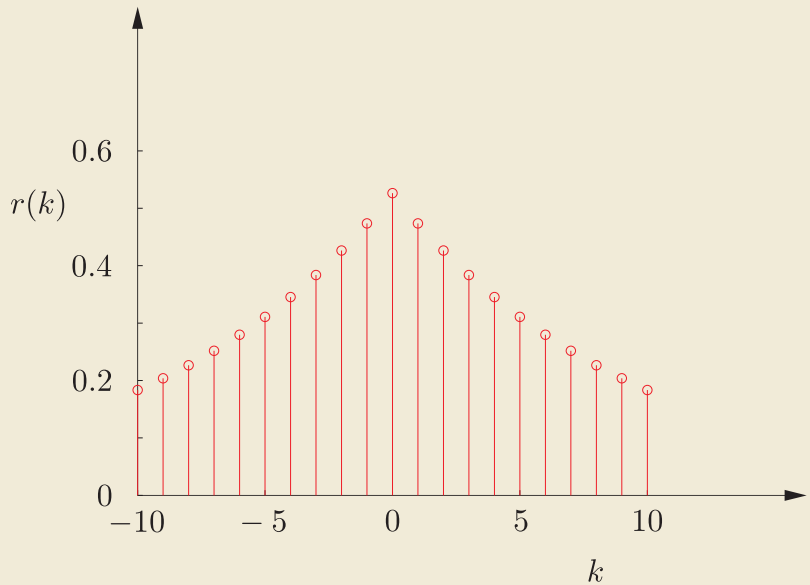
Example: AR processes

- Plots of a realization (left) and the autocorrelation sequence (right) corresponding to the value $a = -0.9$.

Example: AR processes

- Plots of a realization (left) and the autocorrelation sequence (right) corresponding to the value $a = -0.9$.



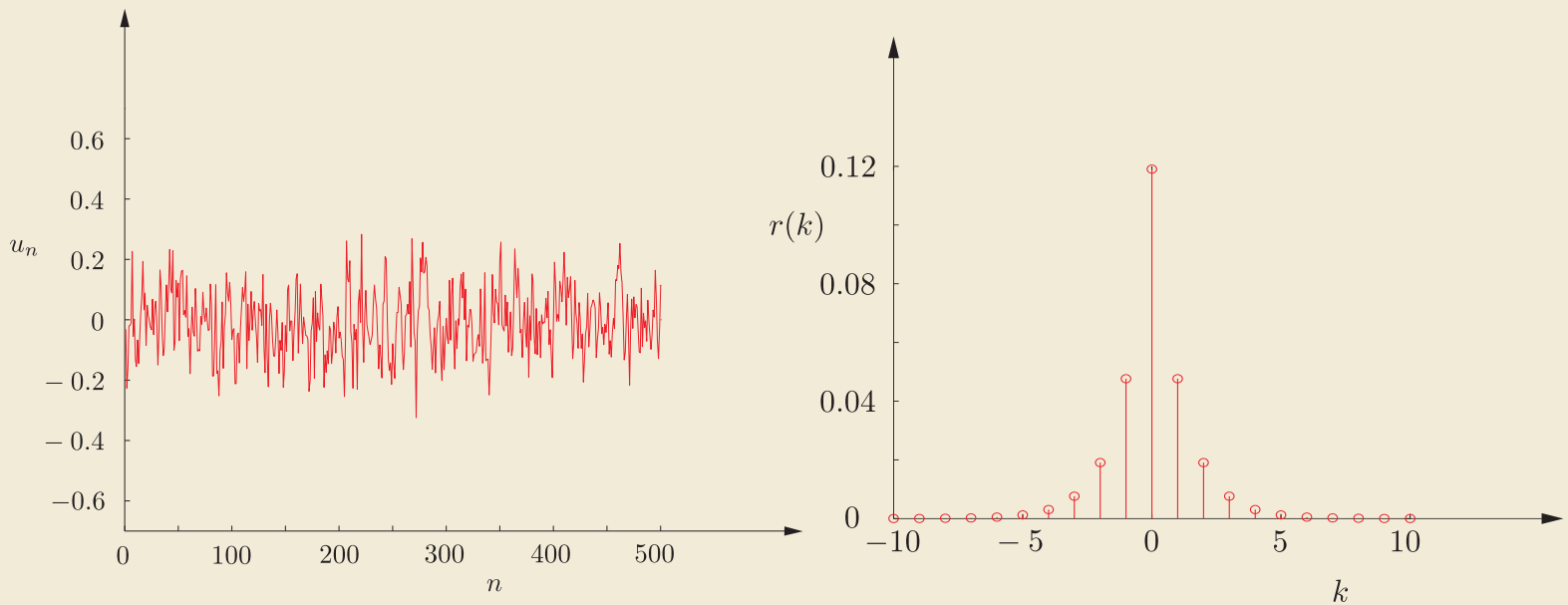


Example: AR processes

- Plots of a realization (left) and the autocorrelation sequence (right) corresponding to the value $a = -0.4$. Compared to the value of $a = -0.9$, the variables at different time instants are less correlated an the autocorrelation sequence fades to zero much faster.

Example: AR processes

- Plots of a realization (left) and the autocorrelation sequence (right) corresponding to the value $a = -0.4$. Compared to the value of $a = -0.9$, the variables at different time instants are less correlated an the autocorrelation sequence fades to zero much faster.

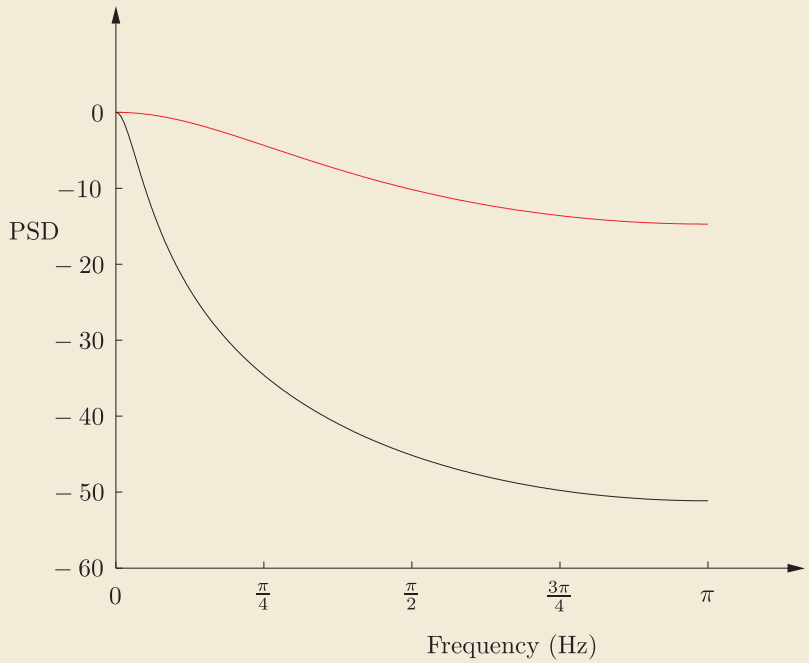


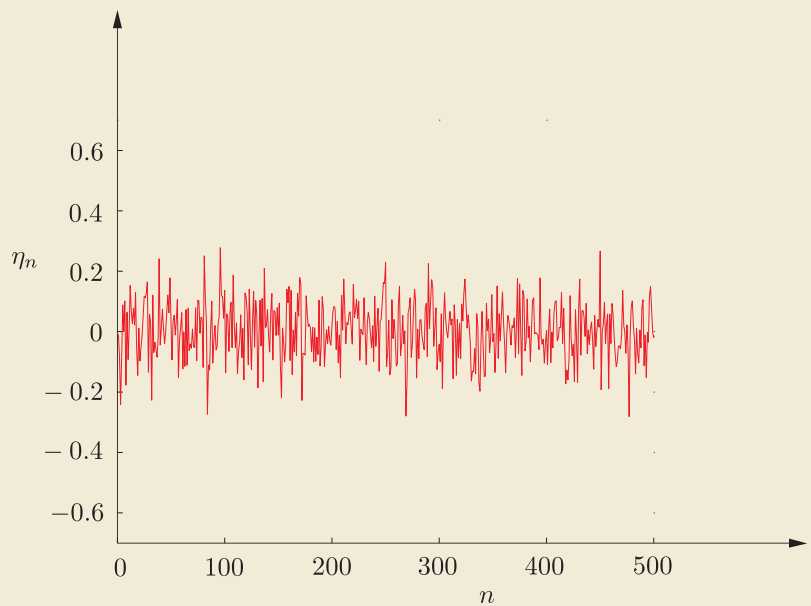
Example: AR processes

- Plots of the PSDs for the two previous cases (left). To the right, a realization of a white noise sequence is given for the sake of comparison with the previously plotted ones.

Example: AR processes

- Plots of the PSDs for the two previous cases (left). To the right, a realization of a white noise sequence is given for the sake of comparison with the previously plotted ones.





$a = -0.9$ (black), $a = -0.4$ (red)