2. Review of probability theory

Statistical Signal Processing
Version 1.7

Prof. Peter Schreier, Ph.D.
Signal and System Theory Group
Faculty of Electrical Engineering, Computer Science, and Mathematics
Universität Paderborn

2.1 Axiomatic approach to probability theory

Sample spaces and set operations

- The **sample space** Ω is the set of all possible distinct **outcomes** ω of a random experiment
  - Roll of a die: $\Omega = \{1, 2, 3, 4, 5, 6\}$
- An **event** is a set, i.e., collection, of outcomes
  - Roll an even number: $E = \{2, 4, 6\}$

Set operations:

- The **complement** of E in Ω is $\overline{E} = \{\omega \in \Omega : \omega \notin E\}$
- The **empty set** or **null set** is $\emptyset = \overline{\Omega}$
- The **union** of E and F is $E \cup F$
- The **intersection** of E and F is $E \cap F$

These operations can be illustrated using **Venn diagrams**.

Two sets E and F are called **mutually exclusive** or **disjoint** if $E \cap F = \emptyset$. Three or more sets E_i are called **mutually exclusive** or **pairwise disjoint** if $E_i \cap E_j = \emptyset$ for all $i \neq j$.

Probability based on axiomatic theory

This is an idealization/abstraction.

Axioms of probability

Probability is a set function $P(\cdot)$ that assigns to every event E a number $P(E)$, called the probability of E such that

1. $P(E) \geq 0$
2. $P(\Omega) = 1$, where Ω is the sample space
3. $P(E \cup F) = P(E) + P(F)$, if $E \cap F = \emptyset$

So what is $P(E \cup F)$ if E and F are not disjoint?
$\Omega = E \cup \overline{E}$; $E \cap \overline{E} = \emptyset$
$E \cup F = (E \cap \overline{F}) \cup (\overline{E} \cap F) \cup (E \cap F)$ (three disjoint sets)
By axiom 3: $P(E \cup F) = P(E \cap \overline{F}) + P(\overline{E} \cap F) + P(E \cap F)$
Also, $P(E \cap F) + P(E \cap \overline{F}) = P(E)$ and $P(E \cap F) + P(\overline{E} \cap F) = P(F)$
Thus, $P(E \cup F) = P(E) - P(E \cap F) + P(F) - P(E \cap F) + P(E \cap F)$
and finally:

$$P(E \cup F) = P(E) + P(F) - P(E \cap F)$$

Joint and conditional probabilities

- $P(EF) \triangleq P(E \cap F)$ is the **joint probability** of E and F
- $P(E|F) \triangleq \dfrac{P(EF)}{P(F)}$ is the **conditional probability** of E given F
  $P(EF) = P(E|F)P(F)$

Statistical independence

Two events are called **statistically independent** if $P(EF) = P(E)P(F)$.

- Then, $P(E|F) = \dfrac{P(EF)}{P(F)} = \dfrac{P(E)P(F)}{P(F)} = P(E)$
- Also, $P(F|E) = \dfrac{P(FE)}{P(E)} = \dfrac{P(E)P(F)}{P(E)} = P(F)$

Bayes' rule

$$P(F|E) = \frac{P(E|F)P(F)}{P(E|F)P(F) + P(E|\overline{F})P(\overline{F})}$$

Example 1. Toss a fair coin three times. What is the probability of
$A = \{$ at least two Heads show $\}$
$B = \{$ Head and Tail alternate $\}$
Are events A and B statistically independent?

Example 2. You have 5 coins. One of them is a fake coin which contains Heads on both sides. You draw one of these coins at random and put it on the table. Assuming that Head shows, what is the probability of Head on the other side of the coin?
The sample space is

coin	1	2	3	4	5
side 1	H	H	H	H	H
side 2	T	T	T	T	H

Define events
$A = \{$ Head shows $\}$
$B = \{$ Head is on the other side of the coin $\}$
What is $P(A)$, $P(B)$, $P(AB)$, $P(B|A)$?

More than two events

Let $E = E_1 \cup E_2 \cup \cdots \cup E_K$, where $E_i \cap E_j = \emptyset$ for $i \neq j$ (mutually exclusive or pairwise disjoint events). Then the **total probability** is

$$P(E) = \sum_{i=1}^{K} P(E_i)$$

In general, however, $P(E_1 \cup E_2 \cup \cdots \cup E_K) \leq \sum_{i=1}^{K} P(E_i)$

Total law of probability and general Bayes' rule

Let $E = E_1 \cup E_2 \cup \cdots \cup E_K$, where $E_i \cap E_j = \emptyset$ for $i \neq j$. Then, for any event F,

$$P(F) = \sum_{i=1}^{K} P(F|E_i)P(E_i)$$

The general form of Bayes' rule is

$$P(E_j|F) = \frac{P(F|E_j)P(E_j)}{\sum_{i=1}^{K} P(F|E_i)P(E_i)}$$

2.2 Discrete random variables

Random variables

Definition

A **random variable** (RV) $X(\omega)$ is a function that assigns a real number $X(\omega)$ to every possible outcome $\omega \in \Omega$ of a random experiment.

Example. Let $X(\omega)$ be the number of heads when tossing a coin three times. The sample space is $\Omega = \{\text{TTT, TTH, THT, THH, HTT, HTH, HHT, HHH}\}$.

$$X(\omega) = \begin{cases} 0, & \omega = \text{TTT}, \\ 1, & \omega \in \{\text{TTH, THT, HTT}\} \\ 2, & \omega \in \{\text{THH, HTH, HHT}\} \\ 3, & \omega = \text{HHH} \end{cases}$$

Each real number $X(\omega)$ then has an associated probability. We write $P(X(\omega) = x) = P_X(x)$.

- We often drop ω for convenience.
- We differentiate between the RV X and the value x it takes.

Example 1. For the example just considered (number of heads when tossing a coin 3x), we get

x	0	1	2	3
$P_X(x)$	1/8	3/8	3/8	1/8

We can also compute

- $P(X \geq 2) = 1/2$ (at least two Heads);
- $\sum_x P_X(x) = 1$

Example 2. Roll two dice. Let Y be the sum of the two dice.

y	2	3	4	5	6	7	8	9	10	11	12
$P_Y(y)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

We can also compute

- $P(Y \leq 4) = 1/6$
- $P(4 \leq Y \leq 6) = P(Y \leq 6) - P(Y \leq 3) = 1/3$

Discrete random variables

The random variables we have considered so far are **discrete RVs**. This means that they only take on a **countable** number of possible values x_i.

- Examples of **discrete RVs**: rolling a die; the number of cars passing over a bridge within one hour
- Examples of **continuous RVs**: spinning a wheel and measuring the angle at which it stops; the amount of time one has to wait in the cafeteria to get served

We call $P_X(x) \triangleq P(X = x)$ the **probability distribution** ("probability mass function" is also common). It satisfies

- $0 \leq P_X(x) \leq 1$
- $\sum_x P_X(x) = 1$

There are important special probability distributions, such as the **Poisson distribution** $P_X(x) = \dfrac{\lambda^x \exp(-\lambda)}{x!}$, $x = 0, 1, 2, \ldots$, with parameter $\lambda > 0$.

Two discrete random variables

The following are fundamental:

- $P_{XY}(x, y) \triangleq P(X = x, Y = y)$: **joint probability distribution** of X and Y
- $P_X(x) = \sum_y P_{XY}(x, y)$: **marginal probability distribution**
- $P_Y(y) = \sum_x P_{XY}(x, y)$: **marginal probability distribution**
- $P_{X|Y}(x|y) = \dfrac{P_{XY}(x, y)}{P_Y(y)}$: **conditional distribution** of X given Y
- $P_{Y|X}(y|x) = \dfrac{P_{XY}(x, y)}{P_X(x)}$: **conditional distribution** of Y given X
- These definitions also generalize to more than two random variables.

Two random variables X and Y are (statistically) **independent** if $P_{XY}(x, y) = P_X(x)P_Y(y)$. This implies $P_{X|Y}(x|y) = P_X(x)$ and $P_{Y|X}(y|x) = P_Y(y)$.

Example

A hat contains 4 balls, which are labelled "AA", "AB", "BA", and "BB". Two balls are drawn randomly from this hat. Define two random variables X and Y. If ball "AA" is among the two balls drawn from the hat, then $X = 1$, otherwise $X = 0$. The RV Y is the total number of "B"s on the two balls drawn.

- Derive the joint and marginal distributions of X and Y.

$P_{XY}(x,y)$	$y=1$	$y=2$	$y=3$	$P_X(x)$
$x=0$	0	1/6	1/3	1/2
$x=1$	1/3	1/6	0	1/2
$P_Y(y)$	1/3	1/3	1/3	

- What is the conditional probability distribution of X given Y?

$P_{X\|Y}(x\|y)$	$y=1$	$y=2$	$y=3$
$x=0$	0	1/2	1
$x=1$	1	1/2	0

- Are X and Y independent?
  No they are not. E.g., $P_{XY}(0,1) = 0 \neq \frac{1}{3} \cdot \frac{1}{2} = P_X(0)P_Y(1)$

Expectation (statistical averages)

The simplest statistical average is the **mean** (first-order moment) of X:

$$E\{X\} = \mu_X \triangleq \sum_x x P_X(x)$$

More generally, we define

$$E\{g(X)\} \triangleq \sum_x g(x) P_X(x)$$

We call $E\{\cdot\}$ the **expectation operator**. It is linear:

$$E\{aX + bY\} = aE\{X\} + bE\{Y\}$$

We can also define the **variance** (central second-order moment)

$$\text{var}(X) \triangleq E\{(X - \mu_X)^2\} = \sum_x (x - \mu_X)^2 P_X(x),$$

the **standard deviation** $\sigma_X = \sqrt{\text{var}(X)}$, and the **second-order moment** $E\{X^2\} = \mu_{X^2}$. They are related as $\text{var}(X) = \sigma_X^2 = \mu_{X^2} - \mu_X^2$.

Why work with moments?

- Moments **summarize** certain important aspects of random variables. However, a **complete** statistical description of a random variable is only given by the probability distribution.
- Remember the difference: $E(X)$ is a number (deterministic) $\longleftrightarrow$ X is a **random** variable
- It is often easier to work with a few numbers than a probability distribution (which is a function)
- Often, moments are all we know or are able to estimate. In many cases, knowledge of the probability distribution is an unrealistic assumption.

There are also **higher-order** moments:

- 3rd order: **skewness** $E\left\{\left(\frac{X-\mu_X}{\sigma_X}\right)^3\right\}$ measures the asymmetry of a distribution
- 4th order: **kurtosis** $\frac{E\{X^4\}}{\sigma_X^4} - 3$ measures how concentrated a distribution is

Expectation of two random variables

For two random variables X and Y the expectation is defined as

$$E\{g(X,Y)\} = \sum_x \sum_y g(x,y) P_{XY}(x,y)$$

In particular, for $g(x,y) = xy$ we obtain the **correlation**

$$\mu_{XY} = E\{XY\} = \sum_x \sum_y xy P_{XY}(x,y),$$

and for $g(x,y) = (x-\mu_X)(y-\mu_y)$ the **covariance**

$$\sigma_{XY} = E\{(X-\mu_X)(Y-\mu_Y)\} = \sum_x \sum_y (x-\mu_X)(y-\mu_Y) P_{XY}(x,y).$$

- There is the relationship $\sigma_{XY} = \mu_{XY} - \mu_X \mu_Y$.
- If $E\{XY\} = E\{X\}E\{Y\}$, then X and Y are **uncorrelated**.
- Statistical independence implies uncorrelatedness, but **not vice versa**.

Correlation coefficient

The Cauchy-Schwarz inequality says

$$|\sigma_{XY}| \leq \sigma_X \sigma_Y$$

with equality iff $X = aY + b$. We may therefore define the

Correlation coefficient

The **correlation coefficient** between two random variables X and Y is

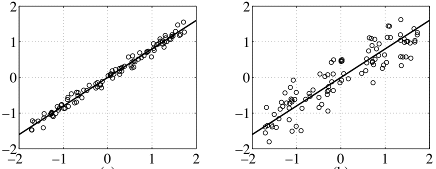
$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

The correlation coefficient satisfies $-1 \leq \rho_{XY} \leq 1$.

- If $|\rho_{XY}| = 1$, then X and Y are linearly related: $X = aY + b$. We say that X is **perfectly linearly predictable/estimable** from Y (or vice versa).
- If $\rho_{XY} = 0$, then $\sigma_{XY} = \mu_{XY} - \mu_X \mu_Y = 0$ and thus X and Y are **uncorrelated**. In this case a linear estimator of X from Y (or vice versa) would be **useless**.

Linear estimation

These figures show samples of two correlated RVs X and Y.



(a) (b)

It seems that one should be able to construct a linear estimator $\hat{Y} = aX + b$ that predicts Y from X. As a measure for the estimation quality we would like to minimize the **mean-squared error** (MSE)

$$\begin{aligned}\mathrm{MSE} &= E\{(Y-\hat{Y})^2\} = E\{(Y-aX-b)^2\}\\ &= E\{Y^2\} - 2aE\{XY\} - 2b\mu_Y + 2ab\mu_X + a^2E\{X^2\} + b^2\end{aligned}$$

In order to minimize the MSE, we set

$$\frac{\partial \mathrm{MSE}}{\partial a} = -2E\{XY\} + 2b\mu_X + 2aE\{X^2\} = 0$$

$$\frac{\partial \mathrm{MSE}}{\partial b} = -2\mu_Y + 2a\mu_X + 2b = 0$$

Linear estimation

After some elementary algebra, we obtain

$$a = \frac{E\{(X-\mu_X)(Y-\mu_Y)\}}{E(X-\mu_X)^2} = \rho_{XY}\frac{\sigma_Y}{\sigma_X}$$
$$b = \mu_Y - \rho_{XY}\frac{\sigma_Y}{\sigma_X}\mu_X$$

Thus, the optimal linear estimator that minimizes the mean-squared error is

$$\hat{Y} = \mu_Y + \rho_{XY}\frac{\sigma_Y}{\sigma_X}(X-\mu_X)$$

The corresponding MSE is

$$\text{MSE} = \sigma_Y^2(1-\rho_{XY}^2)$$

We observe that

- $|\rho_{XY}| = 1 \implies \text{MSE} = 0$
- $\rho_{XY} = 0 \implies \text{MSE} = \sigma_Y^2$ and the best estimator is $\hat{Y} = \mu_Y$

Conditional expectation

Conditioning on an **event** $Y = y$:

$$E\{X|Y=y\} = \sum_x x P_{X|Y}(x|y)$$

Conditioning on a random variable

The conditional expectation $E\{X|Y\}$ is itself a random variable.

We can write

$$\begin{aligned} E\{X\} &= \sum_x x P_X(x) = \sum_x x \sum_y P_{XY}(x,y) = \sum_y \left[\sum_x x P_{X|Y}(x|y) \right] P_Y(y) \\ &= \sum_y E\{X|Y=y\} P_Y(y) = E\{E\{X|Y\}\}. \end{aligned}$$

Note that the inner expectation is with respect to X and the outer with respect to Y.

2.3 Continuous random variables

Continuous random variables

Example. Let's say we spin a wheel and measure the angle X at which it stops. It seems to be a reasonable assumption that every angle $x \in [0, 2\pi)$ is equally likely. The probability that the wheel stops between angles x_1 and x_2 is therefore

$$P(x_1 < X \leq x_2) = \frac{x_2 - x_1}{2\pi}$$

Note that this probability is greater than 0 only if $x_2 = x_1 + \delta$ for $\delta > 0$. That is, $P(X = x) = 0$. This means that for continuous random variables we always have to work with probabilities of the form $P(X \leq x)$. We may, however, differentiate this function to obtain

$$f_X(x) = \frac{d}{dx} P(X \leq x)$$

In this example, $f_X(x) = \begin{cases} 1/(2\pi) & 0 \leq x < 2\pi \\ 0 & \text{else.} \end{cases}$

Probability density and cumulative distribution functions

Definitions

The function

$$F_X(x) \triangleq P(X \leq x) = \int_{-\infty}^{x} f_X(\xi)\, d\xi$$

is called **cumulative distribution function (cdf)**. Its derivative

$$f_X(x) = \frac{d}{dx} F_X(x)$$

is the **probability density function (pdf)**.

The following are easy to establish:

- $P(a < X \leq b) = F_X(b) - F_X(a)$
- $F_X(x)$ is a nondecreasing function
- $f_X(x) \geq 0$ but **not necessarily** $f_X(x) \leq 1$
- $F_X(-\infty) = 0$ and $F_X(\infty) = \int_{-\infty}^{\infty} f_X(x)\, dx = 1$

Example

Consider a model train running on a circular track with radius r at constant speed. Let X be the x-position at a randomly chosen time instant, and let Θ be the corresponding angle on the track. The angle Θ is *uniformly distributed* on $[-\pi, \pi)$.

- Consider the event $\{X \le x\}$. If $x < -r$, then this event has probability zero. If $x \ge r$, then this event has probability 1.
- If $-r \le x \le r$, then $\{X \le x\} = \{\arccos(\frac{x}{r}) \le \Theta \le \pi\} \cup \{-\pi \le \Theta \le -\arccos(\frac{x}{r})\}$, so

$$F_X(x) = 2\int_{\arccos(\frac{x}{r})}^{\pi} \frac{1}{2\pi}\, d\xi = \frac{1}{\pi}[\pi - \arccos(\tfrac{x}{r})] = \frac{1}{\pi}\arccos(-\tfrac{x}{r})$$

- For the probability density, we obtain

$$f_X(x) = \frac{d}{dx}F_X(x) = \begin{cases} \frac{1}{r\pi\sqrt{1-(x/r)^2}} & -r < x < r \\ 0 & x < -r \text{ or } x > r \end{cases}$$

Two random variables

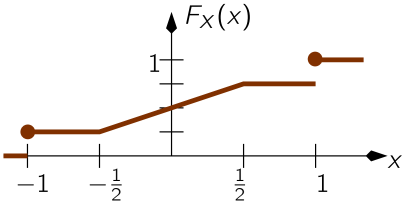
We define:

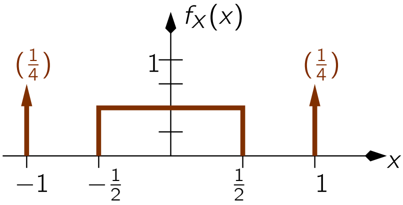
- $F_{XY}(x,y) = P(X \leq x, Y \leq y)$: **joint cdf** of X and Y
- $f_{XY}(x,y) = \frac{\partial}{\partial x}\frac{\partial}{\partial y} F_{XY}(x,y)$: **joint pdf** of X and Y
- $f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)\, dy$: **marginal pdf**
- $f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x,y)\, dx$: **marginal pdf**
- $f_{X|Y}(x|y) = \dfrac{f_{XY}(x,y)}{f_Y(y)}$: **conditional pdf** of X given Y
- $f_{Y|X}(y|x) = \dfrac{f_{XY}(x,y)}{f_X(x)}$: **conditional pdf** of Y given X
- These definitions also generalize to more than two random variables.

Two random variables X and Y are (statistically) **independent** if $F_{XY}(x,y) = F_X(x)F_Y(y)$. This implies $f_{X|Y}(x|y) = f_X(x)$ and $f_{Y|X}(y|x) = f_Y(y)$.

Mixed discrete and continuous random variables

Example. Consider a random variable X that takes on $x = -1$ and $x = 1$ with probability $1/4$ each, and values in the interval $x \in (-1/2, 1/2)$ with probability $1/2$. Each value in the interval is equally likely.





The pdf is the **generalized derivative** of the cdf:

$$f_X(x) = \tfrac{1}{4}\delta(x-1) + \tfrac{1}{4}\delta(x+1) + \tfrac{1}{2}\mathrm{rect}(x)$$

where $\delta(x)$ is the **Dirac delta-impulse**:

$$\int_{-\infty}^{\infty} \delta(x - x_0) g(x)\, dx = g(x_0)$$

Expectations

Expectation operator

$$E\{g(X)\} \triangleq \int_{-\infty}^{\infty} g(x) f_X(x)\, dx$$

- For $g(x) = x$, we obtain the **mean** (first-order moment):

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

- For $g(x) = x^2$, we obtain the **second-order moment**:

$$\mu_{X^2} = E\{X^2\} = \int_{-\infty}^{\infty} x^2 f_X(x)\, dx$$

- For $g(x) = (x - \mu_X)^2$, we obtain the **variance** (central second-order moment):

$$\mathrm{var}(X) = \sigma_X^2 = E\{(X - \mu_X)^2\} = \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x)\, dx$$

Joint and conditional expectations

For two continuous random variables X and Y the expectation is defined as

$$E\{g(X,Y)\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f_{XY}(x,y)\,dx\,dy$$

- For $g(x,y) = xy$ we obtain the **correlation**

$$\mu_{XY} = E\{XY\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy f_{XY}(x,y)\,dx\,dy$$

- For $g(x,y) = (x-\mu_X)(y-\mu_y)$ we obtain the **covariance**

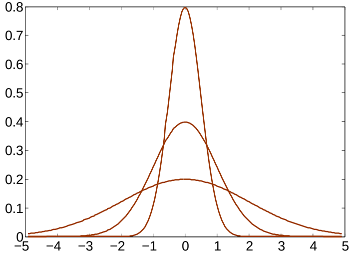
$$\sigma_{XY} = E\{(X-\mu_X)(Y-\mu_Y)\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-\mu_X)(y-\mu_Y) f_{XY}(x,y)\,dx\,dy$$

The **conditional expectation** $E\{X|Y\}$ is itself a random variable. The remarks made for discrete RVs also apply to continuous RVs.

2.4 The Gaussian distribution

Gaussian pdf

This plot shows the Gaussian pdf for $\mu_X = 0$ and $\sigma_X = \{0.5, 1, 2\}$.



Gaussian pdf $X \sim N(\mu_X, \sigma_X^2)$

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma_X} \exp\left\{ -\frac{1}{2} \left(\frac{x - \mu_X}{\sigma_X} \right)^2 \right\}$$

with mean μ_X and standard deviation σ_X

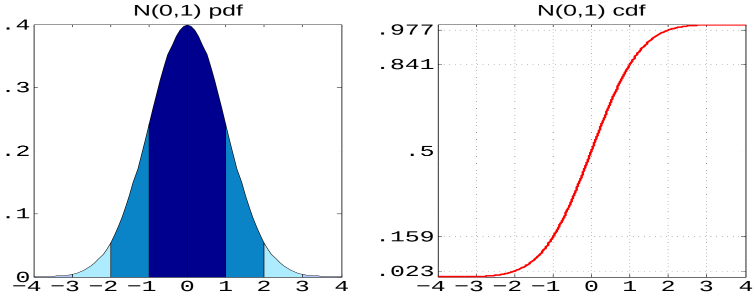
The Gaussian pdf is completely determined by its mean and variance.

If $X \sim N(\mu_X, \sigma_X^2)$, then

$$F_X(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma_X} \exp\left[-\frac{1}{2}\left(\frac{\xi - \mu_X}{\sigma_X}\right)^2\right] d\xi$$

There is no closed-form expression for this integral. The standard normal cdf for $N(0,1)$ is

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-s^2/2} ds$$



Gaussian cdf (2)

For arbitrary mean and variance, the Gaussian cdf can be expressed as

$$F_X(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{(x-\mu_X)/\sigma_X} e^{-s^2/2} ds = \Phi\left(\frac{x-\mu_X}{\sigma_X}\right)$$

There are the following closely related functions:

- The **error function**

$$\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-s^2} ds$$

- The **complementary error function**

$$\mathrm{erfc}(z) = 1 - \mathrm{erf}(z) = \frac{2}{\sqrt{\pi}} \int_z^\infty e^{-s^2} ds = 2\Phi(-\sqrt{2}z)$$

- The **Q-function**

$$Q(z) = 1 - \Phi(z) = \frac{1}{2}\mathrm{erfc}\left(\frac{z}{\sqrt{2}}\right)$$

Central limit theorem

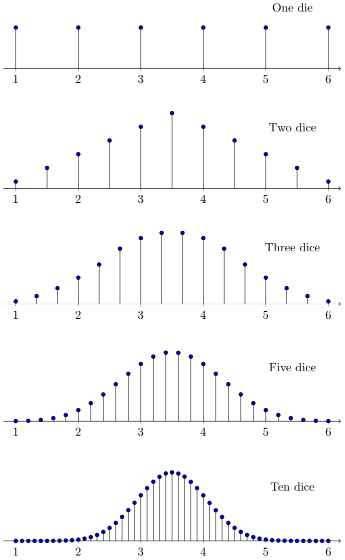
Central limit theorem (CLT)

Let $X_1, X_2, ..., X_N$ be **independent and identically distributed (i.i.d.)** random variables with mean μ and variance σ^2. The random variable

$$Z = \frac{1}{\sqrt{N}} \sum_{n=1}^{N} (X_n - \mu)$$

has mean $\mu_Z = 0$ and variance $\sigma_Z^2 = \sigma^2$. As $N \to \infty$, Z approaches a **Gaussian** distributed random variable.

The random variables X_n can have arbitrary distributions. They may even be discrete random variables, as illustrated on the right.



Bivariate Gaussian density

Let X and Y be **jointly Gaussian**. Then their joint pdf is

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

- We need to specify five parameters: σ_X, σ_Y, μ_X, μ_Y, and ρ_{XY}.
- The **marginals** of $f_{XY}(x,y)$ are also **Gaussian**. The pdf of X is

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)\,dy = \frac{1}{\sqrt{2\pi}\sigma_X}\exp\left\{-\frac{1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2\right\}$$

  and the pdf of Y is analogous.
- We also see that if X and Y are **uncorrelated**, i.e., $\rho_{XY} = 0$, then $f_{XY}(x,y) = f_X(x)f_Y(y)$, i.e., X and Y are **independent**.

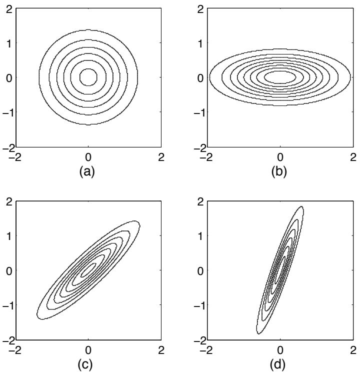
Level contours of the bivariate Gaussian

Contours of constant pdf:
Let's assume $\mu_X = \mu_Y = 0$.

- If $\sigma_X = \sigma_Y$ and $\rho_{XY} = 0$, we have **circular contours** (plot a).
- If $\rho_{XY} = 0$, but $\sigma_X \neq \sigma_Y$, then the contours are of the form

$$(\frac{x}{\sigma_X})^2 + (\frac{y}{\sigma_Y})^2 = c^2.$$

  This an **ellipse** whose major axis is aligned with the x-axis for $\sigma_X > \sigma_Y$ (plot b) or the y-axis for $\sigma_X < \sigma_Y$.



- If $\sigma_X = \sigma_Y = \sigma$, then the contours are of the form $x^2 - 2\rho_{XY}xy + y^2 = c^2\sigma^2$. This is a 45° **tilted ellipse** for $\rho_{XY} > 0$ (plot c) and a 135° **tilted ellipse** for $\rho_{XY} < 0$.
- In general, for $\rho_{XY} \neq 0$ and $\sigma_X \neq \sigma_Y$, the ellipse can have an **arbitrary orientation** (plot d).

Summary of important properties

The Gaussian distribution has a number of important properties, which account for its widespread use:

- The Gaussian pdf is completely determined by its mean and variance.
- If X and Y are jointly Gaussian and uncorrelated, they are independent.
- If X and Y are jointly Gaussian, then the marginals of X and Y are Gaussian as well. (The converse does not hold.)
- If X and Y are jointly Gaussian, then the conditional distribution of X given Y (or vice versa) is Gaussian as well.
- If X is Gaussian, then $aX + b$ is Gaussian with mean $a\mu_X + b$ and variance $a^2\sigma_X^2$.
- If X_n, $n = 1, \ldots, N$, are independent Gaussians with means μ_n and variances σ_n^2, then $Z = \sum_{n=1}^{N} X_n$ is Gaussian with mean $\mu_Z = \sum_{n=1}^{N} \mu_n$ and variance $\sigma_Z^2 = \sum_{n=1}^{N} \sigma_n^2$.
- Central limit theorem (see corresponding slide): The sum of a large number of i.i.d. RVs tends toward a Gaussian.

2.5 Random vectors

Joint distributions and densities

Let's arrange n random variables $X_1, X_2, ..., X_n$ in a random vector $\mathbf{X} = [X_1, ..., X_n]^T$. The cdf of $\mathbf{X}$ is interpreted as the joint cdf of its components:

$$F_{\mathrm{X}}(\mathbf{x}) = P(X_1 \leq x_1, ..., X_n \leq x_n)$$

The pdf of $\mathbf{X}$ is the joint pdf of its components:

$$f_{\mathrm{X}}(\mathbf{x}) = \frac{\partial}{\partial x_1} \cdots \frac{\partial}{\partial x_n} F_{\mathrm{X}}(\mathbf{x})$$

The joint cdf of two random vectors $\mathbf{X} = [X_1, ..., X_n]^T$ and $\mathbf{Y} = [Y_1, ..., Y_m]^T$ is

$$F_{\mathrm{XY}}(\mathbf{x}, \mathbf{y}) = P(X_1 \leq x_1, ..., X_n \leq x_n, Y_1 \leq y_1, ..., Y_m \leq y_m)$$

The joint pdf of $\mathbf{X}$ and $\mathbf{Y}$ is

$$f_{\mathrm{XY}}(\mathbf{x}, \mathbf{y}) = \frac{\partial}{\partial x_1} \cdots \frac{\partial}{\partial x_n} \frac{\partial}{\partial y_1} \cdots \frac{\partial}{\partial y_m} F_{\mathrm{XY}}(\mathbf{x}, \mathbf{y})$$

Mean vectors and covariance matrices

If the expectation operator $E\{\cdot\}$ is applied to a vector or matrix, it is applied to each component individually.

Mean vector and covariance matrix

$$\boldsymbol{\mu}_X = E\{\mathbf{X}\}$$
$$\mathbf{R}_{XX} = E\{(\mathbf{X} - \boldsymbol{\mu}_X)(\mathbf{X} - \boldsymbol{\mu}_X)^T\} = E\{\mathbf{X}\mathbf{X}^T\} - \boldsymbol{\mu}_X\boldsymbol{\mu}_X^T$$

The (i,j)th element of $\mathbf{R}_{XX}$ is the covariance between X_i and X_j:
$(\mathbf{R}_{XX})_{ij} = E\{(X_i - \mu_{X_i})(X_j - \mu_{X_j})\} = \sigma_{X_iX_j} = \sigma_{X_jX_i}$
The covariance matrix can be written as

$$\mathbf{R}_{XX} = \begin{bmatrix} \sigma^2_{X_1} & \sigma_{X_1X_2} & \cdots & \sigma_{X_1X_n} \\ \sigma_{X_1X_2} & \sigma^2_{X_2} & \cdots & \vdots \\ \vdots & & \ddots & \vdots \\ \sigma_{X_1X_n} & \cdots & \cdots & \sigma^2_{X_n} \end{bmatrix}$$

This matrix is **symmetric**: $\mathbf{R}_{XX} = \mathbf{R}_{XX}^T$.

More on covariance matrices

- The matrix $E\{\mathbf{X}\mathbf{X}^T\}$ is called the **correlation matrix**.
- We can also define the **cross-covariance matrix** $\mathbf{R}_{XY} = E\{(\mathbf{X} - \boldsymbol{\mu}_X)(\mathbf{Y} - \boldsymbol{\mu}_Y)^T\}$. This matrix is not symmetric.
- We call two n-dimensional vectors **uncorrelated** if $E\{\mathbf{X}\mathbf{Y}^T\} = E\{\mathbf{X}\}E\{\mathbf{Y}^T\}$. Uncorrelatedness does not imply independence!
- We call $\mathbf{X}$ and $\mathbf{Y}$ **orthogonal** if $E\{\mathbf{X}\mathbf{Y}^T\} = \mathbf{0}$.

Covariance/correlation matrices (but not cross-covariance matrices) have the following important property:

> Covariance matrices are symmetric and **positive semidefinite**:
>
> $$\mathbf{z}^T\mathbf{R}_{XX}\mathbf{z} \geq 0 \text{ for all } \mathbf{z} \neq \mathbf{0}$$

Proof:

$$0 \leq E\{[\mathbf{z}^T(\mathbf{X} - \boldsymbol{\mu}_X)]^2\} = \mathbf{z}^T E\{(\mathbf{X} - \boldsymbol{\mu}_X)(\mathbf{X} - \boldsymbol{\mu}_X)^T\}\mathbf{z} = \mathbf{z}^T\mathbf{R}_{XX}\mathbf{z}$$

Multivariate Gaussian

Probability density

The n-dimensional Gaussian pdf is

$$f_{\mathrm{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}\sqrt{\det \mathbf{R}_{\mathrm{XX}}}} \exp\{-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_{\mathrm{X}})^T \mathbf{R}_{\mathrm{XX}}^{-1}(\mathbf{x}-\boldsymbol{\mu}_{\mathrm{X}})\}$$

This requires an invertible covariance matrix $\mathbf{R}_{\mathrm{XX}}$.
Exercise: Can you derive the form of the bivariate Gaussian used before from this general pdf?

Behavior under linear transformation

Let $\mathbf{A}$ be an $m \times n$ matrix of rank m. Then the random vector generated by $\mathbf{Y} = \mathbf{AX}$ has an m-dimensional Gaussian pdf with mean vector $\boldsymbol{\mu}_{\mathrm{Y}} = \mathbf{A}\boldsymbol{\mu}_{\mathrm{X}}$ and covariance matrix $\mathbf{R}_{\mathrm{YY}} = \mathbf{A}\mathbf{R}_{\mathrm{XX}}\mathbf{A}^T$.

Gaussian RVs are among the very few that do not change their type of distribution under linear transformation.