Probability and Random Processes

Serik Sagitov, Chalmers University of Technology and Gothenburg University

Abstract

Lecture notes based on the book Probability and Random Processes by Geoffrey Grimmett and David Stirzaker. Last updated June 3, 2014.

Contents

1 Random events and random variables

1.1 Probability space

A random experiment is modeled in terms of a *probability space* $(\Omega, \mathcal{F}, \mathbb{P})$

- the *sample space* Ω is the set of all possible outcomes of the experiment,

- the *σ-field* (or sigma-algebra) $\mathcal{F}$ is a collection of measurable subsets $A \subset \Omega$ (which are called *random events*) satisfying

  1. $\emptyset \in \mathcal{F}$,

  2. if $A_i \in \mathcal{F}$, $0 = 1, 2, \ldots$, then $\cup_{i=1}^{\infty} A_i \in \mathcal{F}$, countable unions,

  3. if $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$, *complementary event*,

- the probability measure $\mathbb{P}$ is a function on $\mathcal{F}$ satisfying three probability axioms

1. if $A \in \mathcal{F}$, then $\mathbb{P}(A) \geq 0$,
2. $\mathbb{P}(\Omega) = 1$,
3. if $A_i \in \mathcal{F}$, $0 = 1, 2, \ldots$ are all disjoint, then $\mathbb{P}(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$.

De Morgan's laws

$$\left(\bigcap_i A_i \right)^c = \bigcup_i A_i^c, \qquad \left(\bigcup_i A_i \right)^c = \bigcap_i A_i^c.$$

Properties derived from the axioms

$$\mathbb{P}(\emptyset) = 0,$$
$$\mathbb{P}(A^c) = 1 - \mathbb{P}(A),$$
$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$$

Inclusion-exclusion rule

$$\mathbb{P}(A_1 \cup \ldots \cup A_n) = \sum_i \mathbb{P}(A_i) - \sum_{i<j} \mathbb{P}(A_i \cap A_j) + \sum_{i<j<k} \mathbb{P}(A_i \cap A_j \cap A_k) - \ldots$$
$$+ (-1)^{n+1} \mathbb{P}(A_1 \cap \ldots \cap A_n).$$

Continuity of the probability measure

- if $A_1 \subset A_2 \subset \ldots$ and $A = \cup_{i=1}^{\infty} A_i = \lim_{i \to \infty} A_i$, then $\mathbb{P}(A) = \lim_{i \to \infty} \mathbb{P}(A_i)$,

- if $B_1 \supset B_2 \supset \ldots$ and $B = \cap_{i=1}^{\infty} B_i = \lim_{i \to \infty} B_i$, then $\mathbb{P}(B) = \lim_{i \to \infty} \mathbb{P}(B_i)$.

1.2 Conditional probability and independence

If $\mathbb{P}(B) > 0$, then the conditional probability of A given B is

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

The law of total probability and the Bayes formula. Let $B_1, \ldots, B_n$ be a partition of Ω, then

$$\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(A|B_i)\mathbb{P}(B_i),$$
$$\mathbb{P}(B_j|A) = \frac{\mathbb{P}(A|B_j)\mathbb{P}(B_j)}{\sum_{i=1}^{n} \mathbb{P}(A|B_i)\mathbb{P}(B_i)}.$$

Definition 1.1 Events $A_1, \ldots, A_n$ are independent, if for any subset of events $(A_{i_1}, \ldots, A_{i_k})$

$$\mathbb{P}(A_{i_1} \cap \ldots \cap A_{i_k}) = \mathbb{P}(A_{i_1}) \ldots \mathbb{P}(A_{i_k}).$$

Example 1.2 Pairwise independence does not imply independence of three events. Toss two coins and consider three events

- $A = \{\text{heads on the first coin}\}$,

- $B = \{\text{tails on the first coin}\}$,

- $C = \{\text{one head and one tail}\}$.

Clearly, $\mathbb{P}(A|C) = \mathbb{P}(A)$ and $\mathbb{P}(B|C) = \mathbb{P}(B)$ but $\mathbb{P}(A \cap B|C) = 0$.

1.3 Random variables

A real random variable is a measurable function $X : \Omega \to \mathbb{R}$ so that different outcomes $\omega \in \Omega$ can give different values $X(\omega)$. Measurability of $X(\omega)$:

$$\{\omega : X(\omega) \leq x\} \in \mathcal{F} \text{ for any real number } x.$$

Probability distribution $\mathbb{P}_X(B) = \mathbb{P}(X \in B)$ defines a new probability space $(R, \mathcal{B}, \mathbb{P}_X)$, where $\mathcal{B} = \sigma(\text{all open intervals})$ is the Borel sigma-algebra.

Definition 1.3 Distribution function (cumulative distribution function)

$$F(x) = F_X(x) = \mathbb{P}_X\{(-\infty, x]\} = \mathbb{P}(X \leq x).$$

In terms of the distribution function we get

$$\mathbb{P}(a < X \leq b) = F(b) - F(a),$$
$$\mathbb{P}(X < x) = F(x-),$$
$$\mathbb{P}(X = x) = F(x) - F(x-).$$

Any monotone right-continuous function with

$$\lim_{x \to -\infty} F(x) = 0 \text{ and } \lim_{x \to \infty} F(x) = 1$$

can be a distribution function.

Definition 1.4 The random variable X is called discrete, if for some countable set of possible values

$$\mathbb{P}(X \in \{x_1, x_2, \ldots\}) = 1.$$

Its distribution is described by the probability mass function $f(x) = \mathbb{P}(X = x)$.

The random variable X is called (absolutely) continuous, if its distribution has a probability density function $f(x)$:

$$F(x) = \int_{-\infty}^{x} f(y)dy, \quad \text{for all } x,$$

so that $f(x) = F'(x)$ almost everywhere.

Example 1.5 The indicator of a random event $1_A = 1_{\{\omega \in A\}}$ with $p = \mathbb{P}(A)$ has a Bernoulli distribution

$$\mathbb{P}(1_A = 1) = p, \quad \mathbb{P}(1_A = 0) = 1 - p.$$

For several events $S_n = \sum_{i=1}^{n} 1_{A_i}$ counts the number of events that occurred. If independent events $A_1, A_2, \ldots$ have the same probability $p = \mathbb{P}(A_i)$, then S_n has a binomial distribution $\text{Bin}(n, p)$

$$\mathbb{P}(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n.$$

Example 1.6 (Cantor distribution) Consider $(\Omega, \mathcal{F}, \mathbb{P})$ with $\Omega = [0, 1]$, $\mathcal{F} = \mathcal{B}_{[0,1]}$, and

$$\mathbb{P}([0, 1]) = 1$$
$$\mathbb{P}([0, 1/3]) = \mathbb{P}([2/3, 1]) = 2^{-1}$$
$$\mathbb{P}([0, 1/9]) = \mathbb{P}([2/9, 1/3]) = \mathbb{P}([2/3, 7/9]) = \mathbb{P}([8/9, 1]) = 2^{-2}$$

and so on. Put $X(\omega) = \omega$, its distribution, called the Cantor distribution, is neither discrete nor continuous. Its distribution function, called the Cantor function, is continuous but not absolutely continuous.

1.4   Random vectors

Definition 1.7 The joint distribution of a random vector $\mathbf{X} = (X_1, \ldots, X_n)$ is the function

$$F_{\mathbf{X}}(x_1, \ldots, x_n) = \mathbb{P}(\{X_1 \leq x_1\} \cap \ldots \cap \{X_n \leq x_n\}).$$

Marginal distributions

$$
\begin{aligned}
F_{X_1}(x) &= F_{\mathbf{X}}(x, \infty, \ldots, \infty), \\
F_{X_2}(x) &= F_{\mathbf{X}}(\infty, x, \infty, \ldots, \infty), \\
&\cdots \\
F_{X_n}(x) &= F_{\mathbf{X}}(\infty, \ldots, \infty, x).
\end{aligned}
$$

The existence of the joint probability density function $f(x_1, \ldots, x_n)$ means that the distribution function

$$F_{\mathbf{X}}(x_1, \ldots, x_n) = \int_{-\infty}^{x_1} \ldots \int_{-\infty}^{x_n} f(y_1, \ldots, y_n) dy_1 \ldots dy_n, \quad \text{for all } (x_1, \ldots, x_n),$$

is absolutely continuous, so that $f(x_1, \ldots, x_n) = \frac{\partial^n F(x_1, \ldots, x_n)}{\partial x_1 \ldots \partial x_n}$ almost everywhere.

Definition 1.8 Random variables $(X_1, \ldots, X_n)$ are called independent if for any $(x_1, \ldots, x_n)$

$$\mathbb{P}(X_1 \leq x_1, \ldots, X_n \leq x_n) = \mathbb{P}(X_1 \leq x_1) \ldots \mathbb{P}(X_n \leq x_n).$$

In the jointly continuous case this equivalent to

$$f(x_1, \ldots, x_n) = f_{X_1}(x_1) \ldots f_{X_n}(x_n).$$

Example 1.9 In general, the joint distribution can not be recovered form the marginal distributions. If

$$F_{X,Y}(x, y) = xy \mathbf{1}_{\{(x,y) \in [0,1]^2\}},$$

then vectors (X, Y) and (X, X) have the same marginal distributions.

Example 1.10 Consider

$$
F(x, y) = \begin{cases}
1 - e^{-x} - xe^{-y} & \text{if } 0 \leq x \leq y, \\
1 - e^{-y} - ye^{-y} & \text{if } 0 \leq y \leq x, \\
0 & \text{otherwise.}
\end{cases}
$$

Show that $F(x, y)$ is the joint distribution function of some pair (X, Y). Find the marginal distribution functions and densities.

Solution. Three properties should be satisfied for $F(x, y)$ to be the joint distribution function of some pair (X, Y):

1. $F(x, y)$ is non-decreasing on both variables,

2. $F(x, y) \to 0$ as $x \to -\infty$ and $y \to -\infty$,

3. $F(x, y) \to 1$ as $x \to \infty$ and $y \to \infty$.

Observe that

$$f(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y} = e^{-y} \mathbf{1}_{\{0 \leq x \leq y\}}$$

is always non-negative. Thus the first property follows from the integral representation:

$$F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(u, v) du dv,$$

which, for $0 \leq x \leq y$, is verifies as

$$\int_{-\infty}^{x} \int_{-\infty}^{y} f(u, v) du dv = \int_{0}^{x} \left(\int_{u}^{y} e^{-v} dv \right) du = 1 - e^{-x} - xe^{-y},$$
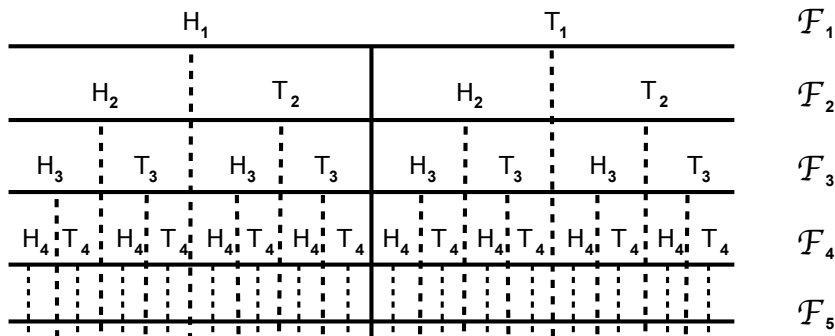
Figure 1: Filtration for four consecutive coin tossings.

and for $0 \leq y \leq x$ as

$$\int_{-\infty}^{x} \int_{-\infty}^{y} f(u,v)dudv = \int_{0}^{y} \Big(\int_{u}^{y} e^{-v}dv \Big) du = 1 - e^{-y} - ye^{-y}.$$

The second and third properties are straightforward. We have shown also that $f(x,y)$ is the joint density.

For $x \geq 0$ and $y \geq 0$ we obtain the marginal distributions as limits

$$F_X(x) = \lim_{y \to \infty} F(x,y) = 1 - e^{-x}, \qquad f_X(x) = e^{-x},$$

$$F_Y(y) = \lim_{x \to \infty} F(x,y) = 1 - e^{-y} - ye^{-y}, \qquad f_Y(y) = ye^{-y}.$$

$X \sim \mathrm{Exp}(1)$ and $Y \sim \mathrm{Gamma}(2,1)$.

1.5 Filtration

Definition 1.11 A sequence of sigma-fields $\{\mathcal{F}_n\}_{n=1}^{\infty}$ such that

$$\mathcal{F}_1 \subset \mathcal{F}_2 \subset \ldots \subset \mathcal{F}_n \subset \ldots, \quad \mathcal{F}_n \subset \mathcal{F} \text{ for all } n$$

is called a filtration.

To illustrate this definition use an infinite sequence of Bernoulli trials. Let S_n be the number of heads in n independent tosses of a fair coin. Figure 1 shows imbedded partitions $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \mathcal{F}_3 \subset \mathcal{F}_4 \subset \mathcal{F}_5$ of the sample space generated by S_1, S_2, S_3, S_4, S_5.

The events representing our knowledge of the first three tosses is given by $\mathcal{F}_3$. From the perspective of $\mathcal{F}_3$ we can not say exactly the value of S_4. Clearly, there is dependence between S_3 and S_4. The joint distribution of S_3 and S_4:

	$S_4 = 0$	$S_4 = 1$	$S_4 = 2$	$S_4 = 3$	$S_4 = 4$	Total
$S_3 = 0$	1/16	1/16	0	0	0	1/8
$S_3 = 1$	0	3/16	3/16	0	0	3/8
$S_3 = 2$	0	0	3/16	3/16	0	3/8
$S_3 = 3$	0	0	0	1/16	1/16	1/8
Total	1/16	1/4	3/8	1/4	1/16	1

The conditional expectation

$$\mathbb{E}(S_4|S_3) = S_3 + 0.5$$

is a discrete random variable with values $0.5, 1.5, 2.5, 3.5$ and probabilities $1/8, 3/8, 3/8, 1/8$.

For finite n the picture is straightforward. For $n = \infty$ it is a non-trivial task to define an overall $(\Omega, \mathcal{F}, \mathbb{P})$ with $\Omega = (0,1]$. One can use the Lebesgue measure $\mathbb{P}(dx) = dx$ and the sigma-field $\mathcal{F}$ of Lebesgue measurable subsets of $(0,1]$. Not all subsets of $(0,1]$ are Lebesgue measurable.

2 Expectation and conditional expectation

2.1 Expectation

The expected value of X is

$$\mathbb{E}(X) = \int_\Omega X(\omega)\mathbb{P}(d\omega).$$

A discrete r.v. X with a finite number of possible values is a simple r.v. in that

$$X = \sum_{i=1}^{n} x_i 1_{A_i}$$

for some partition $A_1, \ldots, A_n$ of Ω. In this case the meaning of the expectation is obvious

$$\mathbb{E}(X) = \sum_{i=1}^{n} x_i \mathbb{P}(A_i).$$

For any non-negative r.v. X there are simple r.v. such that $X_n(\omega) \nearrow X(\omega)$ for all $\omega \in \Omega$, and the expectation is defined as a possibly infinite limit $\mathbb{E}(X) = \lim_{n\to\infty} \mathbb{E}(X_n)$.

Any r.v. X can be written as a difference of two non-negative r.v. $X^+ = X \vee 0$ and $X^- = -X \wedge 0$. If at least one of $\mathbb{E}(X^+)$ and $\mathbb{E}(X^-)$ is finite, then $\mathbb{E}(X) = \mathbb{E}(X^+) - \mathbb{E}(X^-)$, otherwise $\mathbb{E}(X)$ does not exist.

Example 2.1 A discrete r.v. with the probability mass function $f(k) = \frac{1}{2k(k-1)}$ for $k = -1, \pm 2, \pm 3, \ldots$ has no expectation.

For a discrete r.v. X with mass function f and any function g

$$\mathbb{E}(g(X)) = \sum_x g(x)f(x).$$

For a continuous r.v. X with density f and any measurable function g

$$\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)dx.$$

In general

$$\mathbb{E}(X) = \int_\Omega X(\omega)\mathbb{P}(d\omega) = \int_{-\infty}^{\infty} x\mathbb{P}_X(dx) = \int_{-\infty}^{\infty} x dF(x).$$

Example 2.2 Turn to the example of $(\Omega, \mathcal{F}, \mathbb{P})$ with $\Omega = [0,1]$, $\mathcal{F} = \mathcal{B}_{[0,1]}$, and a random variable $X(\omega) = \omega$ having the Cantor distribution. A sequence of simple r.v. monotonely converging to X is $X_n(\omega) = k3^{-n}$ for $\omega \in [(k-1)3^{-n}, k3^{-n})$, $k = 1, \ldots, 3^n$ and $X_n(1) = 1$.

$$X_1(\omega) = 0, \quad \mathbb{E}(X_1) = 0,$$
$$X_2(\omega) = (1/3)1_{\{\omega \in [1/3, 2/3)\}} + (2/3)1_{\{\omega \in [2/3, 1]\}}, \quad \mathbb{E}(X_2) = (2/3) * (1/2) = 1/3,$$
$$\mathbb{E}(X_3) = (2/9) * (1/4) + (2/3) * (1/4) + (8/9) * (1/4) = 4/9,$$

and so on, gives $\mathbb{E}(X_n) \nearrow 1/2 = \mathbb{E}(X)$.

Lemma 2.3 *Cauchy-Schwartz inequality. For r.v. X and Y we have*

$$\left(\mathbb{E}(XY)\right)^2 \leq \mathbb{E}(X^2)\mathbb{E}(Y^2)$$

with equality if only if $aX + bY = 1$ a.s. for some non-trivial pair of constants (a, b).

Definition 2.4 Variance, standard deviation, covariance and correlation

$$\mathbb{V}ar(X) = \mathbb{E}(X - \mathbb{E}X)^2 = \mathbb{E}(X^2) - (\mathbb{E}X)^2, \quad \sigma_X = \sqrt{\mathbb{V}ar(X)},$$
$$\mathbb{C}ov(X, Y) = \mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y) = \mathbb{E}(XY) - (\mathbb{E}X)(\mathbb{E}Y),$$
$$\rho(X, Y) = \frac{\mathbb{C}ov(X, Y)}{\sigma_X \sigma_Y}.$$

The covariance matrix of a random vector $(X_1, \ldots, X_n)$ with means $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$

$$\mathbf{V} = \mathbb{E}\big(\mathbf{X} - \boldsymbol{\mu}\big)^{\mathrm{t}}\big(\mathbf{X} - \boldsymbol{\mu}\big) = \|\mathbb{C}ov(X_i, X_j)\|$$

is symmetric and nonnegative-definite. For any vector $\mathbf{a} = (a_1, \ldots, a_n)$ the r.v. $a_1 X_1 + \ldots + a_n X_n$ has mean $\mathbf{a}\boldsymbol{\mu}^{\mathrm{t}}$ and variance

$$\mathbb{V}ar(a_1 X_1 + \ldots + a_n X_n) = \mathbb{E}\big(\mathbf{a}\mathbf{X}^{\mathrm{t}} - \mathbf{a}\boldsymbol{\mu}^{\mathrm{t}}\big)\big(\mathbf{X}\mathbf{a}^{\mathrm{t}} - \boldsymbol{\mu}\mathbf{a}^{\mathrm{t}}\big) = \mathbf{a}\mathbf{V}\mathbf{a}^{\mathrm{t}}.$$

If $(X_1, \ldots, X_n)$ are independent, then they are uncorrelated: $\mathbb{C}ov(X_i, X_j) = 0$.

2.2 Conditional expectation and prediction

Definition 2.5 For a pair of discrete random variables (X, Y) the conditional expectation $\mathbb{E}(Y|X)$ is defined as $\psi(X)$, where

$$\psi(x) = \sum_y y \mathbb{P}(Y = y | X = x).$$

Definition 2.6 Consider a pair of random variables (X, Y) with joint density $f(x, y)$, marginal densities

$$f_1(x) = \int_{-\infty}^{\infty} f(x, y) dy, \quad f_2(x) = \int_{-\infty}^{\infty} f(x, y) dx,$$

and conditional densities

$$f_1(x|y) = \frac{f(x, y)}{f_2(y)}, \qquad f_2(y|x) = \frac{f(x, y)}{f_1(x)}.$$

The conditional expectation $\mathbb{E}(Y|X)$ is defined as $\psi(X)$, where

$$\psi(x) = \int_{-\infty}^{\infty} y f_2(y|x) dy.$$

Properties of conditional expectations:
 (i) linearity: $\mathbb{E}(aY + bZ|X) = a\mathbb{E}(Y|X) + b\mathbb{E}(Z|X)$ for any constants (a, b),
 (ii) pull-through property: $\mathbb{E}(Yg(X)|X) = g(X)\mathbb{E}(Y|X)$ for any measurable function $g(x)$,
 (iii) $\mathbb{E}(Y\mathbb{1}_G) = \mathbb{E}(\psi(X)\mathbb{1}_G)$ for $G = \{\omega : X(\omega) \in B\}$, where $B \in \mathcal{R}$,
 (iv) tower property: $\mathbb{E}(\mathbb{E}(Y|X, Z)|X) = \mathbb{E}(Y|X)$,
 (v) total expectation: $\mathbb{E}(\mathbb{E}(Y|X)) = \mathbb{E}(Y)$,
 (vi) total variance: $\mathbb{V}ar(Y) = \mathbb{V}ar(\mathbb{E}(Y|X)) + \mathbb{E}(\mathbb{V}ar(Y|X))$.

Proof of (ii) in the discrete case:

$$\mathbb{E}(Yg(X)|X) = \sum_{x,y} yg(x)\mathbb{P}(Y = y, X = x) = \sum_{x,y} g(x)\mathbb{P}(X = x)y\mathbb{P}(Y = y|X = x)$$

$$= \sum_x g(x)\psi(x)\mathbb{P}(X = x) = g(X)\mathbb{E}(Y|X).$$

Definition 2.7 General definition. Let Y be a r.v. on $(\Omega, \mathcal{F}, \mathbb{P})$ and let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$. If there exists a $\mathcal{G}$-measurable r.v. Z such that

$$\mathbb{E}((Y - Z)\mathbb{1}_G) = 0 \text{ for all } G \in \mathcal{G},$$

then Z is called the conditional expectation of Y given $\mathcal{G}$ and is written $Z = \mathbb{E}(Y|\mathcal{G})$.

Properties of conditional expectations:
 (vii) if $\mathbb{E}(Y|\mathcal{G})$ exists, then it is a.s. unique,
 (viii) if $\mathbb{E}|Y| < \infty$, then $\mathbb{E}(Y|\mathcal{G})$ exists due to the Radon-Nikodym theorem,
 (ix) if $\mathcal{G} = \sigma(X)$, then $\mathbb{E}(Y|X) := \mathbb{E}(Y|\mathcal{G})$ and $\mathbb{E}(Y|X) = \psi(X)$ for some measurable function ψ.

Proof of (viii). Consider the probability space $(\Omega, \mathcal{G}, \mathbb{P})$ and define a finite signed measure $\mathbb{P}_1(G) = \mathbb{E}(Y\mathbb{1}_G) = \int_G Y(\omega)\mathbb{P}(d\omega)$ which is absolutely continuous with respect to $\mathbb{P}$. Thus $\mathbb{P}_1(G) = \int_G Z(\omega)\mathbb{P}(d\omega)$ with $Z = \partial\mathbb{P}_1/\partial\mathbb{P}$ being the Radon-Nikodym derivative.

Definition 2.8 Let X and Y be random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ such that $\mathbb{E}(Y^2) < \infty$. The best predictor of Y given the knowledge of X is the function $\hat{Y} = h(X)$ that minimizes $\mathbb{E}((Y - \hat{Y})^2)$.

Let $L^2(\Omega, \mathcal{F}, \mathbb{P})$ be the set of random variables Z on $(\Omega, \mathcal{F}, \mathbb{P})$ such that $\mathbb{E}(Z^2) < \infty$. Define a scalar product on the linear space $L^2(\Omega, \mathcal{F}, \mathbb{P})$ by $\langle U, V \rangle = \mathbb{E}(UV)$ leading to the norm

$$\|Z\| = \langle Z, Z \rangle^{1/2} = (\mathbb{E}(Z^2))^{1/2}.$$

Let H be the subspace of $L^2(\Omega, \mathcal{F}, \mathbb{P})$ of all functions of X having finite second moment

$$H = \{h(X) : \mathbb{E}(h(X)^2) < \infty\}.$$

Geometrically, the best predictor of Y given X is the projection $\hat{Y}$ of Y on H so that

$$\mathbb{E}((Y - \hat{Y})Z) = 0, \text{ for all } Z \in H. \tag{1}$$

Theorem 2.9 Let X and Y be random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ such that $\mathbb{E}(Y^2) < \infty$. The best predictor of Y given X is the conditional expectation $\hat{Y} = \mathbb{E}(Y|X)$.

Proof. Put $\hat{Y} = \mathbb{E}(Y|X)$. We have due to the Jensen inequality $\hat{Y}^2 \leq \mathbb{E}(Y^2|X)$ and therefore

$$\mathbb{E}(\hat{Y}^2) \leq \mathbb{E}(\mathbb{E}(Y^2|X)) = \mathbb{E}(Y^2) < \infty,$$

implying $\hat{Y} \in H$. To verify (1) we observe that

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

To prove uniqueness assume that there is another predictor $\bar{Y}$ with $\mathbb{E}((Y - \bar{Y})^2) = \mathbb{E}((Y - \hat{Y})^2) = d^2$. Then $\mathbb{E}((Y - \frac{\hat{Y}+\bar{Y}}{2})^2) \geq d^2$ and according to the parallelogram rule

$$2\left(\|Y - \hat{Y}\|^2 + \|Y - \bar{Y}\|^2\right) = 4\|Y - \frac{\hat{Y} + \bar{Y}}{2}\|^2 + \|\bar{Y} - \hat{Y}\|^2$$

we have

$$\|\bar{Y} - \hat{Y}\|^2 \leq 2\left(\|Y - \hat{Y}\|^2 + \|Y - \bar{Y}\|^2\right) - 4d^2 = 0.$$

2.3 Multinomial distribution

De Moivre trials: each trial has r possible outcomes with probabilities $(p_1, \ldots, p_r)$. Consider n such independent trials and let $(X_1, \ldots, X_r)$ be the counts of different outcomes. Multinomial distribution $\text{Mn}(n, p_1, \ldots, p_r)$

$$\mathbb{P}(X_1 = k_1, \ldots, X_r = k_r) = \frac{n!}{k_1! \ldots k_r!} p_1^{k_1} \ldots p_r^{k_r}.$$

Marginal distributions $X_i \sim \text{Bin}(n, p_i)$, also

$$(X_1 + X_2, X_3 \ldots, X_r) \sim \text{Mn}(n, p_1 + p_2, p_3, \ldots, p_r).$$

Conditionally on X_1

$$(X_2, \ldots, X_r) \sim \text{Mn}(n - X_1, \frac{p_2}{1 - p_1}, \ldots, \frac{p_r}{1 - p_1}),$$

so that $(X_i|X_j) \sim \text{Bin}(n - X_j, \frac{p_i}{1-p_j})$ and $\mathbb{E}(X_i|X_j) = (n - X_j)\frac{p_i}{1-p_j}$. It follows

$$\mathbb{E}(X_i X_j) = \mathbb{E}(\mathbb{E}(X_i X_j | X_j))$$
$$= \mathbb{E}(X_j \mathbb{E}(X_i | X_j)) = \mathbb{E}(nX_j - X_j^2)\frac{p_i}{1 - p_j}$$
$$= (n^2 p_j - np_j(1 - p_j) + n^2 p_j^2)\frac{p_i}{1 - p_j} = n(n-1)p_i p_j$$

and $\mathbb{C}ov(X_i, X_j) = -np_i p_j$ so that

$$\rho(X_i, X_j) = -\sqrt{\frac{p_i p_j}{(1 - p_i)(1 - p_j)}}.$$

2.4   Multivariate normal distribution

Bivariate normal distribution with parameters $(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$

$$f(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{ -\frac{(\frac{x-\mu_1}{\sigma_1})^2 - 2\rho(\frac{x-\mu_1}{\sigma_1})(\frac{y-\mu_2}{\sigma_2}) + (\frac{y-\mu_2}{\sigma_2})^2}{2(1-\rho^2)} \right\}.$$

Marginal distributions

$$f_1(x) = \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}, \quad f_2(y) = \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{(y-\mu_2)^2}{2\sigma_2^2}},$$

and conditional distributions

$$f_1(x|y) = \frac{f(x,y)}{f_2(y)} = \frac{1}{\sigma_1\sqrt{2\pi(1-\rho^2)}} \exp\left\{ -\frac{(x - \mu_1 - \frac{\rho\sigma_1}{\sigma_2}(y - \mu_2))^2}{2\sigma_1^2(1-\rho^2)} \right\},$$

$$f_2(y|x) = \frac{f(x,y)}{f_1(x)} = \frac{1}{\sigma_2\sqrt{2\pi(1-\rho^2)}} \exp\left\{ -\frac{(y - \mu_2 - \frac{\rho\sigma_2}{\sigma_1}(x - \mu_1))^2}{2\sigma_2^2(1-\rho^2)} \right\}.$$

Exercise: check the total variance formula for this example.

A multivariate normal distribution with mean vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$ and covariance matrix $\mathbf{V}$ has density

$$f(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n \det \mathbf{V}}} e^{-(\mathbf{x}-\boldsymbol{\mu})\mathbf{V}^{-1}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{t}}}.$$

For any vector $(a_1, \ldots, a_n)$ the r.v. $a_1 X_1 + \ldots + a_n X_n$ is normally distributed. Application in statistics: in the IID case: $\boldsymbol{\mu} = (\mu, \ldots, \mu)$ and $\mathbf{V} = \mathrm{diag}\{\sigma^2, \ldots, \sigma^2\}$ the sample mean and sample variance

$$\bar{X} = \frac{X_1 + \ldots + X_n}{n}, \quad s^2 = \frac{(X_1 - \bar{X})^2 + \ldots + (X_n - \bar{X})^2}{n-1}$$

are independent and $\frac{\sqrt{n}(\bar{X}-\mu)}{s}$ has a t-distribution with $n-1$ degrees of freedom.

If Y and Z are independent r.v. with standard normal distribution, their ratio $X = Y/Z$ has a Cauchy distribution with density

$$f(x) = \frac{1}{\pi(1+x^2)}, \quad -\infty < x < \infty.$$

In the Cauchy distribution case the mean is undefined and $\bar{X} \overset{d}{=} X$. Cauchy and normal distributions are examples of stable distributions. The Cauchy distribution provides with a counterexample for the law of large numbers.

2.5   Sampling from a distribution

Computers generate pseudo-random numbers $U_1, U_2, \ldots$ which we consider as IID r.v. with $\mathrm{U}_{[0,1]}$ distribution.

Inverse transform sampling: if F is a cdf and $U \sim \mathrm{U}_{[0,1]}$, then $X = F_{-1}(U)$ has cdf F. It follows from

$$\{F_{-1}(U) \le x\} = \{U \le F(x)\}.$$

Example 2.10 Examples of the inverse transform sampling.
   (i) Bernoulli distribution $X = 1_{\{U \le p\}}$,
   (ii) Binomial sampling: $S_n = X_1 + \ldots + X_n$, $X_k = 1_{\{U_k \le p\}}$,
   (iii) Exponential distribution $X = -\log(U)/\lambda$,
   (iv) Gamma sampling: $S_n = X_1 + \ldots + X_n$, $X_k = -\log(U_k)/\lambda$.

Lemma 2.11 *Rejection sampling. Suppose that we know how to sample from density $g(x)$ but we want to sample from density $f(x)$ such that $f(x) \leq ag(x)$ for some $a > 0$. Algorithm*
  step 1: sample x from $g(x)$ and u from $U_{[0,1]}$,
  step 2: if $u \leq \frac{f(x)}{ag(x)}$, accept x as a realization of sampling from $f(x)$,
  step 3: if not, reject the value of x and repeat the sampling step.

Proof. Let Z and U be independent, Z has density $g(x)$ and $U \sim \mathrm{U}_{[0,1]}$. Then

$$
\mathbb{P}\Big(Z \leq x \Big| U \leq \frac{f(Z)}{ag(Z)}\Big) = \frac{\int_{-\infty}^{x} \mathbb{P}\Big(U \leq \frac{f(y)}{ag(y)}\Big)g(y)dy}{\int_{-\infty}^{\infty} \mathbb{P}\Big(U \leq \frac{f(y)}{ag(y)}\Big)g(y)dy} = \int_{-\infty}^{x} f(y)dy.
$$

2.6   Probability generating, moment generating, and characteristic functions

Definition 2.12 If X takes values $k = 0, 1, 2, \ldots$ with probabilities p_k and $\sum_{k=0}^{\infty} p_k = 1$, then the distribution of X is fully described by its probability generating function

$$
G(s) = \mathbb{E}(s^X) = \sum_{k=0}^{\infty} p_k s^k.
$$

Properties of pgf:
  (i) $p_0 = G(0), \quad p_k = \frac{1}{k!}\frac{d^k G(s)}{ds^k}|_{s=0}$,
  (ii) $\mathbb{E}(X) = G'(1), \quad \mathbb{E}(X(X-1)) = G''(1)$.
  (iiii) if X and Y are independent, then $G_{X+Y}(s) = G_X(s)G_Y(s)$,

Example 2.13 Examples of probability generating functions
  (i) Bernoulli distribution $G(s) = q + ps$,
  (ii) Binomial distribution $G(s) = (q + ps)^n$,
  (iii) Geometric distribution $G(s) = \frac{1-p}{1-ps}$,
  (iv) Poisson distribution $G(s) = e^{\lambda(s-1)}$.

Definition 2.14 Moment generating function of X is $M(\theta) = \mathbb{E}(e^{\theta X})$. In the continuous case $M(\theta) = \int e^{\theta x} f(x)dx$. Moments $\mathbb{E}(X) = M'(0)$, $\mathbb{E}(X^k) = M^{(k)}(0)$.

Example 2.15 Examples of moment generating functions
  (i) Normal distribution $M(\theta) = e^{\theta\mu + \frac{1}{2}\theta^2\sigma^2}$,
  (ii) Exponential distribution $M(\theta) = \frac{\lambda}{\lambda-\theta}$ for $\theta < \lambda$,
  (ii) Gamma(α, λ) distribution has density $f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\lambda x}$ and $M(\theta) = \left(\frac{\lambda}{\lambda-\theta}\right)^\alpha$ for $\theta < \lambda$, it follows that the sum of k exponentials with parameter λ has a Gamma(k, λ) distribution,
  (iii) Cauchy distribution $M(0) = 1$, $M(t) = \infty$ for $t \neq 0$.

Definition 2.16 The characteristic function of X is complex valued $\phi(\theta) = \mathbb{E}(e^{i\theta X})$. The joint characteristic function for $\mathbf{X} = (X_1, \ldots, X_n)$ is $\phi(\boldsymbol{\theta}) = \mathbb{E}(e^{i\boldsymbol{\theta}\mathbf{X}^t})$.

Example 2.17 Examples of characteristic functions
  (i) Normal distribution $\phi(\theta) = e^{i\theta\mu - \frac{1}{2}\theta^2\sigma^2}$,
  (ii) Gamma distribution $\phi(\theta) = \left(\frac{\lambda}{\lambda-i\theta}\right)^\alpha$,
  (iii) Cauchy distribution $\phi(\theta) = e^{-|\theta|}$,
  (iv) Multinomial distribution $\phi(\boldsymbol{\theta}) = \left(\sum_{j=1}^{r} p_j e^{i\theta_j}\right)^n$.
  (v) Multivariate normal distribution $\phi(\boldsymbol{\theta}) = e^{i\boldsymbol{\theta}\boldsymbol{\mu}^t - \frac{1}{2}\boldsymbol{\theta}\mathbf{V}\boldsymbol{\theta}^t}$.

Example 2.18 Given a vector $\mathbf{X} = (X_1, \ldots, X_n)$ with a multivariate normal distribution any linear combination $\mathbf{aX}^t = a_1 X_1 + \ldots + a_n X_n$ is normally distributed since

$$
\mathbb{E}(e^{\theta\mathbf{aX}^t}) = \phi(\theta\mathbf{a}) = e^{i\theta\mu - \frac{1}{2}\theta^2\sigma^2}, \quad \mu = \mathbf{a}\boldsymbol{\mu}^t, \quad \sigma^2 = \mathbf{aVa}^t.
$$

3 Convergence of random variables

3.1 Borel-Cantelli lemmas

Given a sequence of random events $A_1, A_2, \ldots$ define new events

$$\sup_n A_n = \bigcup_n A_n, \qquad\qquad \inf_n A_n = \bigcap_n A_n,$$

$$\limsup_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} A_m, \qquad \liminf_{n\to\infty} A_n = \bigcup_{n=1}^{\infty} \bigcap_{m=n}^{\infty} A_m.$$

Observe that

$$\limsup_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} A_m = \{\forall n\ \exists m \geq n \text{ such that } A_m \text{ occurs}\} = \{\text{events } A_n \text{ occur infinitely often}\},$$

$$\liminf_{n\to\infty} A_n^c = \bigcup_{n=1}^{\infty} \bigcap_{m=n}^{\infty} A_m^c = \{\exists n \text{ such that } A_m^c \text{ occur } \forall m \geq n\} = \{\text{events } A_n \text{ occur finitely often}\}.$$

Theorem 3.1 *Borel-Cantelli lemmas. Let $\{A_n\ i.o.\} := \{$infinitely many of events $A_1, A_2, \ldots$ occur$\}$.*

 1. *If $\sum_{n=1}^{\infty} P(A_n) < \infty$, then $P(A_n\ i.o.) = 0$,*

 2. *If $\sum_{n=1}^{\infty} P(A_n) = \infty$ and events $A_1, A_2, \ldots$ are independent, then $P(A_n\ i.o.) = 1$.*

Proof. (1) Put $A = \{A_n \text{ i.o.}\}$. We have $A \subset \cup_{m \geq n} A_m$, and so $\mathbb{P}(A) \leq \sum_{m \geq n} \mathbb{P}(A_m) \to 0$ as $n \to \infty$.
(2) By independence

$$\mathbb{P}(\bigcap_{m \geq n} A_m^c) = \lim_{N\to\infty} \mathbb{P}(\bigcap_{m=n}^{N} A_m^c) = \prod_{m \geq n}(1 - \mathbb{P}(A_m)) \leq \exp\Big(-\sum_{m \geq n} \mathbb{P}(A_m)\Big) = 0.$$

It follows $\mathbb{P}(A^c) = \lim_n \mathbb{P}(\cap_{m \geq n} A_m^c) = 0$ which gives $\mathbb{P}(A) = 1$.

If events $A_1, A_2, \ldots$ are independent, then either $P(A_n \text{ i.o.}) = 0$ or $P(A_n \text{ i.o.}) = 1$. This is an example of a general zero-one law.

Definition 3.2 Let $X_1, X_2, \ldots$ be a sequence of random variables defined on the same probability space and $\mathcal{H}_n = \sigma(X_{n+1}, X_{n+2}, \ldots)$. Then $\mathcal{H}_n \supset \mathcal{H}_{n+1} \supset \ldots$, and we define the tail σ-algebra as $\mathcal{T} = \cap_n \mathcal{H}_n$.

Event H is in the tail σ-algebra if and only if changing the values of $X_1, \ldots, X_N$ does not affect the occurrence of H for any finite N.

Theorem 3.3 *Kolmogorov's zero-one law. Let $X_1, X_2, \ldots$ be independent random variables defined on the same probability space. For all events $H \in \mathcal{T}$ from the tail σ-algebra we have either $P(H) = 0$ or $P(H) = 1$.*

Proof. A standard result of measure theory asserts that for any $H \in \mathcal{H}_1$ there exists a sequence of events $C_n = \sigma(X_1, \ldots, X_n)$ such that $\mathbb{P}(H \Delta C_n) \to 0$ as $n \to \infty$. If $H \in \mathcal{T}$, then by independence

$$\mathbb{P}(H \cap C_n) = \mathbb{P}(H)\mathbb{P}(C_n) \to \mathbb{P}(H)^2$$

implying $\mathbb{P}(H) = \mathbb{P}(H)^2$.

Example 3.4 Examples of events belonging to the tail σ-algebra $\mathcal{T}$:

$$\{X_n > 0 \text{ i.o.}\}, \quad \{\limsup_{n\to\infty} X_n = \infty\}, \quad \{\sum_n X_n \text{ converges}\}.$$

These events are not affected by $X_1, \ldots, X_N$ for any fixed N.

Example 3.5 An example of an event $A \notin \mathcal{T}$ not belonging to the tail σ-algebra: suppose X_n may take only two values 1 and -1, and consider

$$A = \{S_n = 0 \text{ i.o.}\}, \text{ where } S_n = X_1 + \ldots + X_n.$$

Whether A occurs or not depends on the value of X_1. Indeed, if $X_2 = X_4 = \ldots = 1$ and $X_3 = X_5 = \ldots = -1$, then A occurs if $X_1 = -1$ and does not occur if $X_1 = 1$.

3.2   Inequalities

Jensen's inequality. Given a convex function $J(x)$ and a random variable X we have

$$J(\mathbb{E}(X)) \leq \mathbb{E}(J(X)).$$

Proof. Put $\mu = \mathbb{E}(X)$. Due to convexity there is λ such that $J(x) \geq J(\mu) + \lambda(x - \mu)$ for all x. Thus

$$\mathbb{E}(J(X)) \geq \mathbb{E}(J(\mu) + \lambda(X - \mu)) = J(\mu).$$

Markov's inequality. For any random variable X and $a > 0$

$$\mathbb{P}(|X| \geq a) \leq \frac{\mathbb{E}|X|}{a}.$$

Proof:

$$\mathbb{E}|X| \geq \mathbb{E}(|X|1_{\{|X| \geq a\}}) \geq a\mathbb{E}(1_{\{|X| \geq a\}}) = a\mathbb{P}(|X| \geq a).$$

Chebyshev's inequality. Given a random variable X with mean μ and variance σ^2 for any $\epsilon > 0$ we have

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

Proof:

$$\mathbb{P}(|X - \mu| \geq \epsilon) = \mathbb{P}((X - \mu)^2 \geq \epsilon^2) \leq \frac{\mathbb{E}((X - \mu)^2)}{\epsilon^2}.$$

Cauchy-Schwartz's inequality. The following inequality becomes an equality if only if $aX + bY = 1$ a.s. for a pair of constants $(a, b) \neq (0, 0)$:

$$\left(\mathbb{E}(XY)\right)^2 \leq \mathbb{E}(X^2)\mathbb{E}(Y^2).$$

Exercise 3.6 For a random variable X define its cumulant generating function by $\Lambda(t) = \log M(t)$, where $M(t) = \mathbb{E}(e^{tX})$ is the moment generating function assumed to be finite on an interval $t \in [0, z)$. Show that $\Lambda(0) = 0$, $\Lambda'(0) = \mu$, and

$$\Lambda''(t) = \frac{\mathbb{E}(e^{tX})\mathbb{E}(X^2 e^{tX}) - (\mathbb{E}(Xe^{tX}))^2}{M^2(t)}.$$

Deduce that $\Lambda(t)$ is convex on $[0, z)$.

Hölder's inequality. If $p, q > 1$ and $p^{-1} + q^{-1} = 1$, then

$$\mathbb{E}|XY| \leq \left(\mathbb{E}|X^p|\right)^{1/p}\left(\mathbb{E}|Y^q|\right)^{1/q}.$$

Lyapunov's inequality. If $0 < s < r$, then

$$\left(\mathbb{E}|X^s|\right)^{1/s} \leq \left(\mathbb{E}|X^r|\right)^{1/r}.$$

Proof. Using Hölder's inequality with $p = r/s$ and $q = (1 - s/r)^{-1}$ we obtain

$$\mathbb{E}\left(|X^s| \cdot 1\right) \leq \left(\mathbb{E}|X^s|^p\right)^{1/p} = \left(\mathbb{E}|X^r|\right)^{s/r}.$$

Minkowski's inequality. If $p \geq 1$, then the following triangle inequality holds

$$\left(\mathbb{E}|X + Y|^p\right)^{1/p} \leq \left(\mathbb{E}|X^p|\right)^{1/p} + \left(\mathbb{E}|Y^p|\right)^{1/p}.$$

Kolmogorov's inequality. Let $\{X_n\}$ be iid with zero means and variances σ_n^2. Then for any $\epsilon > 0$

$$\mathbb{P}(\max_{1 \leq i \leq n} |X_1 + \ldots + X_i| \geq \epsilon) \leq \frac{\sigma_1^2 + \ldots + \sigma_n^2}{\epsilon^2}.$$

3.3 Modes of convergence

Theorem 3.7 *If* $X_1, X_2, \ldots$ *are random variables defined on the same probability space, then so are*

$$\inf_n X_n, \quad \sup_n X_n, \quad \liminf_n X_n, \quad \limsup_n X_n.$$

Proof. For any real x

$$\{\omega : \inf_n X_n \leq x\} = \bigcup_n \{\omega : X_n \leq x\} \in \mathcal{F}, \quad \{\omega : \sup_n X_n \leq x\} = \bigcap_n \{\omega : X_n \leq x\} \in \mathcal{F}.$$

It remains to observe that

$$\liminf_n X_n = \sup_n \inf_{m \geq n} X_m, \qquad \limsup_n X_n = \inf_n \sup_{m \geq n} X_m.$$

Definition 3.8 Let $X, X_1, X_2, \ldots$ be random variables on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Define four modes of convergence of random variables
   (i) almost sure convergence $X_n \overset{\text{a.s.}}{\to} X$, if $\mathbb{P}(\omega : \lim X_n = X) = 1$,
   (ii) convergence in r-th mean $X_n \overset{L^r}{\to} X$ for a given $r \geq 1$, if $\mathbb{E}|X_n^r| < \infty$ for all n and

$$\mathbb{E}(|X_n - X|^r) \to 0,$$

   (iii) convergence in probability $X_n \overset{\text{P}}{\to} X$, if $\mathbb{P}(|X_n - X| > \epsilon) \to 0$ for all positive ϵ,
   (iv) convergence in distribution $X_n \overset{d}{\to} X$ (does not require a common probability space), if

$$\mathbb{P}(X_n \leq x) \to \mathbb{P}(X \leq x) \text{ for all } x \text{ such that } \mathbb{P}(X = x) = 0.$$

Theorem 3.9 *Let* $q \geq r \geq 1$. *The following implications hold*

$$X_n \overset{\text{a.s.}}{\to} X \Rightarrow_{(ii)}$$
$$X_n \overset{L^q}{\to} X \Rightarrow_{(i)} \quad X_n \overset{L^r}{\to} X \Rightarrow_{(iii)} \quad X_n \overset{\text{P}}{\to} X \Rightarrow_{(iv)} \quad X_n \overset{d}{\to} X.$$

Proof. The implication (i) is due to the Lyapunov inequality, while (iii) follows from the Markov inequality. For (ii) observe that the a.s. convergence is equivalent to $\mathbb{P}(\cup_{n \geq m}\{|X_n - X| > \epsilon\}) \to 0$, $m \to \infty$ for any $\epsilon > 0$. To prove (iv) use

$$\mathbb{P}(X_n \leq x) = \mathbb{P}(X_n \leq x, X \leq x + \epsilon) + \mathbb{P}(X_n \leq x, X > x + \epsilon) \leq \mathbb{P}(X \leq x + \epsilon) + \mathbb{P}(|X - X_n| > \epsilon),$$
$$\mathbb{P}(X \leq x - \epsilon) = \mathbb{P}(X_n \leq x, X \leq x - \epsilon) + \mathbb{P}(X_n > x, X \leq x - \epsilon) \leq \mathbb{P}(X_n \leq x) + \mathbb{P}(|X - X_n| > \epsilon).$$

Theorem 3.10 *Reverse implications:*
   (i) $X_n \overset{d}{\to} c \Rightarrow X_n \overset{\text{P}}{\to} c$ *for a constant limit,*
   (ii) $X_n \overset{\text{P}}{\to} X \Rightarrow X_n \overset{L^r}{\to} X$, *if* $\mathbb{P}(|X_n| \leq a) = 1$ *for all* n *and some positive constant* a,
   (iii) $X_n \overset{\text{P}}{\to} X \Rightarrow X_n \overset{\text{a.s.}}{\to} X$ *if* $\mathbb{P}(|X_n - X| > \epsilon) \to 0$ *so fast that*

$$\sum_n \mathbb{P}(|X_n - X| > \epsilon) < \infty \text{ for any } \epsilon > 0. \tag{2}$$

   (iv) $X_n \overset{\text{P}}{\to} X \Rightarrow X_{n'} \overset{\text{a.s.}}{\to} X$ *along a subsequence.*

Proof. To prove (i) use

$$\mathbb{P}(|X_n - c| > \epsilon) = \mathbb{P}(X_n < c - \epsilon) + \mathbb{P}(X_n > c + \epsilon).$$

To prove (ii) use $\mathbb{P}(|X| \leq a) = 1$ and

$$|X_n - X|^r \leq \epsilon^r 1_{\{|X_n - X| \leq \epsilon\}} + (2a)^r 1_{\{|X_n - X| > \epsilon\}}.$$

The implication (iii) follows from the first Borel-Cantelli lemma. Indeed, put $B_m = \{|X_n - X| > m^{-1}$ i.o.$\}$. Due to the Borel-Cantelli lemma, condition (2) implies $\mathbb{P}(B_m) = 0$ for any natural m. Remains to observe that $\{\omega : \lim X_n \neq X\} = \sup_m B_m$. The implication (iv) follows from (iii).

Example 3.11 Let $\Omega = [0, 2]$ and put $A_n = [a_n, a_{n+1}]$, where a_n is the fractional part of $1 + 1/2 + \ldots + 1/n$.

- The random variables 1_{A_n} converge to zero in mean (and therefore in probability) but not a.s.

- The random variables $n1_{A_n}$ converge to zero in probability but not in mean and not a.s.

- $\mathbb{P}(A_n \text{ i.o.}) = 0.5$ and $\sum_n \mathbb{P}(A_n) = \infty$.

Example 3.12 Let $\Omega = [0, 1]$ and put $B_n = [0, 1/n]$.

- The random variables $n1_{B_n}$ converge to zero a.s but not in mean.

- Put $X_{2n} = 1_{B_2}$ and $X_{2n+1} = 1 - X_{2n}$. The random variables X_n converge to 1_{B_2} in distribution but not in probability. The random variables X_n converge in distribution to $1 - 1_{B_2}$ as well.

- Both $X_{2n} = 1_{B_2}$ and $X_{2n+1} = 1 - X_{2n}$ converge to 1_{B_2} in distribution but their sum does not converge to 21_{B_2}.

Exercise 3.13 Suppose $X_n \xrightarrow{L^r} X$, where $r \geq 1$. Show that $\mathbb{E}(|X_n|^r) \to \mathbb{E}(|X|^r)$. (Hint: use Minkowski's inequality two times as you need two estimate $\limsup \mathbb{E}(|X_n|^r)$ from above and $\liminf \mathbb{E}(|X_n|^r)$ from below.)

Exercise 3.14 Suppose $X_n \xrightarrow{L^1} X$. Show that $\mathbb{E}(X_n) \to \mathbb{E}(X)$. (Hint: use Jensen's inequality.) Is the converse true?

Exercise 3.15 Suppose $X_n \xrightarrow{L^2} X$. Show that $\mathbb{V}ar(X_n) \to \mathbb{V}ar(X)$. (Hint :use the previous two exercises.)

3.4 Continuity of expectation

Theorem 3.16 *Bounded convergence. Suppose $|X_n| \leq M$ almost surely and $X_n \xrightarrow{\text{P}} X$. Then $\mathbb{E}(X) = \lim \mathbb{E}(X_n)$.*

Proof. Let $\epsilon > 0$ and use

$$|\mathbb{E}(X_n) - \mathbb{E}(X)| \leq \mathbb{E}|X_n - X| = \mathbb{E}\big(|X_n - X| \cdot 1_{\{|X_n - X| \leq \epsilon\}}\big) + \mathbb{E}\big(|X_n - X| \cdot 1_{\{|X_n - X| > \epsilon\}}\big)$$
$$\leq \epsilon + M\mathbb{P}(|X_n - X| > \epsilon).$$

Lemma 3.17 *Fatou's lemma. If almost surely $X_n \geq 0$, then $\liminf \mathbb{E}(X_n) \geq \mathbb{E}(\liminf X_n)$. In particular, applying this to $X_n = 1_{\{A_n\}}$ and $X_n = 1 - 1_{\{A_n\}}$ we get*

$$\mathbb{P}(\liminf A_n) \leq \liminf \mathbb{P}(A_n) \leq \limsup \mathbb{P}(A_n) \leq \mathbb{P}(\limsup A_n).$$

Proof. Put $Y_n = \inf_{m \geq n} X_m$. We have $Y_n \leq X_n$ and $Y_n \nearrow Y = \liminf X_n$. It suffices to show that $\liminf \mathbb{E}(Y_n) \geq \mathbb{E}(Y)$. Since, $|Y_n \wedge M| \leq M$, the bounded convergence theorem implies

$$\liminf_{n \to \infty} \mathbb{E}(Y_n) \geq \liminf_{n \to \infty} \mathbb{E}(Y_n \wedge M) = \mathbb{E}(Y \wedge M).$$

The convergence $\mathbb{E}(Y \wedge M) \to \mathbb{E}(Y)$ as $M \to \infty$ can be shown using the definition of expectation.

Theorem 3.18 *Monotone convergence. If $0 \leq X_n(\omega) \leq X_{n+1}(\omega)$ for all n and ω, then, clearly, for all ω, there exists a limit (possibly infinite) for the sequence $X_n(\omega)$. In this case $\mathbb{E}(\lim X_n) = \lim \mathbb{E}(X_n)$.*

Proof. From $\mathbb{E}(X_n) \leq \mathbb{E}(\lim X_n)$ we have $\limsup \mathbb{E}(X_n) \leq \mathbb{E}(\lim X_n)$. Now use Fatou's lemma.

Lemma 3.19 *Let X be a non-negative random variable with finite mean. Show that*

$$\mathbb{E}(X) = \int_0^\infty \mathbb{P}(X > x)dx.$$

Proof. Put $F(x) = P(X \le x)$. Integrating by parts we get

$$\mathbb{E}(X) = \int_0^\infty x dF(x) = \int_0^\infty x d(F(x) - 1) = x(F(x) - 1)|_0^\infty - \int_0^\infty (F(x) - 1) dx.$$

Thus it suffices to prove that $n(1 - F(n)) \to 0$ as $n \to \infty$. This follows from the monotone convergence theorem since

$$n(1 - F(n)) \le \mathbb{E}(X 1_{X > n}) \to 0.$$

Exercise 3.20 Let X be a non-negative random variable. Show that

$$\mathbb{E}(X^r) = r \int_0^\infty x^{r-1} \mathbb{P}(X > x) dx$$

for any $r \ge 1$ for which the expectation is finite.

Theorem 3.21 *Dominated convergence. If $X_n \overset{\text{a.s.}}{\to} X$ and almost surely $|X_n| \le Y$ and $\mathbb{E}(Y) < \infty$, then $\mathbb{E}|X| < \infty$ and $\mathbb{E}(X_n) \to \mathbb{E}(X)$.*

Proof. Apply Fatou's lemma twice.

Definition 3.22 A sequence X_n of random variables is said to be uniformly integrable if

$$\sup_n \mathbb{E}(|X_n|; |X_n| > a) \to 0, \quad a \to \infty.$$

Exercise 3.23 Put $Z = \sup_n |X_n|$. If $\mathbb{E}(Z) < \infty$, then the sequence X_n is uniformly integrable. Hint: $\mathbb{E}(|X_n| 1_A) \le \mathbb{E}(Z 1_A)$ for any event A.

Lemma 3.24 *A sequence X_n of random variables is uniformly integrable if and only if both of the following hold:*
(i) $\sup_n \mathbb{E}|X_n| < \infty$,
(ii) for all $\epsilon > 0$, there is $\delta > 0$ such that, for any event A such that $\mathbb{P}(A) < \delta$,

$$\sup_n \mathbb{E}(|X_n| 1_A) < \epsilon.$$

Proof. Step 1. Assume that (X_n) is uniformly integrable. For any positive a,

$$\sup_n \mathbb{E}|X_n| = \sup_n \left(\mathbb{E}(|X_n|; |X_n| \le a) + \mathbb{E}(|X_n|; |X_n| > a) \right) \le a + \sup_n \mathbb{E}(|X_n|; |X_n| > a).$$

Thus $\sup_n \mathbb{E}|X_n| < \infty$.
Step 2. Assume that (X_n) is uniformly integrable. Pick a positive a such that

$$\sup_n \mathbb{E}(|X_n| 1_{\{B_n\}}) < \epsilon/2, \qquad B_n = \{|X_n| > a\},$$

and put $\delta = \frac{\epsilon}{2a}$. If event A is such that $\mathbb{P}(A) < \delta$, then for all n,

$$\mathbb{E}(|X_n| 1_A) = \mathbb{E}(|X_n| 1_{\{A \cap B_n\}}) + \mathbb{E}(|X_n| 1_{\{A \cap B_n^c\}}) \le \mathbb{E}(|X_n| 1_{\{B_n\}}) + a \mathbb{P}(A) < \epsilon.$$

Step 3. Assume that (i) and (ii) hold. Let $\epsilon > 0$ and pick δ according to (ii). To prove that (X_n) is uniformly integrable it suffices to verify, see (ii), that

$$\sup_n \mathbb{P}(|X_n| > a) < \delta$$

for sufficiently large a such that $a > \delta^{-1} \sup_n \mathbb{E}|X_n|$. But this is an easy consequence of the Markov inequality

$$\sup_n \mathbb{P}(|X_n| > a) \le a^{-1} \sup_n \mathbb{E}|X_n| < \delta.$$

Theorem 3.25 *Let $X_n \xrightarrow{\text{P}} X$. The following three statements are equivalent to one another.*
   (a) The sequence X_n is uniformly integrable.

   (b) $\mathbb{E}|X_n| < \infty$ for all n, $\mathbb{E}|X| < \infty$, and $X_n \xrightarrow{L^1} X$.
   (c) $\mathbb{E}|X_n| < \infty$ for all n, $\mathbb{E}|X| < \infty$, and $\mathbb{E}|X_n| \to \mathbb{E}|X|$.

Theorem 3.26 *Let $r \geq 1$. If $X_n \xrightarrow{\text{P}} X$ and the sequence $|X_n^r|$ is uniformly integrable, then $X_n \xrightarrow{L^r} X$.*

Proof. Step 1. Show that $\mathbb{E}|X^r| < \infty$. There is a subsequence $(X_{n'})$ which converges to X almost surely. By Fatou's lemma and Lemma 3.24,

$$\mathbb{E}|X^r| = \mathbb{E}\Big(\liminf_{n' \to \infty} |X_{n'}^r| \Big) \leq \liminf_{n' \to \infty} \mathbb{E}|X_{n'}^r| \leq \sup_n \mathbb{E}|X_n^r| < \infty.$$

Step 2. Fix an arbitrary $\epsilon > 0$ and put $B_n = \{|X_n - X| > \epsilon\}$. We have $\mathbb{P}(B_n) \to 0$ as $n \to 0$, and

$$\mathbb{E}|X_n - X|^r \leq \epsilon^r + \mathbb{E}\Big(|X_n - X|^r 1_{B_n} \Big).$$

By the Minkowski inequality

$$\Big[\mathbb{E}\Big(|X_n - X|^r 1_{B_n} \Big)\Big]^{1/r} \leq \Big[\mathbb{E}\Big(|X_n|^r 1_{B_n} \Big)\Big]^{1/r} + \Big[\mathbb{E}\Big(|X|^r 1_{B_n} \Big)\Big]^{1/r}.$$

It remains to see that the last two expectations both go to 0 as $n \to \infty$: the first by Lemma 3.24 and the second by step 1.

4 Limit theorems for the sums if IID random variables

4.1 Weak law of large numbers

Definition 4.1 Convergence in distribution $X_n \xrightarrow{d} X$ means

$$\mathbb{P}(X_n \leq x) \to \mathbb{P}(X \leq x) \text{ for all } x \text{ such that } \mathbb{P}(X = x) = 0.$$

This is equvalent to the weak convergence $F_n \xrightarrow{d} F$ of distribution functions when $F_n(x) \to F(x)$ at each point x where F is continuous.

Theorem 4.2 *Weak convergence and convergence of characteristic functions:*
   (i) two r.v. have the same characteristic function iff they have the same distribution function,
   (ii) if $X_n \xrightarrow{d} X$, then $\phi_n(t) \to \phi(t)$ for all t,
   (iii) conversely, if $\phi(t) = \lim \phi_n(t)$ exists and continuous at $t = 0$, then ϕ is cf of some F, and $F_n \xrightarrow{d} F$.

Theorem 4.3 *If $X_1, X_2, \ldots$ are iid with finite mean μ and $S_n = X_1 + \ldots + X_n$, then*

$$S_n/n \xrightarrow{d} \mu, \quad n \to \infty.$$

Proof. Let F_n and ϕ_n be the df and cf of $n^{-1}S_n$. To prove $F_n(x) \xrightarrow{d} 1_{\{x \geq \mu\}}$ we have to see that $\phi_n(t) \to e^{it\mu}$ which is obtained using a Taylor expansion

$$\phi_n(t) = \Big(\phi_1(tn^{-1}) \Big)^n = \Big(1 + i\mu t n^{-1} + o(n^{-1}) \Big)^n \to e^{it\mu}.$$

Example 4.4 Statistical application: the sample mean is a consistent estimate of the population mean. Counterexample: if $X_1, X_2, \ldots$ are iid with the Cauchy distribution, then $S_n/n \xlongequal{d} X_1$ since $\phi_n(t) = \phi_1(t)$.

4.2 Central limit theorem

The LLN says that $|S_n - n\mu|$ is much smaller than n. The CLT says that this difference is of order $\sqrt{n}$.

Theorem 4.5 *If $X_1, X_2, \ldots$ are iid with finite mean μ and positive finite variance σ^2, then for any x*

$$\mathbb{P}\Big(\frac{S_n - n\mu}{\sigma\sqrt{n}} \le x\Big) \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2} dy, \quad n \to \infty.$$

Proof. Let ψ_n be the cf of $\frac{S_n - n\mu}{\sigma\sqrt{n}}$. Using a Taylor expansion we obtain

$$\psi_n(t) = \Big(1 - \frac{t^2}{2n} + o(n^{-1})\Big)^n \to e^{-t^2/2}.$$

Example 4.6 Important example: simple random walks. 280 years ago de Moivre (1733) obtained the first CLT in the symmetric case with $p = 1/2$.

Statistical application: the standardized sample mean has the sampling distribution which is approximately $N(0, 1)$. Approximate 95% confidence interval formula for the mean $\bar{X} \pm 1.96 \frac{s}{\sqrt{n}}$.

Theorem 4.7 *Let $(X_1^n, \ldots, X_r^n)$ have the multinomial distribution $Mn(n, p_1, \ldots, p_r)$. Then the normalized vector $\Big(\frac{X_1^n - np_1}{\sqrt{n}}, \ldots, \frac{X_r^n - np_r}{\sqrt{n}}\Big)$ converges in distribution to the multivariate normal distribution with zero means and the covariance matrix*

$$\mathbf{V} = \begin{pmatrix} p_1(1-p_1) & -p_1 p_2 & -p_1 p_3 & \ldots & -p_1 p_r \\ -p_2 p_1 & p_2(1-p_2) & -p_2 p_3 & \ldots & -p_2 p_r \\ -p_3 p_1 & -p_3 p_2 & p_3(1-p_3) & \ldots & -p_3 p_r \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ -p_r p_1 & -p_r p_2 & -p_r p_3 & \ldots & p_r(1-p_r) \end{pmatrix}.$$

Proof. To apply the continuity property of the multivariate characteristic functions consider

$$\mathbb{E}\exp\Big(i\theta_1\frac{X_1^n - np_1}{\sqrt{n}} + \ldots + i\theta_r\frac{X_r^n - np_r}{\sqrt{n}}\Big) = \Big(\sum_{j=1}^{r} p_j e^{i\tilde{\theta}_j/\sqrt{n}}\Big)^n,$$

where $\tilde{\theta}_j = \theta_j - (\theta_1 p_1 + \ldots + \theta_r p_r)$. Similarly to the classical case we have

$$\Big(\sum_{j=1}^{r} p_j e^{i\tilde{\theta}_j/\sqrt{n}}\Big)^n = \Big(1 - \frac{1}{2n}\sum_{j=1}^{r} p_j \tilde{\theta}_j^2 + o(n^{-1})\Big)^n \to e^{-\frac{1}{2}\sum_{j=1}^{r} p_j \tilde{\theta}_j^2} = e^{-\frac{1}{2}(\sum_{j=1}^{r} p_j \theta_j^2 - (\sum_{j=1}^{r} p_j \theta_j)^2)}.$$

It remains to see that the right hand side equals $e^{-\frac{1}{2}\boldsymbol{\theta}\mathbf{V}\boldsymbol{\theta}^{\mathrm{t}}}$ which follows from the representation

$$\mathbf{V} = \begin{pmatrix} p_1 & & 0 \\ & \ddots & \\ 0 & & p_r \end{pmatrix} - \begin{pmatrix} p_1 \\ \vdots \\ p_r \end{pmatrix} \begin{pmatrix} p_1, \ldots, p_r \end{pmatrix}.$$

4.3 Strong LLN

Theorem 4.8 *Let $X_1, X_2, \ldots$ be iid random variables defined on the same probability space with mean μ and finite second moment. Then*

$$\frac{X_1 + \ldots + X_n}{n} \xrightarrow{L^2} \mu.$$

Proof. Since $\sigma^2 := \mathbb{E}(X_1^2) - \mu^2 < \infty$, we have

$$\mathbb{E}\Big(\Big(\frac{X_1 + \ldots + X_n}{n} - \mu\Big)^2\Big) = \mathbb{V}ar\Big(\frac{X_1 + \ldots + X_n}{n}\Big) = \frac{n\sigma^2}{n^2} \to 0.$$

Theorem 4.9 *Strong LLN. Let $X_1, X_2, \ldots$ be iid random variables defined on the same probability space. Then*

$$\frac{X_1 + \ldots + X_n}{n} \overset{\text{a.s.}}{\to} \mu$$

for some constant μ iff $\mathbb{E}|X_1| < \infty$. In this case $\mu = \mathbb{E}X_1$ and $\frac{X_1 + \ldots + X_n}{n} \overset{L^1}{\to} \mu$.

There are cases when convergence in probability holds but not a.s. In those cases of course $\mathbb{E}|X_1| = \infty$.

Theorem 4.10 *The law of the iterated logarithm. Let $X_1, X_2, \ldots$ be iid random variables with mean 0 and variance 1. Then*

$$\mathbb{P}\left(\limsup_{n \to \infty} \frac{X_1 + \ldots + X_n}{\sqrt{2n \log \log n}} = 1\right) = 1$$

and

$$\mathbb{P}\left(\liminf_{n \to \infty} \frac{X_1 + \ldots + X_n}{\sqrt{2n \log \log n}} = -1\right) = 1.$$

Proof sketch. The second assertion of Theorem 4.10 follows from the first one after applying it to $-X_i$. The proof of the first part is difficult. One has to show that the events

$$A_n(c) = \{X_1 + \ldots + X_n \geq c\sqrt{2n \log \log n}\}$$

occur for infinitely many values of n if $c < 1$ and for only finitely many values of n if $c > 1$.

4.4 Large deviations

Let $X_1, X_2, \ldots$ be iid random variables with mean μ, variance σ^2, and partial sums $S_n = X_1 + \ldots + X_n$. Let $\Lambda(t)$ be the cumulant generating function for a typical X_i, see Exercise 3.6. For the convex function $\Lambda(t)$ we define the Fenchel-Legendre transform by

$$\Lambda^*(a) = \sup_{t \in \mathbb{R}}\{at - \Lambda(t)\}, \qquad a \in \mathbb{R}.$$

Observe that $\Lambda^*(a) > 0$ for $a > \mu$. This follows from the representation

$$at - \Lambda(t) = \log \frac{e^{at}}{M(t)} = \log \frac{1 + at + o(t)}{1 + \mu t + \frac{1}{2}\sigma^2 t^2 + o(t^2)}$$

showing that $at - \Lambda(t)$ is positive for sufficiently small $t > 0$.

Theorem 4.11 *Let $X_1, X_2, \ldots$ be iid random variables with mean μ, and suppose that their moment generating function $M(t) = \mathbb{E}(e^{tX})$ is finite in some neighborhood of the origin $t = 0$. Let $a > \mu$ be such that $\mathbb{P}(X > a)$ is positive. Then $\Lambda^*(a) > 0$ and*

$$\frac{1}{n} \log \mathbb{P}(S_n > na) \to -\Lambda^*(a), \quad n \to \infty.$$

Without proof.

Exercise 4.12 What happens if a is such that $\mathbb{P}(X > a) = 0$?

Corollary 4.13 *Let $X_1, X_2, \ldots$ be iid random variables with mean μ, and suppose that their moment generating function $M(t) = \mathbb{E}(e^{tX})$ is finite in some neighborhood of the origin $t = 0$. Let $a < \mu$ be such that $\mathbb{P}(X < a) > 0$. Then $\Lambda^*(a) > 0$ and*

$$\frac{1}{n} \log \mathbb{P}(S_n < na) \to -\Lambda^*(a), \quad n \to \infty.$$

Proof. Replace X_i by $-\bar{X}_i$ and put $\bar{a} = -a$. Then $\{S_n < na\} = \{\bar{S}_n > n\bar{a}\}$ and $\bar{a} > \bar{\mu}$. Moreover, $\bar{\Lambda}(t) = \Lambda(-t)$ and therefore, $\bar{\Lambda}^*(a) = \Lambda^*(-a)$. Thus according to the theorem

$$\frac{1}{n} \log \mathbb{P}(S_n < na) = \frac{1}{n} \log \mathbb{P}(\bar{S}_n > n\bar{a}) \to -\bar{\Lambda}^*(\bar{a}) = -\Lambda^*(a), \quad n \to \infty.$$

Example 4.14 For the normally distributed X_i with a common density $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, the normalized partial sums S_n/n are also normally distributed with mean μ and variance σ^2/n so that

$$\mathbb{P}(S_n > na) = \frac{1}{\sqrt{2\pi}} \int_{\frac{(a-\mu)\sqrt{n}}{\sigma}}^{\infty} e^{-\frac{y^2}{2}} \, dy.$$

For $x > 0$, we have

$$(x^{-1} - x^{-3})e^{-\frac{x^2}{2}} \le \int_x^{\infty} e^{-\frac{y^2}{2}} \, dy \le x^{-1} e^{-\frac{x^2}{2}}.$$

Thus for $a > \mu$ we obtain

$$\mathbb{P}(S_n > na) \sim \frac{\sigma}{\sqrt{2\pi n}(a-\mu)} e^{-\frac{(a-\mu)^2}{2\sigma^2} \cdot n}.$$

It follows,

$$\frac{1}{n} \log \mathbb{P}(S_n < na) \to -\frac{(a-\mu)^2}{2\sigma^2}, \quad n \to \infty.$$

From $M(t) = e^{t\mu + \frac{1}{2}t^2\sigma^2}$, we find $\Lambda(t) = t\mu + \frac{1}{2}t^2\sigma^2$ and

$$at - \Lambda(t) = t(a-\mu) - \frac{1}{2}t^2\sigma^2 = \frac{(a-\mu)^2}{2\sigma^2} - \frac{1}{2}\left(\frac{a-\mu}{\sigma} - t\sigma\right)^2.$$

Clearly, this quadratic function reaches its maximum at $t = \frac{a-\mu}{\sigma^2}$. The maximum is

$$\Lambda^*(a) = \frac{(a-\mu)^2}{2\sigma^2}.$$

Example 4.15 For X_i with $\mathrm{Ber}(p)$ distribution, the partial sum S_n has a binomial distribution with parameters (n, p)

$$\mathbb{P}(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

Put

$$H(a) = a \log \frac{a}{p} + (1-a) \log \frac{1-a}{1-p}, \quad a \in (p, 1).$$

Since $H(p) = 0$ and $H'(a) = \log \frac{a(1-p)}{p(1-a)}$, we conclude that $H(a) > 0$. Using the Stirling formula

$$\sqrt{2\pi n} \, n^n e^{-n} e^{\frac{1}{12n+1}} \le n! \le \sqrt{2\pi n} \, n^n e^{-n} e^{\frac{1}{12n}},$$

one can show that

$$\mathbb{P}(S_n > na) \sim \frac{a(1-p)}{a-p} \frac{1}{\sqrt{2\pi a(1-a)n}} e^{-nH(a)}.$$

On the other hand $M(t) = pe^t + 1 - p$ and

$$f(t) = at - \Lambda(t) = at - \log(pe^t + 1 - p)$$

has the following derivatives

$$f'(t) = a - \frac{pe^t}{pe^t + 1 - p} = a - 1 + \frac{1-p}{pe^t + 1 - p}, \quad f''(t) = -\frac{(1-p)pe^t}{(pe^t + 1 - p)^2}.$$

Thus the maximum of $f(t)$ is achieved at t satisfying

$$pe^t + 1 - p = \frac{1-p}{1-a}, \quad t = \log \frac{a(1-p)}{p(1-a)},$$

and the maximum equals

$$\Lambda^*(a) = a \log \frac{a(1-p)}{p(1-a)} - \log \frac{1-p}{1-a} = H(a).$$

5 Markov chains

5.1 Simple random walks

Let $S_n = a + X_1 + \ldots + X_n$ where $X_1, X_2, \ldots$ are IID r.v. taking values 1 and -1 with probabilities p and $q = 1 - p$. This Markov chain is homogeneous both in space and time. We have $S_n = 2Z_n - n$, with $Z_n \sim \text{Bin}(n, p)$. Symmetric random walk if $p = 0.5$. Drift upwards $p > 0.5$ or downwards $p < 0.5$ (like in casino).

(i) The ruin probability $p_k = p_k(N)$: your starting capital k against casino's $N - k$. The difference equation

$$p_k = p \cdot p_{k+1} + q \cdot p_{k-1}, \quad p_N = 0, \quad p_0 = 1$$

gives

$$p_k(N) = \begin{cases} \frac{(q/p)^N - (q/p)^k}{(q/p)^N - 1}, & \text{if } p \neq 0.5, \\ \frac{N-k}{N}, & \text{if } p = 0.5. \end{cases}$$

Start from zero and let τ_b be the first hitting time of b, then for $b > 0$

$$\mathbb{P}(\tau_{-b} < \infty) = \lim_{N \to \infty} p_b(N) = \begin{cases} 1, & \text{if } p \leq 0.5, \\ (q/p)^b, & \text{if } p > 0.5, \end{cases}$$

and

$$\mathbb{P}(\tau_b < \infty) = \begin{cases} 1, & \text{if } p \geq 0.5, \\ (p/q)^b, & \text{if } p < 0.5. \end{cases}$$

(ii) The mean number $D_k = D_k(N)$ of steps before hitting either 0 or N. The difference equation

$$D_k = p \cdot (1 + D_{k+1}) + q \cdot (1 + D_{k-1}), \quad D_0 = D_N = 0$$

gives

$$D_k(N) = \begin{cases} \frac{1}{q-p} \left[k - N \cdot \frac{1 - (q/p)^k}{1 - (q/p)^N} \right], & \text{if } p \neq 0.5, \\ k(N-k), & \text{if } p = 0.5. \end{cases}$$

If $p < 0.5$, then the expected ruin time is computed as $D_k(N) \to \frac{k}{q-p}$ as $N \to \infty$.

(iii) There are

$$N_n(a, b) = \binom{n}{k}, \quad k = \frac{n + b - a}{2}$$

paths from a to b in n steps. Each path has probability $p^k q^{n-k}$. Thus

$$\mathbb{P}(S_n = b | S_0 = a) = \binom{n}{k} p^k q^{n-k}, \quad k = \frac{n + b - a}{2}.$$

In particular, $\mathbb{P}(S_{2n} = a | S_0 = a) = \binom{2n}{n}(pq)^n$. Reflection principle: the number of n-paths visiting r is

$$N_n^r(a, b) = N_n(2r - a, b), \quad a \geq r, b \geq r,$$
$$N_n^r(a, b) = N_n(a, 2r - b), \quad a < r, b < r.$$

(iv) Ballot theorem: if $b > 0$, then the number of n-paths $0 \to b$ not revisiting zero is

$$N_{n-1}(1, b) - N_{n-1}^0(1, b) = N_{n-1}(1, b) - N_{n-1}(-1, b)$$
$$= \binom{n-1}{\frac{n+b}{2} - 1} - \binom{n-1}{\frac{n+b}{2}} = (b/n) N_n(0, b).$$

Thus (by default we will assume $S_0 = 0$)

$$\mathbb{P}(S_1 > 0, \ldots S_{n-1} > 0 | S_n = b) = \frac{b}{n}, \quad b > 0,$$

$$\mathbb{P}(S_1 \neq 0, \ldots S_{n-1} \neq 0, S_n = b) = \frac{|b|}{n} \mathbb{P}(S_n = b),$$

$$\mathbb{P}(S_1 \neq 0, \ldots S_n \neq 0) = n^{-1} \mathbb{E}|S_n|.$$

It follows that

$$\mathbb{P}(S_1 \neq 0, \dots S_{2n} \neq 0) = \mathbb{P}(S_{2n} = 0) \text{ for } p = 0.5. \tag{3}$$

Indeed,

$$\mathbb{P}(S_1 \neq 0, \dots S_{2n} \neq 0) = 2 \sum_{k=1}^{n} \frac{2k}{2n} \mathbb{P}(S_{2n} = 2k) = 2 \sum_{k=1}^{n} \frac{2k}{2n} \binom{2n}{n+k} 2^{-2n}$$

$$= 2^{-2n+1} \sum_{k=1}^{n} \binom{2n-1}{n+k-1} - \binom{2n-1}{n+k} = 2^{-2n+1} \binom{2n-1}{n} = \mathbb{P}(S_{2n} = 0).$$

(v) For the maximum $M_n = \max\{S_0, \dots, S_n\}$ using $N_n^r(0, b) = N_n(0, 2r - b)$ for $r > b$ and $r > 0$ we get

$$\mathbb{P}(M_n \geq r, S_n = b) = (q/p)^{r-b} \mathbb{P}(S_n = 2r - b),$$

implying for $b > 0$

$$\mathbb{P}(S_1 < b, \dots S_{n-1} < b, S_n = b) = \frac{b}{n} \mathbb{P}(S_n = b).$$

The obtained equality

$$\mathbb{P}(S_1 > 0, \dots S_{n-1} > 0, S_n = b) = \mathbb{P}(S_1 < b, \dots S_{n-1} < b, S_n = b)$$

can be explained in terms of the reversed walk also starting at zero: the initial walk comes to b without revisiting zero means that the reversed walk reaches its maximum on the final step.

(vi) The first hitting time τ_b has distribution

$$\mathbb{P}(\tau_b = n) = \frac{|b|}{n} \mathbb{P}(S_n = b), \quad n > 0.$$

The mean number of visits of $b \neq 0$ before revisiting zero

$$\mathbb{E} \sum_{n=1}^{\infty} 1_{\{S_1 \neq 0, \dots S_{n-1} \neq 0, S_n = b\}} = \sum_{n=1}^{\infty} \mathbb{P}(\tau_b = n) = \mathbb{P}(\tau_b < \infty).$$

Theorem 5.1 *Arcsine law for the last visit to the origin. Let $p = 0.5$, $S_0 = 0$, and T_{2n} be the time of the last visit to zero up to time $2n$. Then*

$$\mathbb{P}(T_{2n} \leq 2xn) \to \int_0^x \frac{dy}{\pi\sqrt{y(1-y)}} = \frac{2}{\pi} \arcsin\sqrt{x}, \quad n \to \infty.$$

Proof sketch. Using (3) we get

$$\mathbb{P}(T_{2n} = 2k) = \mathbb{P}(S_{2k} = 0)\mathbb{P}(S_{2k+1} \neq 0, \dots, S_{2n} \neq 0 | S_{2k} = 0)$$
$$= \mathbb{P}(S_{2k} = 0)\mathbb{P}(S_{2(n-k)} = 0),$$

and it remains to apply Stirling's formula.

Theorem 5.2 *Arcsine law for sojourn times. Let $p = 0.5$, $S_0 = 0$, and T_{2n}^+ be the number of time intervals spent on the positive side up to time $2n$. Then $T_{2n}^+ \stackrel{d}{=} T_{2n}$.*

Proof sketch. First using

$$\mathbb{P}(S_1 > 0, \dots, S_{2n} > 0) = \mathbb{P}(S_1 = 1, S_2 \geq 1, \dots, S_{2n} \geq 1) = \frac{1}{2}\mathbb{P}(T_{2n}^+ = 2n)$$

and (3) observe that

$$\mathbb{P}(T_{2n}^+ = 0) = \mathbb{P}(T_{2n}^+ = 2n) = \mathbb{P}(S_{2n} = 0).$$

Then by induction over n one can show that

$$\mathbb{P}(T_{2n}^+ = 2k) = \mathbb{P}(S_{2k} = 0)\mathbb{P}(S_{2(n-k)} = 0)$$

for $k = 1, \dots, n-1$, applying the following useful relation

$$\mathbb{P}(S_{2n} = 0) = \sum_{k=1}^{n} \mathbb{P}(S_{2(n-k)} = 0)\mathbb{P}(\tau_0 = 2k),$$

where τ_0 is the time of first return to zero.

5.2 Markov chains

Conditional on the present value, the future of the system is independent of the past. A Markov chain $\{X_n\}_{n=0}^{\infty}$ with countably many states and transition matrix $\mathbf{P}$ with elements p_{ij}

$$\mathbb{P}(X_n = j | X_{n-1} = i, X_{n-2} = i_{n-2}, \ldots, X_0 = i_0) = p_{ij}.$$

The n-step transition matrix with elements $p_{ij}^{(n)} = \mathbb{P}(X_{n+m} = j | X_m = i)$ equals $\mathbf{P}^n$. Given the initial distribution $\mathbf{a}$ as the vector with components $a_i = \mathbb{P}(X_0 = i)$, the distribution of X_n is given by the vector $\mathbf{aP}^n$ since

$$\mathbb{P}(X_n = j) = \sum_{i=-\infty}^{\infty} \mathbb{P}(X_n = j | X_0 = i)\mathbb{P}(X_0 = i) = \sum_{i=-\infty}^{\infty} a_i p_{ij}^{(n)}.$$

Example 5.3 Examples of Markov chains:

- IID chain has transition probabilities $p_{ij} = p_j$,

- simple random walk has transition probabilities $p_{ij} = p1_{\{j=i+1\}} + q1_{\{j=i-1\}}$,

- Bernoulli process has transition probabilities $p_{ij} = p1_{\{j=i+1\}} + q1_{\{j=i\}}$ and state space $S = \{0, 1, 2, \ldots\}$.

Lemma 5.4 *Hitting times. Let $T_i = \min\{n \geq 1 : X_n = i\}$ and put $f_{ij}^{(n)} = \mathbb{P}(T_j = n | X_0 = i)$. Define the generating functions*

$$P_{ij}(s) = \sum_{n=0}^{\infty} s^n p_{ij}^{(n)}, \qquad F_{ij}(s) = \sum_{n=1}^{\infty} s^n f_{ij}^{(n)}.$$

It is not difficult to see that

$$P_{ij}(s) = 1_{\{j=i\}} + F_{ij}(s)P_{jj}(s)$$
$$P_{ii}(s) = \frac{1}{1 - F_{ii}(s)}.$$

Definition 5.5 Classification of states

- state i is called recurrent (persistent), if $\mathbb{P}(T_i < \infty | X_0 = i) = 1$,

- a non-recurrent state is called a transient state,

- a recurrent state i is called null-recurrent, if $\mathbb{E}(T_i | X_0 = i) = \infty$,

- state i is called positive-recurrent, if $\mathbb{E}(T_i | X_0 = i) < \infty$.

Theorem 5.6 *State i is recurrent iff $\sum_{n=1}^{\infty} p_{ii}^{(n)} = \infty$. A recurrent state i is null-recurrent iff $p_{ii}^{(n)} \to 0$. In the latter case $p_{ij}^{(n)} \to 0$ for all j.*

Proof sketch. Since $F_{ii}(1) = \mathbb{P}(T_i < \infty | X_0 = i)$, we conclude that state i is recurrent iff the expected number of visits of the state is infinite $P_{ii}(1) = \infty$. See also the ergodic Theorem 5.14.

Example 5.7 For a simple random walk

$$\mathbb{P}(S_{2n} = i | S_0 = i) = \binom{2n}{n}(pq)^n.$$

Using the Stirling formula $n! \sim n^n e^{-n}\sqrt{2\pi n}$ we get

$$p_{ii}^{(2n)} \sim \frac{(4pq)^n}{\sqrt{\pi n}}, \quad n \to \infty.$$

Criterium of recurrence $\sum p_{ii}^{(n)} = \infty$ holds only if $p = 0.5$ when $p_{ii}^{(2n)} \sim \frac{1}{\sqrt{\pi n}}$. The one and two-dimensional symmetric simple random walks are null-recurrent but the three-dimensional walk is transient!

Definition 5.8 The period $d(i)$ of state i is the greatest common divisor of n such that $p_{ii}^{(n)} > 0$. We call i periodic if $d(i) \geq 2$ and aperiodic if $d(i) = 1$.

If two states i and j communicate with each other, then

- i and j have the same period,

- i is transient iff j is transient,

- i is null-recurrent iff j is null-recurrent.

Definition 5.9 A chain is called irreducible if all states communicate with each other.

All states in an irreducible chain have the same period d. It is called the period of the chain. Example: a simple random walk is periodic with period 2. Irreducible chains are classified as transient, recurrent, positively recurrent, or null-recurrent.

Definition 5.10 State i is absorbing if $p_{ii} = 1$. More generally, C is called a closed set of states, if $p_{ij} = 0$ for all $i \in C$ and $j \notin C$.

The state space S can be partitioned uniquely as

$$S = T \cup C_1 \cup C_2 \cup \dots,$$

where T is the set of transient states, and the C_i are irreducible closed sets of recurrent states. If S is finite, then at least one state is recurrent and all recurrent states are positively recurrent.

5.3 Stationary distributions

A vector of probabilities $\boldsymbol{\pi} = (\pi_j, j \in S)$ is a stationary distribution for the Markov chain X_n, if given X_0 has distribution $\boldsymbol{\pi}$, X_n has the same distribution $\boldsymbol{\pi}$ for any n, or in other words $\boldsymbol{\pi}$ is a left eigenvector of the transition matrix

$$\boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi}.$$

Theorem 5.11 *An irreducible chain (aperiodic or periodic) has a stationary distribution $\boldsymbol{\pi}$ iff the chain is positively recurrent; in this case $\boldsymbol{\pi}$ is the unique stationary distribution and is given by $\pi_i = 1/\mu_i$, where $\mu_i = \mathbb{E}(T_i|X_0 = i)$ and T_i is the time of first return to i.*

Proof sketch. Let $\boldsymbol{\rho}(k) = (\rho_j(k), j \in S)$ where $\rho_k(k) = 1$ and

$$\rho_j(k) = \sum_{n=1}^{\infty} \mathbb{P}(X_n = j, T_k \geq n | X_0 = k)$$

is the mean number of visits of the chain to the state j between two consecutive visits to state k. Then

$$\sum_{j \in S} \rho_j(k) = \sum_{j \in S} \sum_{n=1}^{\infty} \mathbb{P}(X_n = j, T_k \geq n | X_0 = k)$$

$$= \sum_{n=1}^{\infty} \mathbb{P}(T_k \geq n | X_0 = k) = \mathbb{E}(T_k | X_0 = k) = \mu_k.$$

If the chain is irreducible recurrent, then $\rho_j(k) < \infty$ for any k and j, and furthermore, $\boldsymbol{\rho}(k)\mathbf{P} = \boldsymbol{\rho}(k)$. Thus there exists a positive root $\mathbf{x}$ of the equation $\mathbf{x}\mathbf{P} = \mathbf{x}$, which is unique up to a multiplicative constant; the chain is positively recurrent iff $\sum_{j \in S} x_j < \infty$.

Theorem 5.12 *Let s be any state of an irreducible chain. The chain is transient iff there exists a non-zero bounded solution $(y_j : j \neq s)$ satisfying $|y_j| \leq 1$ for all j to the equations*

$$y_i = \sum_{j \in S \setminus \{s\}} p_{ij} y_j, \quad i \in S \setminus \{s\}. \tag{4}$$

Proof sketch. Main step. Let τ_j be the probability of no visit to s ever for a chain started at state j. Then the vector $(\tau_j : j \neq s)$ satisfies (4).

Example 5.13 Random walk with retaining barrier. Transition probabilities

$$p_{00} = q, \quad p_{i-1,i} = p, \quad p_{i,i-1} = q, \quad i \geq 1.$$

Let $\rho = p/q$.

- If $q < p$, take $s = 0$ to see that $y_j = 1 - \rho^{-j}$ satisfies (4). The chain is transient.

- Solve the equation $\boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi}$ to find that there exists a stationary distribution, with $\pi_j = \rho^j(1 - \rho)$, if and only if $q > p$.

- If $q > p$, the chain is positively recurrent, and if $q = p = 1/2$, the chain is null recurrent.

Theorem 5.14 *Ergodic theorem. For an irreducible aperiodic chain we have that*

$$p_{ij}^{(n)} \to \frac{1}{\mu_j} \text{ as } n \to \infty \text{ for all } (i,j).$$

More generally, for an aperiodic state j and any state i we have that $p_{ij}^{(n)} \to \frac{f_{ij}}{\mu_j}$, where f_{ij} is the probability that the chain ever visits j starting at i.

5.4 Reversibility

Theorem 5.15 *Put $Y_n = X_{N-n}$ for $0 \leq n \leq N$ where X_n is a stationary Markov chain. Then Y_n is a Markov chain with*

$$\mathbb{P}(Y_{n+1} = j | Y_n = i) = \frac{\pi_j p_{ji}}{\pi_i}.$$

The chain Y_n is called the time-reversal of X_n. If $\boldsymbol{\pi}$ exists and $\frac{\pi_j p_{ji}}{\pi_i} = p_{ij}$, the chain X_n is called reversible (in equilibrium). The detailed balance equations

$$\pi_i p_{ij} = \pi_j p_{ji} \text{ for all } (i,j). \tag{5}$$

Theorem 5.16 *Consider an irreducible chain and suppose there exists a distribution $\boldsymbol{\pi}$ such that (5) holds. Then $\boldsymbol{\pi}$ is a stationary distribution of the chain. Furthermore, the chain is reversible.*

Proof. Using (5) we obtain

$$\sum_i \pi_i p_{ij} = \sum_i \pi_j p_{ji} = \pi_j.$$

Example 5.17 Ehrenfest model of diffusion: flow of m particles between two connected chambers. Pick a particle at random and move it to another chamber. Let X_n be the number of particles in the first chamber. State space $S = \{0, 1, \ldots, m\}$ and transition probabilities

$$p_{i,i+1} = \frac{m-i}{m}, \qquad p_{i,i-1} = \frac{i}{m}.$$

The detailed balance equations

$$\pi_i \frac{m-i}{m} = \pi_{i+1} \frac{i+1}{m}$$

imply

$$\pi_i = \frac{m-i+1}{i} \pi_{i-1} = \binom{m}{i} \pi_0.$$

Using $\sum_i \pi_i = 1$ we find that the stationary distribution $\pi_i = \binom{m}{i} 2^{-n}$ is a symmetric binomial.

5.5 Branching process

We introduce here a basic branching process (Z_n) called the Galton-Watson process. It models the sizes of a population of particles which reproduce independently of each other. Suppose the population stems from a single particle: $Z_0 = 1$. Denote by X the offspring number for the ancestor and assume that all particles arising in this process reproduce independently with the numbers of offspring having the same distribution as X. Let $\mu = \mathbb{E}(X)$ and always assume that

$$\mathbb{P}(X = 1) < 1.$$

Consider the number of particles Z_n in generation n. Clearly, (Z_n) is a Markov chain with the state space $\{0, 1, 2, \ldots\}$ where 0 is an absorbing state. The sequence $Q_n = \mathbb{P}(Z_n = 0)$ never decreases. Its limit

$$\eta = \lim_{n \to \infty} \mathbb{P}(Z_n = 0)$$

is called the extinction probability of the Galton-Watson process. Using the branching property

$$Z_{n+1} = X_1^{(n)} + \ldots + X_{Z_n}^{(n)},$$

where the $X_j^{(n)}$ are independent copies of X, we obtain $\mathbb{E}(Z_n) = \mu^n$.

Definition 5.18 The Galton-Watson process is called subcritical if $\mu < 1$. It is called critical if $\mu = 1$, and supercritical if $\mu > 1$.

Theorem 5.19 *Put $h(s) = \mathbb{E}(s^X)$. If $\mu \leq 1$, then $\eta = 1$. If $\mu > 1$, then the extinction probability η is the unique solution of the equation $h(x) = x$ in the interval $x \in [0, 1)$.*

Proof. Let $Z_n^{(j)}$, $j = 1, \ldots, X$ be the branching processes stemming from the offspring of the progenitor particle. Then

$$Z_{n+1} = Z_n^{(1)} + \ldots + Z_n^{(X)},$$

and

$$Q_{n+1} = \mathbb{P}(Z_n^{(1)} = 0, \ldots, Z_n^{(X)} = 0) = \mathbb{E}(Q_n^X) = h(Q_n).$$

Letting $n \to \infty$ in the relation $Q_{n+1} = h(Q_n)$ we get $Q = h(Q)$.

Now, if $\mu \leq 1$, then there is only one root of $h(x) = x$ in the interval $x \in [0, 1]$ which is $x = 1$. Thus $\eta = 1$. If $\mu > 1$, then there are two roots of $h(x) = x$ in the interval $x \in [0, 1]$: one of them is $x = 1$ and the other $x = x_0$ is less than 1. To see that $\eta = x_0$ is the smaller root, it is enough to observe that

$$Q_1 = \mathbb{P}(X = 0) = h(0) < x_0,$$

so that, by induction,

$$Q_{n+1} = h(Q_n) \leq h(x_0) < x_0.$$

Theorem 5.20 *If σ^2 stands for the variance of the offspring number X, then*

$$\mathbb{V}ar(Z_n) = \begin{cases} \frac{\sigma^2 \mu^{n-1}(1 - \mu^n)}{1 - \mu} & \text{if } \mu < 1, \\ n\sigma^2 & \text{if } \mu = 1, \\ \frac{\sigma^2 \mu^{n-1}(\mu^n - 1)}{\mu - 1} & \text{if } \mu > 1. \end{cases}$$

Proof. The variance of the generation size $x_n = \mathbb{V}ar(Z_n)$ satisfies iteration

$$x_{n+1} = \mu^n \sigma^2 + \mu^2 x_n, \ x_1 = \sigma^2,$$

because

$$\mathbb{V}ar(Z_{n+1}) = \mathbb{E}(\mathbb{V}ar(Z_{n+1}|Z_n)) + \mathbb{V}ar(\mathbb{E}(Z_{n+1}|Z_n))$$
$$= \mathbb{E}(Z_n \sigma^2) + \mathbb{V}ar(\mu Z_n).$$

Solving this iteration we arrive to the asserted formula.

5.6 Poisson process and continuous-time Markov chains

Definition 5.21 A pure birth process $X(t)$ with intensities $\{\lambda_i\}_{i=0}^{\infty}$
    (i) holds at state i an exponential time with parameter λ_i,
    (ii) after the holding time it jumps up from i to $i+1$.

Exponential holding times has no memory and therefore imply the Markov property in the continuous time setting.

Example 5.22 A Poisson process $N(t)$ with intensity λ is the number of events observed up to time t given that the inter-arrival times are independent exponentials with parameter λ:

$$\mathbb{P}(N(t) = k) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}, \quad k \geq 0.$$

To find the last formula observe that

$$\mathbb{P}(N(t) = k) = \mathbb{P}(T_1 + \ldots + T_k \leq t) - \mathbb{P}(T_1 + \ldots + T_{k+1} \leq t),$$

where $T_1 + \ldots + T_k$ has a gamma distribution with parameters (k, λ) so that

$$\mathbb{P}(T_1 + \ldots + T_k \leq t) = \int_0^t \frac{\lambda^k}{(k-1)!} x^{k-1} e^{-\lambda x} dx = \frac{(\lambda t)^k}{(k-1)!} \int_0^1 y^{k-1} e^{-\lambda t y} dy$$

$$= \frac{(\lambda t)^k}{k!} e^{-\lambda t} + \frac{(\lambda t)^{k+1}}{k!} \int_0^1 y^k e^{-\lambda t y} dy = \frac{(\lambda t)^k}{k!} e^{-\lambda t} + \mathbb{P}(T_1 + \ldots + T_{k+1} \leq t).$$

Explosion: $\mathbb{P}(X(t) = \infty) > 0$ for a finite t. It is possible iff $\sum 1/\lambda_i < \infty$.

Definition 5.23 A continuous-time process $X(t)$ with a countable state space S satisfies the Markov property if

$$\mathbb{P}(X(t_n) = j | X(t_1) = i_1, \ldots, X(t_{n-1}) = i_{n-1}) = \mathbb{P}(X(t_n) = j | X(t_{n-1}) = i_{n-1})$$

for any states $j, i_1, \ldots, i_{n-1} \in S$ and any times $t_1 < \ldots < t_n$.

In the time homogeneous case compared to the discrete time case instead of transition matrices $\mathbf{P}^n$ with elements $p_{ij}^{(n)}$ we have transition matrices $\mathbf{P}_t$ with elements

$$p_{ij}(t) = \mathbb{P}(X(u + t) = j | X(u) = i).$$

Chapman-Kolmogorov: $\mathbf{P}_{t+s} = \mathbf{P}_t \mathbf{P}_s$ for all $t \geq 0$ and $s \geq 0$. Here $\mathbf{P}_0 = \mathbf{I}$ is the identity matrix.

Example 5.24 For the Poisson process we have $p_{ij}(t) = \frac{(\lambda t)^{j-i}}{(j-i)!} e^{-\lambda t}$, and

$$\sum_k p_{ik}(t) p_{kj}(s) = \sum_k \frac{(\lambda t)^{k-i}}{(k-i)!} e^{-\lambda t} \frac{(\lambda s)^{j-k}}{(j-k)!} e^{-\lambda s} = \frac{(\lambda t + \lambda s)^{j-i}}{(j-i)!} e^{-\lambda(t+s)} = p_{ij}(t+s).$$

5.7 The generator of a continuous-time Markov chain

A generator $\mathbf{G} = (g_{ij})$ is a matrix with non-negative off-diagonal elements such that $\sum_j g_{ij} = 0$. A Markov chain $X(t)$ with generator $\mathbf{G}$

- holds at state i an exponential time with parameter $\lambda_i = -g_{ii}$,

- after the holding time it jumps from i to $j \neq i$ with probability $h_{ij} = \frac{g_{ij}}{\lambda_i}$.

The embedded discrete Markov chain is governed by transition matrix $\mathbf{H} = (h_{ij})$ satisfying $h_{ii} = 0$. A continuous-time MC is a discrete MC plus holding intensities (λ_i).

Example 5.25 The Poisson process and birth process have the same embedded MC with $h_{i,i+1} = 1$. For the birth process $g_{ii} = -\lambda_i$, $g_{i,i+1} = \lambda_i$ and all other $g_{ij} = 0$.

Kolmogorov equations. Forward equation: for any $i, j \in S$

$$p'_{ij}(t) = \sum_k p_{ik}(t) g_{kj}$$

or in the matrix form $\mathbf{P}'_t = \mathbf{P}_t \mathbf{G}$. It is obtained from $\mathbf{P}_{t+\epsilon} - \mathbf{P}_t = \mathbf{P}_t(\mathbf{P}_\epsilon - \mathbf{P}_0)$ watching for the last change. Backward equation $\mathbf{P}'_t = \mathbf{G}\mathbf{P}_t$ is obtained from $\mathbf{P}_{t+\epsilon} - \mathbf{P}_t = (\mathbf{P}_\epsilon - \mathbf{P}_0)\mathbf{P}_t$ watching for the initial change. These equations often have a unique solution

$$\mathbf{P}_t = e^{t\mathbf{G}} := \sum_{n=0}^{\infty} \frac{t^n}{n!} \mathbf{G}^n.$$

Theorem 5.26 *Stationary distribution:* $\boldsymbol{\pi}\mathbf{P}_t = \boldsymbol{\pi}$ *for all* t *iff* $\boldsymbol{\pi}\mathbf{G} = \mathbf{0}$.

Proof:

$$\boldsymbol{\pi}\mathbf{P}_t \stackrel{\forall t}{=} \boldsymbol{\pi} \quad \Leftrightarrow \quad \sum_{n=0}^{\infty} \frac{t^n}{n!} \boldsymbol{\pi}\mathbf{G}^n \stackrel{\forall t}{=} \boldsymbol{\pi} \quad \Leftrightarrow \quad \sum_{n=1}^{\infty} \frac{t^n}{n!} \boldsymbol{\pi}\mathbf{G}^n \stackrel{\forall t}{=} \mathbf{0} \quad \Leftrightarrow \quad \boldsymbol{\pi}\mathbf{G}^n \stackrel{\forall n}{=} \mathbf{0}.$$

Example 5.27 Check that the birth process has no stationary distribution.

Theorem 5.28 *Let* $X(t)$ *be irreducible with generator* $\mathbf{G}$. *If there exists a stationary distribution* $\boldsymbol{\pi}$, *then it is unique and for all* (i, j)

$$p_{ij}(t) \to \pi_j, \quad t \to \infty.$$

If there is no stationary distribution, then $p_{ij}(t) \to 0$ *as* $t \to \infty$.

Example 5.29 Poisson process holding times $\lambda_i = \lambda$. Then $\mathbf{G} = \lambda(\mathbf{H} - \mathbf{I})$ and $\mathbf{P}_t = e^{\lambda t(\mathbf{H}-\mathbf{I})}$.

6 Stationary processes

6.1 Weakly and strongly stationary processes

Definition 6.1 The real-valued process $\{X(t), t \geq 0\}$ is called strongly stationary if the vectors $(X(t_1), \ldots, X(t_n))$ and $(X(t_1 + h), \ldots, X(t_n + h))$ have the same joint distribution for all $t_1, \ldots, t_n$ and $h > 0$.

Definition 6.2 The real-valued process $\{X(t), t \geq 0\}$ with $\mathbb{E}(X^2(t)) < \infty$ for all t is called weakly stationary if for all t_1, t_2 and $h > 0$

$$\mathbb{E}(X(t_1)) = \mathbb{E}(X(t_2)), \qquad \mathbb{C}ov(X(t_1), X(t_2)) = \mathbb{C}ov(X(t_1 + h), X(t_2 + h)).$$

Its autocovariance and autocorrelation functions are

$$c(t) = \mathbb{C}ov(X(s), X(s+t)), \qquad \rho(t) = \frac{c(t)}{c(0)}.$$

Example 6.3 Consider an irreducible Markov chain $\{X(t), t \geq 0\}$ with countably many states and a stationary distribution $\boldsymbol{\pi}$ as the initial distribution. This is a strongly stationary process since

$$\mathbb{P}(X(h + t_1) = i_1, X(h + t_1 + t_2) = i_2, \ldots, X(h + t_1 + \ldots + t_n) = i_n) = \pi_{i_1} p_{i_1, i_2}(t_2) \ldots p_{i_{n-1}, i_n}(t_n).$$

Example 6.4 The process $\{X_n, n = 1, 2, \ldots\}$ formed by iid Cauchy r.v is strongly stationary but not a weakly stationary process.

6.2 Linear prediction

Task: knowing the past $(X_r, X_{r-1}, \ldots, X_{r-s})$ predict a future value X_{r+k} by choosing $(a_0, \ldots, a_s)$ that minimize $\mathbb{E}(X_{r+k} - \hat{X}_{r+k})^2$, where

$$\hat{X}_{r+k} = \sum_{j=0}^{s} a_j X_{r-j}. \tag{6}$$

Theorem 6.5 *For a stationary sequence with zero mean and autocovariance $c(m)$ the best linear predictor* (6) *satisfies the equations*

$$\sum_{j=0}^{s} a_j c(|j - m|) = c(k + m), \quad 0 \le m \le s.$$

Proof. Geometrically, the best linear predictor $\hat{X}_{r+k}$ makes an error, $X_{r+k} - \hat{X}_{r+k}$, which is *orthogonal* to the past $(X_r, X_{r-1}, \ldots, X_{r-s})$, meaning that the covariances are equal to zero:

$$\mathbb{E}((X_{r+k} - \hat{X}_{r+k})X_{r-m}) = 0, \quad m = 0, \ldots, s.$$

Plugging (6) into the last relation, we arrive at the claimed equations.

To justify the geometric intuition let H be a linear space generated by $(X_r, X_{r-1}, \ldots, X_{r-s})$. We have to show that if $M \in H$ is such that for all $Z \in H$,

$$\mathbb{E}(X_{r+k} - M)^2 \le \mathbb{E}(X_{r+k} - Z)^2,$$

then

$$\mathbb{E}((X_{r+k} - M)X_{r-m}) = 0, \quad m = 0, \ldots, s.$$

Suppose that $\mathbb{E}((X_{r+k} - M)X_{r-m}) = dc(0)$ is positive for some m. Then writing $M' = M + d \cdot X_{r-m}$ we arrive at a contradiction

$$\begin{aligned}
\mathbb{E}(X_{r+k} - M')^2 &= \mathbb{E}(X_{r+k} - M - d \cdot X_{r-m})^2 \\
&= \mathbb{E}(X_{r+k} - M)^2 - 2d\mathbb{E}(X_{r+k} - M)X_{r-m} + d^2 c(0) \\
&= \mathbb{E}(X_{r+k} - M)^2 - d^2 c(0).
\end{aligned}$$

Example 6.6 AR(1) process Y_n satisfies

$$Y_n = \alpha Y_{n-1} + Z_n, \quad -\infty < n < \infty,$$

where Z_n are independent r.v. with zero means and unit variance. If $|\alpha| < 1$, then $Y_n = \sum_{m \ge 0} \alpha^m Z_{n-m}$ is weakly stationary with zero mean and autocovariance for $m \ge 0$,

$$\begin{aligned}
c(m) = \mathbb{E}(Y_n Y_{n+m}) &= \mathbb{E}(Z_n + Z_{n-1}\alpha + Z_{n-2}\alpha^2 + \ldots)(Z_{n+m} + Z_{n+m-1}\alpha + Z_{n+m-2}\alpha^2 + \ldots) \\
&= \mathbb{E}(Z_n^2)\alpha^m + \mathbb{E}(Z_{n-1}^2)\alpha\alpha^{m+1} + \mathbb{E}(Z_{n-2}^2)\alpha^2\alpha^{m+2} + \ldots \\
&= \alpha^m + \alpha^{m+2} + \alpha^{m+4} + \ldots = \frac{\alpha^m}{1 - \alpha^2}.
\end{aligned}$$

The best linear predictor is $\hat{Y}_{r+k} = \alpha^k Y_r$. This follows from the equations

$$a_0 + a_1\alpha + a_2\alpha^2 + \ldots + a_s\alpha^s = \alpha^k,$$
$$a_0\alpha + a_1 + a_2\alpha + \ldots + a_s\alpha^{s-1} = \alpha^{k+1}.$$

The mean squared error of the best prediction is

$$\mathbb{E}(\hat{Y}_{r+k} - Y_{r+k})^2 = \mathbb{E}(\alpha^k Y_r - Y_{r+k})^2 = \alpha^{2k}c(0) + c(0) - 2\alpha^k c(k) = \frac{1 - \alpha^{2k}}{1 - \alpha^2}.$$

Example 6.7 Let $X_n = (-1)^n X_0$, where X_0 is -1 or 1 equally likely. The best linear predictor is $\hat{X}_{r+k} = (-1)^k X_r$. The mean squared error of prediction is zero.

29

6.3   Linear combination of sinusoids

Example 6.8 For a sequence of fixed frequencies $0 \le \lambda_1 < \ldots < \lambda_k < \infty$ define a continuous time stochastic process by

$$X(t) = \sum_{j=1}^{k}(A_j \cos(\lambda_j t) + B_j \sin(\lambda_j t)),$$

where $A_1, B_1, \ldots, A_k, B_k$ are uncorrelated r.v. with zero means and $\mathbb{V}ar(A_j) = \mathbb{V}ar(B_j) = \sigma_j^2$. Its mean is zero and its autocovariancies are

$$\mathbb{C}ov(X(t), X(s)) = \mathbb{E}(X(t)X(s)) = \sum_{j=1}^{k} \mathbb{E}(A_j^2 \cos(\lambda_j t) \cos(\lambda_j s) + B_j^2 \sin(\lambda_j t) \sin(\lambda_j s))$$

$$= \sum_{j=1}^{k} \sigma_j^2 \cos(\lambda_j(s-t)),$$

$$\mathbb{V}ar(X(t)) = \sum_{j=1}^{k} \sigma_j^2.$$

Thus $X(t)$ is weakly stationary with autocovariance and autocorrelation functions

$$c(t) = \sum_{j=1}^{k} \sigma_j^2 \cos(\lambda_j t), \quad c(0) = \sum_{j=1}^{k} \sigma_j^2,$$

$$\rho(t) = \frac{c(t)}{c(0)} = \sum_{j=1}^{k} g_j \cos(\lambda_j t) = \int_0^{\infty} \cos(\lambda t) dG(\lambda),$$

where G is a distribution function defined as

$$g_j = \frac{\sigma_j^2}{\sigma_1^2 + \ldots + \sigma_k^2}, \qquad G(\lambda) = \sum_{j:\lambda_j \le \lambda} g_j.$$

We can write

$$X(t) = \int_0^{\infty} \cos(t\lambda) dU(\lambda) + \int_0^{\infty} \sin(t\lambda) dV(\lambda),$$

where

$$U(\lambda) = \sum_{j:\lambda_j \le \lambda} A_j, \qquad V(\lambda) = \sum_{j:\lambda_j \le \lambda} B_j.$$

Example 6.9 Let specialize further and put $k = 1$, $\lambda_1 = \frac{\pi}{4}$, assuming that A_1 and B_1 are iid with

$$\mathbb{P}(A_1 = \frac{1}{\sqrt{2}}) = \mathbb{P}(A_1 = -\frac{1}{\sqrt{2}}) = \frac{1}{2}.$$

Then $X(t) = \cos(\frac{\pi}{4}(t + \tau))$ with

$$\mathbb{P}(\tau = 1) = \mathbb{P}(\tau = -1) = \mathbb{P}(\tau = 3) = \mathbb{P}(\tau = -3) = \frac{1}{4}.$$

This stochastic process has only four possible trajectories. This is not a strongly stationary process since

$$\mathbb{E}(X^4(t)) = \frac{1}{2}\Big(\cos^4\Big(\frac{\pi}{4}t + \frac{\pi}{4}\Big) + \sin^4\Big(\frac{\pi}{4}t + \frac{\pi}{4}\Big)\Big) = \frac{1}{4}\Big(2 - \sin^2\Big(\frac{\pi}{2}t + \frac{\pi}{2}\Big)\Big) = \frac{1 + \sin^2(\frac{\pi}{2}t)}{2}.$$

Example 6.10 Put

$$X(t) = \cos(t + Y) = \cos(t)\cos(Y) - \sin(t)\sin(Y),$$

where Y is uniformly distributed over $[0, 2\pi]$. In this case $k = 1, \lambda = 1, \sigma_1^2 = \frac{1}{2}$. What is the distribution of $X(t)$? For an arbitrary bounded measurable function $\phi(x)$ we have

$$\mathbb{E}(\phi(X(t))) = \mathbb{E}(\phi(\cos(t + Y))) = \frac{1}{2\pi} \int_0^{2\pi} \phi(\cos(t + y)) dy = \frac{1}{2\pi} \int_t^{t+2\pi} \phi(\cos(z)) dz = \frac{1}{2\pi} \int_0^{2\pi} \phi(\cos(z)) dz$$

$$= \frac{1}{2\pi} \left(\int_0^\pi \phi(\cos(z)) dz + \int_\pi^{2\pi} \phi(\cos(z)) dz \right) = \frac{1}{2\pi} \left(\int_0^\pi \phi(\cos(\pi - y)) dy + \int_0^\pi \phi(\cos(\pi + y)) dy \right)$$

$$= \frac{1}{\pi} \int_0^\pi \phi(-\cos(y)) dy.$$

The change of variables $x = -\cos(y)$ yields $dx = \sin(y) dy = \sqrt{1 - x^2} dy$, hence

$$\mathbb{E}(\phi(X)) = \frac{1}{\pi} \int_{-1}^1 \frac{\phi(x) dx}{\sqrt{1 - x^2}}.$$

Thus $X(t)$ has the so-called arcsine density $f(x) = \frac{1}{\pi\sqrt{1-x^2}}$ over the interval $[-1, 1]$. Notice that $Z = \frac{X+1}{2}$ has a Beta$(\frac{1}{2}, \frac{1}{2})$ distribution, since

$$\mathbb{E}(\phi(Z)) = \frac{1}{\pi} \int_{-1}^1 \frac{\phi(\frac{x+1}{2}) dx}{\sqrt{1 - x^2}} = \frac{1}{\pi} \int_0^1 \frac{\phi(z) dz}{\sqrt{z(1 - z)}}.$$

This is a strongly stationary process, since $X(t + h) = \cos(t + Y')$, where Y' is uniformly distributed over $[h, 2\pi + h]$, and

$$(X(t_1 + h), \ldots, X(t_n + h)) = (\cos(t_1 + Y'), \ldots, \cos(t_n + Y')) \overset{d}{=} (\cos(t_1 + Y), \ldots, \cos(t_n + Y)).$$

Example 6.11 In the *discrete time setting* for $n \in \mathbb{Z}$ put

$$X_n = \sum_{j=1}^k (A_j \cos(\lambda_j n) + B_j \sin(\lambda_j n)),$$

where $0 \le \lambda_1 < \ldots < \lambda_k \le \pi$ is a set of fixed frequencies, and again, $A_1, B_1, \ldots, A_k, B_k$ are uncorrelated r.v. with zero means and $\mathbb{V}ar(A_j) = \mathbb{V}ar(B_j) = \sigma_j^2$. Similarly to the continuous time case we get

$$\mathbb{E}(X_n) = 0, \quad c(n) = \sum_{j=1}^k \sigma_j^2 \cos(\lambda_j n), \quad \rho(n) = \int_0^\pi \cos(\lambda n) dG(\lambda),$$

$$X_n = \int_0^\pi \cos(n\lambda) dU(\lambda) + \int_0^\pi \sin(n\lambda) dV(\lambda).$$

6.4 The spectral representation

Any weakly stationary process $\{X(t) : -\infty < t < \infty\}$ with zero mean can be approximated by a linear combination of sinusoids. Indeed, its autocovariance function $c(t)$ is non-negative definite since for any $t_1, \ldots, t_n$ and $z_1, \ldots, z_n$

$$\sum_{j=1}^n \sum_{k=1}^n c(t_k - t_j) z_j z_k = \mathbb{V}ar\left(\sum_{k=1}^n z_k X(t_k) \right) \ge 0.$$

Thus due to the Bochner theorem, given that $c(t)$ is continuous at zero, there is a probability distribution function G on $[0, \infty)$ such that

$$\rho(t) = \int_0^\infty \cos(t\lambda) dG(\lambda).$$

In the discrete time case there is a probability distribution function G on $[0, \pi]$ such that

$$\rho(n) = \int_0^\pi \cos(n\lambda) dG(\lambda).$$

Definition 6.12 The function G is called the spectral distribution function of the corresponding stationary random process, and the set of real numbers λ such that

$$G(\lambda + \epsilon) - G(\lambda - \epsilon) > 0 \text{ for all } \epsilon > 0$$

is called the spectrum of the random process. If G has density g it is called the spectral density function.

In the discrete time case, if

$$\sum_{n=-\infty}^{\infty} |\rho(n)| < \infty,$$

then the spectral density exists and is given by

$$g(\lambda) = \frac{1}{\pi} + \frac{2}{\pi} \sum_{n=1}^{\infty} \rho(n)\cos(n\lambda), \quad 0 \le \lambda \le \pi.$$

Example 6.13 Consider an irreducible continuous time Markov chain $\{X(t), t \ge 0\}$ with two states $\{1, 2\}$ and generator

$$\mathbf{G} = \begin{pmatrix} -\alpha & \alpha \\ \beta & -\beta \end{pmatrix}.$$

Its stationary distribution is $\boldsymbol{\pi} = (\frac{\beta}{\alpha+\beta}, \frac{\alpha}{\alpha+\beta})$ will be taken as the initial distribution. From

$$\begin{pmatrix} p_{11}(t) & p_{12}(t) \\ p_{21}(t) & p_{22}(t) \end{pmatrix} = e^{t\mathbf{G}} = \sum_{n=0}^{\infty} \frac{t^n}{n!}\mathbf{G}^n = \mathbf{I} + \mathbf{G}\sum_{n=1}^{\infty}\frac{t^n}{n!}(-\alpha-\beta)^{n-1} = \mathbf{I} + (\alpha+\beta)^{-1}(1 - e^{-t(\alpha+\beta)})\mathbf{G}$$

we see that

$$p_{11}(t) = 1 - p_{12}(t) = \frac{\beta}{\alpha+\beta} + \frac{\alpha}{\alpha+\beta}e^{-t(\alpha+\beta)},$$

$$p_{22}(t) = 1 - p_{21}(t) = \frac{\alpha}{\alpha+\beta} + \frac{\beta}{\alpha+\beta}e^{-t(\alpha+\beta)},$$

and we find for $t \ge 0$

$$c(t) = \frac{\alpha\beta}{(\alpha+\beta)^2}e^{-t(\alpha+\beta)}, \qquad \rho(t) = e^{-t(\alpha+\beta)}.$$

Thus this process has a spectral density corresponding to a one-sided Cauchy distribution:

$$g(\lambda) = \frac{2(\alpha+\beta)}{\pi((\alpha+\beta)^2 + \lambda^2)}, \quad \lambda \ge 0.$$

Example 6.14 Discrete white noise: a sequence $X_0, X_1, \ldots$ of independent r.v. with zero means and unit variances. This stationary sequence has the uniform spectral density:

$$\rho(n) = 1_{\{n=0\}} = \pi^{-1}\int_0^{\pi} \cos(n\lambda)d\lambda.$$

Theorem 6.15 *If $\{X(t) : -\infty < t < \infty\}$ is a weakly stationary process with zero mean, unit variance, continuous autocorrelation function and spectral distribution function G, then there exists a pair of orthogonal zero mean random process $(U(\lambda), V(\lambda))$ with uncorrelated increments such that*

$$X(t) = \int_0^{\infty} \cos(t\lambda)dU(\lambda) + \int_0^{\infty} \sin(t\lambda)dV(\lambda)$$

and $\mathbb{V}ar(U(\lambda)) = \mathbb{V}ar(V(\lambda)) = G(\lambda)$.

Theorem 6.16 *If $\{X_n : -\infty < n < \infty\}$ is a discrete-time weakly stationary process with zero mean, unit variance, and spectral distribution function G, then there exists a pair of orthogonal zero mean random process $(U(\lambda), V(\lambda))$ with uncorrelated increments such that*

$$X_n = \int_0^{\pi} \cos(n\lambda)dU(\lambda) + \int_0^{\pi} \sin(n\lambda)dV(\lambda)$$

and $\mathbb{V}ar(U(\lambda)) = \mathbb{V}ar(V(\lambda)) = G(\lambda)$.

6.5 Stochastic integral

Let $\{S(t) : t \in \mathbb{R}\}$ be a complex-valued process on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that

- $\mathbb{E}(|S(t)|^2) < \infty$ for all t,

- $\mathbb{E}(|S(t+h) - S(t)|^2) \to 0$ as $h \searrow 0$ for all t,

- orthogonal increments: $\mathbb{E}([S(v) - S(u)][\bar{S}(t) - \bar{S}(s)]) = 0$ whenever $u < v \le s < t$.

Put

$$F(t) := \left\{ \begin{array}{ll} \mathbb{E}(|S(t) - S(0)|^2), & \text{if } t \ge 0, \\ -\mathbb{E}(|S(t) - S(0)|^2), & \text{if } t < 0. \end{array} \right.$$

Since the process has orthogonal increments we obtain

$$\mathbb{E}(|S(t) - S(s)|^2) = F(t) - F(s), \quad s < t \tag{7}$$

implying that F is monotonic and right-continuous.

Example 6.17 The Wiener process $S(t) = W(t)$ with $F(t) = t$ is an example of such a process with orthogonal increments.

Let $\psi : \mathbb{R} \to \mathbb{C}$ be a measurable complex-valued function for which

$$\int_{-\infty}^{\infty} |\psi(t)|^2 dF(t) < \infty.$$

Next comes a two-step definition of a stochastic integral of ψ with respect to S,

$$I(\psi) = \int_{-\infty}^{\infty} \psi(t) dS(t),$$

possessing the following important property

$$\mathbb{E}(I(\psi_1) I(\psi_2)) = \int_{-\infty}^{\infty} \psi_1(t) \psi_2(t) dF(t). \tag{8}$$

1. For an arbitrary step function

$$\phi(t) = \sum_{j=1}^{n-1} c_j 1_{\{a_j \le t < a_{j+1}\}}, \quad -\infty < a_1 < \ldots < a_n < \infty$$

   put

$$I(\phi) := \sum_{j=1}^{n-1} c_j (S(a_{j+1}) - S(a_j)).$$

   Due to orthogonality of increments we obtain (8) and find that "integration is distance preserving"

$$\mathbb{E}(|I(\phi_1) - I(\phi_2)|^2) = \mathbb{E}((I(\phi_1 - \phi_2))^2) = \int_{-\infty}^{\infty} |\phi_1 - \phi_2|^2 dF(t).$$

2. There exists a sequence of step functions such that

$$\|\phi_n - \psi\| := \left(\int_{-\infty}^{\infty} |\phi_n - \psi|^2 dF(t) \right)^{1/2} \to 0.$$

   Thus $I(\phi_n)$ is a mean-square Cauchy sequence and there exists a mean-square limit $I(\phi_n) \to I(\psi)$.

A sketch of the proof of Theorem 6.16 for the complex-valued processes.

Step 1. Let H_X be the set of all r.v of the form $\sum_{j=1}^{n} a_j X_{m_j}$ for $a_1, a_2, \ldots \in \mathbb{C}$, $n \in \mathbb{N}$, $m_1, m_2, \ldots \in \mathbb{Z}$. Similarly, let H_F be the set of linear combinations of sinusoids $f_n(x) := e^{inx}$. Define the linear mapping $\mu : H_F \to H_X$ by $\mu(f_n) := X_n$.

Step 2. The closure $\overline{H}_X$ of H_X is defined to be the space H_X together with all limits of mean-square Cauchy-convergent sequences in H_X. Define the closure $\overline{H}_F$ of H_F as the space H_F together with all limits of Cauchy-convergent sequences $u_n \in H_F$, with the latter meaning by definition that

$$\int_{(-\pi,\pi]} (u_n(\lambda) - u_m(\lambda))(\overline{u_n(\lambda) - u_m(\lambda)})dF(\lambda) \to 0, \qquad n, m \to \infty.$$

For $u = \lim u_n$, where $u_n \in H_F$, define $\mu(u) = \lim \mu(u_n)$ thereby defining a mapping $\mu : \overline{H}_F \to \overline{H}_X$.

Step 3. Define the process $S(\lambda)$ by

$$S(\lambda) = \mu(h_\lambda), \quad -\pi < \lambda \leq \pi, \quad h_\lambda(x) := 1_{\{x \in (-\pi, \lambda]\}}$$

and show that it has orthogonal increments and satisfies (7). Prove that

$$\mu(\psi) = \int_{(-\pi,\pi]} \psi(t)dS(t)$$

first for step-functions and then for $\psi(x) = e^{inx}$. It follows that

$$X_n = \int_{(-\pi,\pi]} e^{int}dS(t).$$

The case of real-valued processes.

For the sequence X_n to be real, the following representation

$$\begin{aligned} X_n &= \int_{(-\pi,\pi]} e^{in\lambda}dS(\lambda) = \int_{(0,\pi]} e^{in\lambda}dS(\lambda) + \int_{[0,\pi)} e^{-in\lambda}dS(-\lambda) \\ &= \int_{[0,\pi]} \cos(n\lambda)(dS(\lambda) + dS(-\lambda)) + i \int_{[0,\pi]} \sin(n\lambda)(dS(\lambda) - dS(-\lambda)) \\ &= \int_{[0,\pi]} \cos(n\lambda)dU(\lambda) + \int_{[0,\pi]} \sin(n\lambda)dV(\lambda) \end{aligned}$$

must hold for real processes $U(\lambda)$ and $V(\lambda)$ such that

$$\begin{aligned} dU(\lambda) &= dS(\lambda) + dS(-\lambda), \\ dV(\lambda) &= i(dS(\lambda) - dS(-\lambda)). \end{aligned}$$

This implies that $dS(\lambda) + dS(-\lambda)$ is purely real and $dS(\lambda) - dS(-\lambda)$ is purely imaginary, which implies a symmetry property: $dS(\lambda) = \overline{dS(-\lambda)}$.

Exercise 6.18 Check the properties of $U(\lambda)$ and $V(\lambda)$ stated in Theorem 6.16 with $dG(\lambda) = 2dF(\lambda)$ for $\lambda > 0$ and $G(0) - G(0-) = F(0) - F(0-)$.

6.6 The ergodic theorem for the weakly stationary processes

Theorem 6.19 *Let $\{X_n, n = 1, 2, \ldots\}$ be a weakly stationary process with mean μ and autocovariance function $c(m)$. There exists a r.v. Y with mean μ and variance*

$$\mathrm{Var}(Y) = \lim_{n \to \infty} n^{-1} \sum_{j=1}^{n} c(j) = c(0)(G(0) - G(0-)),$$

such that

$$\frac{X_1 + \ldots + X_n}{n} \xrightarrow{L^2} Y, \qquad n \to \infty.$$

Sketch proof. Suppose that $\mu = 0$ and $c(0) = 1$, then using a spectral representation

$$X_n = \int_0^\pi \cos(n\lambda)dU(\lambda) + \int_0^\pi \sin(n\lambda)dV(\lambda),$$

we get

$$\bar{X}_n := \frac{X_1 + \ldots + X_n}{n} = \int_0^\pi g_n(\lambda)dU(\lambda) + \int_0^\pi h_n(\lambda)dV(\lambda),$$

where in terms of complex numbers

$$g_n(\lambda) + ih_n(\lambda) = \frac{e^{i\lambda} + \ldots + e^{in\lambda}}{n}.$$

which equals 1 for $\lambda = 0$. For $\lambda \neq 0$ we have

$$g_n(\lambda) + ih_n(\lambda) = \frac{e^{i\lambda} - e^{in\lambda}}{n(1 - e^{i\lambda})} = \frac{(e^{i\lambda} - e^{in\lambda})(1 - e^{-i\lambda})}{2n(1 - \cos(\lambda))}$$
$$= \frac{e^{i\lambda} - 1 + e^{i(n-1)\lambda} - e^{-in\lambda}}{4n\sin^2(\lambda/2)},$$

so that

$$g_n(\lambda) = \frac{\cos(\lambda) - 1 + \cos((n-1)\lambda) - \cos(n\lambda)}{4n\sin^2(\lambda/2)} = \frac{\sin((n-1)\lambda/2) - \sin(\lambda/2)}{2n\sin(\lambda/2)},$$
$$h_n(\lambda) = \frac{\sin(\lambda) + \sin((n-1)\lambda) - \sin(n\lambda)}{4n\sin^2(\lambda/2)} = \frac{\cos(\lambda/2) - \cos((n-1)\lambda/2)}{2n\sin(\lambda/2)}.$$

If $\lambda \in [0, \pi]$, then $|g_n(\lambda)| \leq 1$, $|h_n(\lambda)| \leq 1$, and $g_n(\lambda) \to 1_{\{\lambda=0\}}$, $h_n(\lambda) \to 0$ as $n \to \infty$. It can be shown that

$$\int_0^\pi g_n(\lambda)dU(\lambda) \overset{L^2}{\to} \int_0^\pi 1_{\{\lambda=0\}}dU(\lambda) = U(0) - U(0-),$$
$$\int_0^\pi h_n(\lambda)dV(\lambda) \overset{L^2}{\to} 0.$$

Thus $\bar{X}_n \overset{L^2}{\to} Y := U(0) - U(0-)$ and it remains to find the mean and variance of Y.

6.7   The ergodic theorem for the strongly stationary processes

Let $\{X_n, n = 1, 2, \ldots\}$ be a strongly stationary process defined on $(\Omega, \mathcal{F}, \mathbb{P})$. The vector $(X_1, X_2, \ldots)$ takes values in $\mathbb{R}^T$, where $T = \{1, 2, \ldots\}$. The natural σ-algebra for $\mathbb{R}^T$ is the product $\mathcal{B}^T$ of the appropriate number of copies of the Borel σ-algebra $\mathcal{B}$ of subsets of $\mathbb{R}$.

Definition 6.20 Consider a shift operator $\tau : \mathbb{R}^T \to \mathbb{R}^T$ such that $\tau(x_1, x_2, \ldots) = (x_2, x_3, \ldots)$. A set $B \in \mathcal{B}^T$ is called *invariant* with respect to the shift operator τ, if $B = \tau^{-1}B$. An event $A \in \mathcal{F}$ is called invariant, if

$$A = \{\omega : (X_1, X_2, \ldots) \in B\}$$

for some invariant set $B \in \mathcal{B}^T$. The collection of all invariant events forms a σ-algebra and denoted $\mathcal{I}$. The strictly stationary process is called *ergodic*, if either $\mathbb{P}(A) = 0$ or $\mathbb{P}(A) = 1$ for any $A \in \mathcal{I}$.

Example 6.21 Consider a reducible Markov chain $\{X_n, n = 0, 1, \ldots\}$ with two states $(0, 1)$ and transition probabilities $p_{00} = p_{11} = 1$, $p_{01} = p_{10} = 0$. The invariant events $X_0 = 0$ and $X_0 = 1$ can have probabilities π_0 and $\pi_1 = 1 - \pi_0$ for any given $\pi_0 \in (0, 1)$. This is an example of a non-ergodic case.

Example 6.22 Any invariant event is a tail event, so that $\mathcal{I} \subset \mathcal{T}$. If $\{X_n, n = 1, 2, \ldots\}$ are iid with a finite mean, then according to Kolmogorov's zero-one law such a stationary process is ergodic. The classical LLN, the sufficiency part of Theorem 4.9, follows from the next ergodic theorem.

Theorem 6.23 *If $\{X_n, n = 1, 2, \ldots\}$ is a strongly stationary sequence with a finite mean, then*

$$\bar{X}_n := \frac{X_1 + \ldots + X_n}{n} \to \mathbb{E}(X_1|\mathcal{I}) \ a.s. \ and \ in \ L^1.$$

In the ergodic case

$$\bar{X}_n \to \mathbb{E}(X_1) \ a.s. \ and \ in \ L^1.$$

PROOF IN THE ERGODIC CASE. To prove the a.s. convergence it suffices to show that

$$\text{if } \mathbb{E}(X_1) < 0, \text{ then } \limsup_{n \to \infty} \bar{X}_n \le 0 \text{ a.s.} \tag{9}$$

Indeed, applying (9) to $X'_n = X_n - \mathbb{E}(X_1) - \epsilon$ and $X''_n = \mathbb{E}(X_1) - X_n - \epsilon$, we obtain

$$\mathbb{E}(X_1) - \epsilon \le \liminf_{n \to \infty} \bar{X}_n \le \limsup_{n \to \infty} \bar{X}_n \le \mathbb{E}(X_1) + \epsilon.$$

To prove (9) assume $\mathbb{E}(X_1) < 0$ and put

$$M_n := \max\{0, X_1, X_1 + X_2, \ldots, X_1 + \ldots + X_n\}.$$

Clearly $\bar{X}_n \le M_n/n$, and it is enough to show that $\mathbb{P}(M_\infty < \infty) = 1$. Suppose the latter is not true. Since $\{M_\infty < \infty\}$ is an invariant event, we get $\mathbb{P}(M_\infty < \infty) = 0$ or in other words $M_n \nearrow \infty$ a.s. To arrive to a contradiction observe that

$$M_{n+1} = \max\{0, X_1 + M'_n\} = M'_n + \max\{-M'_n, X_1\},$$

where

$$M'_n := \max\{0, X_2, X_2 + X_3, \ldots, X_2 + \ldots + X_{n+1}\}$$

has the same distribution as M_n. Since $\mathbb{E}(M_{n+1}) \ge \mathbb{E}(M_n) = \mathbb{E}(M'_n)$, it follows that

$$\mathbb{E}(\max\{-M'_n, X_1\}) \ge 0.$$

This contradicts the assumption $\mathbb{E}(X_1) < 0$ as due to the monotone convergence Theorem 3.18

$$\mathbb{E}(\max\{-M'_n, X_1\}) \to \mathbb{E}(X_1).$$

Thus almost-sure convergence is proved.

To prove the convergence in mean we apply Theorem 3.25. It suffices to verify that the family $\{\bar{X}_n\}_{n \ge 1}$ is uniformly integrable, that is for all $\epsilon > 0$, there is $\delta > 0$ such that, for all n, $\mathbb{E}(|\bar{X}_n|1_A) < \epsilon$ for any event A such that $\mathbb{P}(A) < \delta$. This follows from

$$\mathbb{E}(|\bar{X}_n|1_A) \le n^{-1} \sum_{i=1}^{n} \mathbb{E}(|X_i|1_A),$$

and the next lemma.

Lemma 6.24 *If $X_1, X_2, \ldots$ have a common distribution with a finite mean, then for all $\epsilon > 0$, there is $\delta > 0$ such that $\mathbb{E}(|X_i|1_A) < \epsilon$ for all i and for any event A such that $\mathbb{P}(A) < \delta$.*

Proof. Put $R(y) = \mathbb{E}(|X_i|1_{\{|X_i|>y\}})$. Since $R(y) \to 0$ as $y \to \infty$, for a given ϵ we can find t such that $R(t) < \epsilon/2$. Writing $\delta = \epsilon/(2t)$, consider any event A such that $\mathbb{P}(A) < \delta$. Denote $B_i = \{|X_i| > t\}$. Then

$$\mathbb{E}(|X_i|1_A) \le \mathbb{E}(|X_i|1_{A \cap B_i^c}) + \mathbb{E}(|X_i|1_{B_i}) \le t\mathbb{P}(A) + R(t) < \epsilon.$$

Example 6.25 Let $Z_1, \ldots, Z_k$ be iid with a finite mean μ. Then the following cyclic process

$$X_1 = Z_1, \ldots, X_k = Z_k,$$
$$X_{k+1} = Z_1, \ldots, X_{2k} = Z_k,$$
$$X_{2k+1} = Z_1, \ldots, X_{3k} = Z_k, \ldots,$$

is a strongly stationary process. The corresponding limit in the ergodic theorem is not the constant μ like in the strong LLN but rather a random variable

$$\frac{X_1 + \ldots + X_n}{n} \to \frac{Z_1 + \ldots + Z_k}{k}.$$

Example 6.26 Let $\{X_n, n = 1, 2, \ldots\}$ be an irreducible positive-recurrent Markov chain with the state space $S = \{0, \pm 1, \pm 2, \ldots\}$. Let $\boldsymbol{\pi} = (\pi_j)_{j \in S}$ be the unique stationary distribution. If X_0 has distribution $\boldsymbol{\pi}$, then X_n is strongly stationary.

For a fixed state $k \in S$ let $I_n = 1_{\{X_n = k\}}$. The strongly stationary process I_n has autocovariance function
$$c(m) = \mathbb{C}ov(I_n, I_{n+m}) = \mathbb{E}(I_n I_{n+m}) - \pi_k^2 = \pi_k(p_{kk}^{(m)} - \pi_k).$$

Since $p_{kk}^{(m)} \to \pi_k$ as $m \to \infty$ we have $c(m) \to 0$ and the limit in Theorem 6.23 has zero variance. It follows that $n^{-1}(I_1 + \ldots + I_n)$, the proportion of $(X_1, \ldots, X_n)$ visiting state k, converges to π_k as $n \to \infty$.

Example 6.27 Binary expansion. Let X be uniformly distributed on $[0, 1]$ and has a binary expansion $X = \sum_{j=1}^\infty X_j 2^{-j}$. Put $Y_n = \sum_{j=n}^\infty X_j 2^{n-j-1}$ so that $Y_1 = X$ and $Y_{n+1} = (2^n X) \bmod 1$. From

$$\mathbb{E}(Y_1 Y_{n+1}) = \sum_{j=0}^{2^n - 1} \int_{j2^{-n}}^{(j+1)2^{-n}} x(2^n x - j) dx = 2^{-2n} \sum_{j=0}^{2^n-1} \int_0^1 (y+j) y \, dy = 2^{-2n} \sum_{j=0}^{2^n-1} \left(\frac{1}{3} + \frac{j}{2}\right) = \frac{1}{4} + \frac{2^{-n}}{12}$$

we get $c(n) = \frac{2^{-n}}{12}$ implying that $n^{-1} \sum_{j=1}^n Y_j \to 1/2$ almost surely.

6.8 Gaussian processes

Definition 6.28 A random process $\{X(t), t \geq 0\}$ is called Gaussian if for any $(t_1, \ldots, t_n)$ the vector $(X(t_1), \ldots, X(t_n))$ has a multivariate normal distribution.

A Gaussian random process is strongly stationary iff it is weakly stationary.

Theorem 6.29 *A Gaussian process $\{X(t), t \geq 0\}$ is Markov iff for any $0 \leq t_1 < \ldots < t_n$*
$$\mathbb{E}(X(t_n)|X(t_1), \ldots, X(t_{n-1})) = \mathbb{E}(X(t_n)|X(t_{n-1})). \tag{10}$$

Proof. Clearly, the Markov property implies (10). To prove the converse we have to show that in the Gaussian case (10) gives
$$\mathbb{V}ar(X(t_n)|X(t_1), \ldots, X(t_{n-1})) = \mathbb{V}ar(X(t_n)|X(t_{n-1})).$$

Indeed, since $X(t_n) - \mathbb{E}\{X(t_n)|X(t_1), \ldots, X(t_{n-1})\}$ is orthogonal to $(X(t_1), \ldots, X(t_{n-1}))$, which in the Gaussian case means independence, we have

$$\mathbb{E}\left\{\left(X(t_n) - \mathbb{E}\{X(t_n)|X(t_1), \ldots, X(t_{n-1})\}\right)^2 | X(t_1), \ldots, X(t_{n-1})\right\}$$
$$= \mathbb{E}\left\{\left(X(t_n) - \mathbb{E}\{X(t_n)|X(t_1), \ldots, X(t_{n-1})\}\right)^2\right\} = \mathbb{E}\left\{\left(X(t_n) - \mathbb{E}\{X(t_n)|X(t_{n-1})\}\right)^2\right\}$$
$$= \mathbb{E}\left\{\left(X(t_n) - \mathbb{E}\{X(t_n)|X(t_{n-1})\}\right)^2 | X(t_{n-1})\right\}.$$

Example 6.30 A stationary Gaussian Markov process is called the Ornstein-Uhlenbeck process. It is characterized by the auto-correlation function $\rho(t) = e^{-\alpha t}$, $t \geq 0$ with a positive α. This follows from the equation $\rho(t+s) = \rho(t)\rho(s)$ which is obtained as follows. From the property of the bivariate normal distribution
$$\mathbb{E}(X(t+s)|X(s)) = \theta + \rho(t)(X(s) - \theta)$$

we derive

$$\rho(t+s) = c(0)^{-1} \mathbb{E}((X(t+s) - \theta)(X(0) - \theta)) = c(0)^{-1} \mathbb{E}\{\mathbb{E}((X(t+s) - \theta)(X(0) - \theta)|X(0), X(s))\}$$
$$= \rho(t) c(0)^{-1} \mathbb{E}((X(s) - \theta)(X(0) - \theta))$$
$$= \rho(t)\rho(s).$$

7 Renewal theory and Queues

7.1 Renewal function and excess life

Let $T_0 = 0$, $T_n = X_1 + \ldots + X_n$, where X_i are iid strictly positive random variables called inter-arrival times. A renewal process $N(t)$ gives the number of renewal events during the time interval $(0, t]$

$$\{N(t) \geq n\} = \{T_n \leq t\}, \quad T_{N(t)} \leq t < T_{N(t)+1}.$$

Put $F(t) = \mathbb{P}(X_1 \leq t)$, $\mu = \mathbb{E}(X_1)$, and define convolutions

$$F^{*0}(t) = 1_{\{t \geq 0\}}, \quad F^{*k}(t) = \int_0^t F^{*(k-1)}(t - u)dF(u).$$

Then $\mathbb{P}(T_k \leq t) = F^{*k}(t)$.

Exercise 7.1 Show that

$$\{N(t) \leq n\} = \{T_{n+1} > t\}.$$

Definition 7.2 Put $m(t) := \mathbb{E}(N(t))$. We will call $U(t) = 1 + m(t)$ the renewal function.

Lemma 7.3 *We have*

$$U(t) = \sum_{k=0}^{\infty} F^{*k}(t),$$

and in terms of the Laplace-Stieltjes transforms $\hat{F}(\theta) := \int_0^\infty e^{-\theta t}dF(t)$

$$\hat{U}(\theta) = \frac{1}{1 - \hat{F}(\theta)}.$$

Proof. Using the recursion

$$N(t) = 1_{\{X_1 \leq t\}}(1 + \tilde{N}(t - X_1)),$$

where $\tilde{N}(t)$ is a renewed copy of the initial process $N(t)$, we find after taking expectations

$$m(t) = F(t) + \int_0^t m(t - u)dF(u).$$

It follows

$$m(t) = F(t) + F^{*2}(t) + \int_0^t m(t - u)dF^{*2}(u) = \ldots = \sum_{k=1}^{\infty} F^{*k}(t),$$

and consequently

$$\hat{m}(\theta) = \int_0^\infty e^{-\theta t}dm(t) = \sum_{k=1}^{\infty} \int_0^\infty e^{-\theta t}dF^{*k}(t) = \sum_{k=1}^{\infty} \hat{F}(\theta)^k = \frac{\hat{F}(\theta)}{1 - \hat{F}(\theta)}.$$

Example 7.4 Poisson process is the Markovian renewal process with exponentially distributed inter-arrival times. Since $F(t) = 1 - e^{-\lambda t}$, we find $\hat{m}(\theta) = \frac{\lambda}{\theta}$, implying $m(t) = \lambda t$. Notice that $\mu = 1/\lambda$ and the rate is $\lambda = 1/\mu$.

Definition 7.5 For a measurable function $g(t)$ the relation

$$A(t) = g(t) + \int_0^t A(t - u)dF(u)$$

is called a renewal equation. Clearly, its solution for $t \geq 0$ takes the form $A(t) = \int_0^t g(t - u)dU(u)$.

Definition 7.6 The excess lifetime at t is $E(t) = T_{N(t)+1} - t$. The current lifetime (or age) at t is $C(t) = t - T_{N(t)}$. The total lifetime at t is the sum $C(t) + E(t)$.

Notice that $\{E(t), 0 \leq t < \infty\}$ is a Markov process of deterministic unit speed descent toward zero with upward jumps when the process hits zero.

Lemma 7.7 *The distribution of the excess life $E(t)$ is given by*

$$\mathbb{P}(E(t) > y) = \int_0^t (1 - F(t + y - u))dU(u).$$

The distribution of the age $C(t)$ is given by

$$\mathbb{P}(C(t) \geq y) = 1_{\{y \leq t\}} \int_0^{t-y} (1 - F(t - u))dU(u).$$

In particular, $\mathbb{P}(C(t) = t) = 1 - F(t)$.

Proof. The first claim follows from the following renewal equation for $b(t) := \mathbb{P}(E(t) > y)$

$$b(t) = \mathbb{E}\Big(\mathbb{E}(1_{\{E(t)>y\}}|X_1)\Big) = \mathbb{E}\Big(1_{\{\tilde{E}(t-X_1)>y\}}1_{\{X_1 \leq t\}} + 1_{\{X_1 > t+y\}}\Big)$$

$$= 1 - F(t + y) + \int_0^t b(t - x)dF(x).$$

It is the case that $C(t) \geq y$ if and only if there are no arrivals in $(t - y, t]$. Thus

$$\mathbb{P}(C(t) \geq y) = \mathbb{P}(E(t - y) > y)$$

and the second claim follows from the first one.

Example 7.8 Return to the Poisson process with rate is λ. In this case

$$\mathbb{P}(E(t) > y) = \int_0^t (1 - F(t + y - u))dU(u) = e^{-\lambda(t+y)} + \lambda \int_0^t e^{-\lambda(t+y-u)}du = e^{-\lambda y}$$

is independent of t. Therefore, for $y \leq t$, we have

$$\mathbb{P}(C(t) \leq y) = 1 - e^{-\lambda y},\ 0 \leq y < t, \qquad \mathbb{P}(C(t) = t) = e^{-\lambda t},$$

implying that the total lifetime has the mean

$$\mathbb{E}(C(t) + E(t)) = 2\mu - \lambda \int_t^\infty (y - t)e^{-\lambda y}dy,$$

which for large t is twice as large as the inter-arrival mean $\mu = 1/\lambda$. This observation is called the "waiting time paradox".

7.2   LLN and CLT for the renewal process

Theorem 7.9 *Law of large numbers: $N(t)/t \overset{a.s.}{\to} 1/\mu$ as $t \to \infty$. So that $1/\mu$ gives the rate of occurrence of the renewal events.*

Proof. Note that $T_{N(t)} \leq t < T_{N(t)+1}$. Thus, if $N(t) > 0$,

$$\frac{T_{N(t)}}{N(t)} \leq \frac{t}{N(t)} < \frac{T_{N(t)+1}}{N(t)+1}\Big(1 + \frac{1}{N(t)}\Big).$$

Since $N(t) \to \infty$ as $t \to \infty$, it remains to apply Theorem 4.9, the classical law of large numbers, saying that $T_n/n \to \mu$ almost surely.

Theorem 7.10 *Central limit theorem. If $\sigma^2 = \mathbb{V}ar(X_1)$ is positive and finite, then*

$$\frac{N(t) - t/\mu}{\sqrt{t\sigma^2/\mu^3}} \overset{d}{\to} N(0, 1)\ as\ t \to \infty.$$

Proof. The usual CLT implies that for any x,

$$\mathbb{P}\Big(\frac{T_{a(t)} - \mu a(t)}{\sigma\sqrt{a(t)}} \le x\Big) \to \Phi(x) \text{ as } a(t) \to \infty.$$

For a given x, put $a(t) = \lfloor t/\mu + x\sqrt{t\sigma^2/\mu^3}\rfloor$, and observe that on one hand

$$\mathbb{P}\Big(\frac{N(t) - t/\mu}{\sqrt{t\sigma^2/\mu^3}} \ge x\Big) = \mathbb{P}(N(t) \ge a(t)) = \mathbb{P}(T_{a(t)} \le t),$$

and on the other hand, $\frac{t - \mu a(t)}{\sigma\sqrt{a(t)}} \to -x$ as $t \to \infty$. We conclude

$$\mathbb{P}\Big(\frac{N(t) - t/\mu}{\sqrt{t\sigma^2/\mu^3}} \ge x\Big) \to \Phi(-x) = 1 - \Phi(x).$$

7.3 Stopping times and Wald's equation

Definition 7.11 Let M be a r.v. taking values in the set $\{1, 2, \ldots\}$. We call it a stopping time with respect to the sequence (X_n) of random variables, if

$$\{M \le m\} \in \mathcal{F}_m, \quad \text{for all } m = 1, 2, \ldots,$$

where $\mathcal{F}_m = \sigma\{X_1, \ldots, X_m\}$ is the σ-algebra of events generated by the events $\{X_1 \le c_1\}, \ldots, \{X_m \le c_m\}$ for all $c_i \in (-\infty, \infty)$.

Theorem 7.12 *Wald's equation. Let $X_1, X_2, \ldots$ be iid r.v. with finite mean μ, and let M be a stopping time with respect to the sequence (X_n) such that $\mathbb{E}(M) < \infty$. Then*

$$\mathbb{E}(X_1 + \ldots + X_M) = \mu\mathbb{E}(M).$$

Proof. Observe that

$$\sum_{i=1}^{n} X_i 1_{\{M \ge i\}} \overset{\text{a.s.}}{\to} \sum_{i=1}^{\infty} X_i 1_{\{M \ge i\}} = X_1 + \ldots + X_M$$

and for all n,

$$|\sum_{i=1}^{n} X_i 1_{\{M \ge i\}}| \le Y, \qquad Y = \sum_{i=1}^{\infty} |X_i| 1_{\{M \ge i\}},$$

where

$$\mathbb{E}(Y) = \sum_{i=1}^{\infty} \mathbb{E}(|X_i| 1_{\{M \ge i\}}) = \sum_{i=1}^{\infty} \mathbb{E}(|X_i|)\mathbb{P}(M \ge i) = \mathbb{E}(|X_1|)\mathbb{E}(M).$$

In the second equality we used independence between $\{M \ge i\}$ and X_i, which follows from the fact that $\{M \ge i\}$ is the complimentary event to $\{M \le i-1\} \in \sigma\{X_1, \ldots, X_{i-1}\}$. By the dominated convergence Theorem 3.21

$$\mathbb{E}(X_1 + \ldots + X_M) = \lim_{n \to \infty} \mathbb{E}\Big(\sum_{i=1}^{n} X_i 1_{\{M \ge i\}}\Big) = \sum_{i=1}^{\infty} \mathbb{E}(X_i)\mathbb{P}(M \ge i) = \mu\mathbb{E}(M).$$

Example 7.13 Observe that $M = N(t)$ is not a stopping time and in general $\mathbb{E}(T_{N(t)}) \ne \mu m(t)$. Indeed, for the Poisson process $\mu m(t) = t$ while $T_{N(t)} = t - C(t)$, where $C(t)$ is the current lifetime.

Theorem 7.14 *Elementary renewal theorem: $m(t)/t \to 1/\mu$ as $t \to \infty$, where $\mu = \mathbb{E}(X_1)$.*

Proof. Since $M = N(t) + 1$ is a stopping time for (X_n):

$$\{M \le m\} = \{N(t) \le m-1\} = \{X_1 + \ldots + X_m > t\},$$

the Wald equation implies
$$\mathbb{E}(T_{N(t)+1}) = \mu\mathbb{E}(N(t)+1) = \mu U(t).$$
From $T_{N(t)+1} = t + E(t)$ we get
$$U(t) = \mu^{-1}(t + \mathbb{E}(E(t))),$$
so that $U(t) \geq \mu^{-1}t$. Moreover, if $\mathbb{P}(X_1 \leq a) = 1$ for some finite a, then $U(t) \leq \mu^{-1}(t+a)$ and the assertion follows.

If X_1 is unbounded, then consider truncated inter-arrival times $\min(X_i, a)$ with mean μ_a and renewal function $U_a(t)$. It remains to observe that $U_a(t) \sim t\mu_a^{-1}$, $U_a(t) \geq U(t)$, and $\mu_a \to \mu$ as $a \to \infty$.

7.4 Stationary excess lifetime distribution

Theorem 7.15 *Renewal theorem. If X_1 is not arithmetic, then for any positive h*
$$U(t + h) - U(t) \to \mu^{-1}h, \quad t \to \infty.$$

Without proof.

Example 7.16 Arithmetic case. A typical arithmetic case is obtained, if we assume that the set of possible values R_X for the inter-arrival times X_i satisfies $R_X \subset \{1, 2, \ldots\}$ and $R_X \not\subset \{k, 2k, \ldots\}$ for any $k = 2, 3, \ldots$, implying $\mu \geq 1$. If again $U(n)$ is the renewal function, then $U(n) - U(n-1)$ is the probability that a renewal event occurs at time n. A discrete time version of Theorem 7.15 claims that $U(n) - U(n-1) \to \mu^{-1}$.

Theorem 7.17 *Key renewal theorem. If X_1 is not arithmetic, $\mu < \infty$, and $g : [0, \infty) \to [0, \infty)$ is a monotone function, then*

$$\int_0^t g(t - u)dU(u) \to \mu^{-1} \int_0^\infty g(u)du, \quad t \to \infty.$$

Sketch of the proof. Using Theorem 7.15, first prove the assertion for indicator functions of intervals, then for step functions, and finally for the limits of increasing sequences of step functions.

Theorem 7.18 *If X_1 is not arithmetic and $\mu < \infty$, then*

$$\lim_{t \to \infty} \mathbb{P}(E(t) \leq y) = \mu^{-1} \int_0^y (1 - F(x))dx.$$

Proof. Apply the key renewal theorem and Lemma 7.7.

Definition 7.19 Let $X_1, X_2, \ldots$ be independent positive r.v. such that $X_2, X_3, \ldots$ have the same distribution. If as before, $T_0 = 0$ and $T_n = X_1 + \ldots + X_n$, then $N^{\mathrm{d}}(t) = \max\{n : T_n \leq t\}$ is called a *delayed renewal process*. It is described by two distributions $F(t) = \mathbb{P}(X_i \leq t)$, $i \geq 2$ and $F^{\mathrm{d}}(t) = \mathbb{P}(X_1 \leq t)$.

Lemma 7.20 *The mean $m^{\mathrm{d}}(t) = \mathbb{E}N^{\mathrm{d}}(t)$ satisfies the renewal equation*

$$m^{\mathrm{d}}(t) = F^{\mathrm{d}}(t) + \int_0^t m^{\mathrm{d}}(t - u)dF(u).$$

Proof. First observe that conditioning on X_1 gives

$$m^{\mathrm{d}}(t) = F^{\mathrm{d}}(t) + \int_0^t m(t - u)dF^{\mathrm{d}}(u).$$

Now since

$$m(t) = F(t) + \int_0^t m(t - u)dF(u),$$

we have

$$m * F^{\mathrm{d}}(t) := \int_0^t m(t - u)dF^{\mathrm{d}}(u) = F * F^{\mathrm{d}}(t) + m * F * F^{\mathrm{d}}(t) = F^{\mathrm{d}} * F(t) + m * F^{\mathrm{d}} * F(t)$$

$$= (F^{\mathrm{d}} + m * F^{\mathrm{d}}) * F(t) = m^{\mathrm{d}} * F(t) = \int_0^t m^{\mathrm{d}}(t - u)dF(u).$$

41

Theorem 7.21 *The process $N^{\mathrm{d}}(t)$ has stationary increments:* $N^{\mathrm{d}}(s+t) - N^{\mathrm{d}}(s) \overset{d}{=} N^{\mathrm{d}}(t)$, *if and only if*

$$F^{\mathrm{d}}(y) = \mu^{-1} \int_0^y (1 - F(x))dx. \tag{11}$$

In this case the renewal function is linear $m^{\mathrm{d}}(t) = t/\mu$ *and* $\mathbb{P}(E^{\mathrm{d}}(t) \le y) = F^{\mathrm{d}}(y)$ *independently of* t.

Proof. Necessity. If $N^{\mathrm{d}}(t)$ has stationary increments, then $m^{\mathrm{d}}(s+t) = m^{\mathrm{d}}(s) + m^{\mathrm{d}}(t)$, and we get $m^{\mathrm{d}}(t) = t m^{\mathrm{d}}(1)$. Substitute this into the renewal equation of Lemma 7.20 to obtain

$$F^{\mathrm{d}}(t) = m^{\mathrm{d}}(1) \int_0^t (1 - F(x))dx$$

using integration by parts. Let $t \to \infty$ to find $m^{\mathrm{d}}(1) = 1/\mu$ as stated in (11).

Sufficiency. Given (11), we have

$$\hat{F}^{\mathrm{d}}(\theta) = \int_0^\infty e^{-\theta y} dF^{\mathrm{d}}(y) = \mu^{-1} \int_0^\infty e^{-\theta y}(1 - F(y))dy = \frac{1 - \hat{F}(\theta)}{\mu \theta}.$$

Taking the Laplace-Stieltjes transforms in Lemma 7.20 yields

$$\hat{m}^{\mathrm{d}}(\theta) = \frac{1 - \hat{F}(\theta)}{\mu \theta} + \hat{m}^{\mathrm{d}}(\theta)\hat{F}(\theta).$$

It follows that $\hat{m}^{\mathrm{d}}(\theta) = 1/(\mu \theta)$, and therefore $m^{\mathrm{d}}(t) = t/\mu$.

Observe that $N^{\mathrm{d}}(\cdot)$ has stationary increments if and only if the distribution of $E^{\mathrm{d}}(t)$ does not depend on t, and it is enough to check that $\mathbb{P}(E^{\mathrm{d}}(t) > y) = 1 - F^{\mathrm{d}}(y)$. This is obtained from

$$\mathbb{P}(E^{\mathrm{d}}(t) > y) = 1 - F^{\mathrm{d}}(t+y) + \int_0^t \mathbb{P}(E^{\mathrm{d}}(t-u) > y)dF^{\mathrm{d}}(u).$$

and

$$F^{\mathrm{d}} * U(t) = m^{\mathrm{d}}(t) = t/\mu.$$

Indeed, writing $g(t) = 1 - F(t+y)$ we have $\mathbb{P}(E^{\mathrm{d}}(t) > y) = g * U(t)$ and therefore

$$\int_0^t \mathbb{P}(E^{\mathrm{d}}(t-u) > y)dF^{\mathrm{d}}(u) = g * U * F^{\mathrm{d}}(t) = g * F^{\mathrm{d}} * U(t) = \mu^{-1} \int_0^t (1 - F(t+y-u))du$$

$$= \mu^{-1} \int_y^{t+y} (1 - F(x))dx = F^{\mathrm{d}}(t+y) - F^{\mathrm{d}}(y).$$

7.5 Renewal-reward processes

Let $(X_i, R_i), i = 1, 2, \ldots$ be iid pairs of possibly dependent random variables: X_i are positive inter-arrival times and R_i the associated rewards. Cumulative reWard process

$$W(t) = R_1 + \ldots + R_{N(t)}.$$

Theorem 7.22 *Renewal-reward theorem. Suppose* (X_i, R_i) *have finite means* $\mu = \mathbb{E}(X)$ *and* $\mathbb{E}(R)$. *Then*

$$\frac{W(t)}{t} \overset{\text{a.s.}}{\to} \frac{\mathbb{E}(R)}{\mu}, \qquad \frac{\mathbb{E}W(t)}{t} \to \frac{\mathbb{E}(R)}{\mu}, \qquad t \to \infty.$$

Proof. We have by the laws of large numbers, Theorem 4.9 and Theorem 7.9, that the first claim is valid

$$\frac{W(t)}{t} = \frac{R_1 + \ldots + R_{N(t)}}{N(t)} \cdot \frac{N(t)}{t} \overset{\text{a.s.}}{\to} \frac{\mathbb{E}(R)}{\mu}.$$

To prove the second claim we apply the argument used in the proof of Theorem 7.14

$$\mathbb{E}W(t) = \mathbb{E}(R_1 + \ldots + R_{N(t)+1}) - \mathbb{E}R_{N(t)+1} = U(t)\mathbb{E}(R) - \mathbb{E}(R_{N(t)+1}).$$
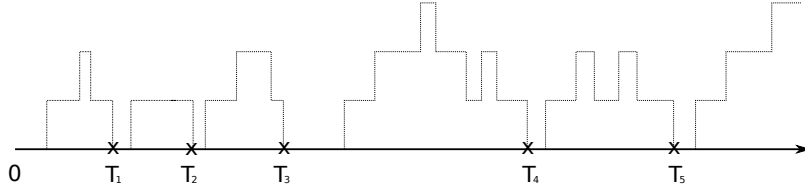
42

Figure 2: The queue length and the regeneration times. Here $N_1 = 2$, $N_2 = 1$, $N_3 = 2$, $N_4 = 4$, $N_5 = 3$.

The result will follow once we have shown that $r(t)/t \to 0$ as $t \to \infty$, where $r(t) = \mathbb{E}(R_{N(t)+1})$. By conditioning on X_1, as usual, we arrive at the renewal equation

$$r(t) = H(t) + \int_0^t r(t-u)dF(u), \qquad H(t) = \mathbb{E}(R_1 1_{\{X_1 > t\}}).$$

Since $H(t) \to 0$, for any $\epsilon > 0$, there exists a finite $M = M(\epsilon)$ such that $|H(t)| < \epsilon$ for $t \geq M$. Therefore, when $t \geq M$,

$$t^{-1}|r(t)| \leq t^{-1} \int_0^{t-M} |H(t-u)|dU(u) + t^{-1} \int_{t-M}^t |H(t-u)|dU(u)$$
$$\leq t^{-1}\Big(\epsilon U(t) + (U(t) - U(t-M))\mathbb{E}|R|\Big).$$

Using the renewal theorems we get $\limsup |r(t)/t| \leq \epsilon/\mu$ and it remains to let $\epsilon \to 0$.

7.6 Regeneration technique for queues

Here we start by three applications (A-C) of Theorem 7.22 to a general queueing system. Customers arrive one by one, and after spending some time in the system they depart from the system. Let $Q(t)$ be the number of customers in the system at time t with $Q(0) = 0$. Let T be the time of first return to zero $Q(T) = 0$. This can be called a regeneration time as starting from time T the future process $Q(T + t)$ is independent of the past and has the same distribution as the process $Q(t)$.

Assuming the traffic is *light* so that $\mathbb{P}(T < \infty) = 1$, we get a renewal process of regeneration times $0 = T_0 < T = T_1 < T_2 < T_3 < \dots$. Write N_i for the number of customers arriving during the cycle $[T_{i-1}, T_i)$ and put $N = N_1$. To be able to apply Theorem 7.22 we shall assume

$$\mathbb{E}(T) < \infty, \qquad \mathbb{E}(N) < \infty, \qquad \mathbb{E}(NT) < \infty.$$

(A) Let the reward associated with the inter-arrival time $X_i = T_i - T_{i-1}$ to be

$$R_i = \int_{T_{i-1}}^{T_i} Q(u)du.$$

Since $R := R_1 \leq NT$, we have $\mathbb{E}(R) \leq \mathbb{E}(NT) < \infty$, and by Theorem 7.22,

$$t^{-1}\int_0^t Q(u)du \overset{\text{a.s.}}{\to} \mathbb{E}(R)/\mathbb{E}(T) =: L \text{ the long run average queue length.}$$

(B) Let the reward associated with the inter-arrival time X_i to be N_i. Denote by $W(t)$ the number of customers arrived by time t. Then by Theorem 7.22,

$$W(t)/t \overset{\text{a.s.}}{\to} \mathbb{E}(N)/\mathbb{E}(T) =: \lambda \text{ the long run rate of arrival.}$$

(C) Consider now the reward-renewal process with discrete inter-arrival times N_i, and the associated rewards S_i defined as the total time spent in system by the customers arrived during $[T_{i-1}, T_i)$. If the n-th customer spends time V_n, then

$$S := S_1 = V_1 + \dots + V_N.$$

43

Since $\mathbb{E}(S) \le \mathbb{E}(NT) < \infty$, again by Theorem 7.22, we have

$$n^{-1} \sum_{i=1}^{n} V_i \overset{\text{a.s.}}{\to} \mathbb{E}(S)/\mathbb{E}(N) =: \nu \text{ the long run average time spent by a customer in the system.}$$

Theorem 7.23 *Little's law. If $\mathbb{E}(T) < \infty$, $\mathbb{E}(N) < \infty$, and $\mathbb{E}(NT) < \infty$, then $L = \lambda\nu$.*

Although it looks intuitively reasonable, it's a quite remarkable result, as the relationship is not influenced by the arrival process distribution, the service distribution, the service order, or practically anything else.

Proof. The mean amount of customer time spent during the first cycle can be represented in two ways

$$\mathbb{E}(S) = \mathbb{E}\Big(\sum_{i=1}^{N} V_i \Big) = \mathbb{E}\Big(\int_0^T Q(u)du \Big) = \mathbb{E}(R).$$

Thus combining (A), (B), (C) we get

$$\frac{L}{\lambda\nu} = \frac{\mathbb{E}(R)}{\mathbb{E}(T)} \cdot \frac{\mathbb{E}(T)}{\mathbb{E}(N)} \cdot \frac{\mathbb{E}(N)}{\mathbb{E}(S)} = 1.$$

7.7   M/M/1 queues

The most common notation scheme annotates the queueing systems by a triple $A/B/s$, where A describes the distribution of inter-arrival times of customers, B describes the distribution of service times, and s is the number of servers. It is assumed that the inter-arrival and service times are two independent iid sequences (X_i) and (S_i). Let $Q(t)$ be the queue size at time t.

The simplest queue M/M/1 has exponential with parameter λ inter-arrival times and exponential with parameter μ service times (M stands for the Markov discipline). It is the only queue with $Q(t)$ forming a Markov chain. In this case $Q(t)$ is a birth-death process with the birth rate λ and the death rate μ for the positive states, and no deaths at state zero.

Theorem 7.24 *The probabilities $p_n(t) = \mathbb{P}(Q(t) = n)$ satisfy the Kolmogorov forward equations*

$$\begin{cases} p_n'(t) & = -(\lambda + \mu)p_n(t) + \lambda p_{n-1}(t) + \mu p_{n+1}(t) \qquad \text{for } n \ge 1, \\ p_0'(t) & = -\lambda p_0(t) + \mu p_1(t), \end{cases}$$

subject to the boundary condition $p_n(0) = 1_{\{n=0\}}$.

Proof. For $n \ge 1$ and a small positive δ, put

$$R_n(t, \delta) = \mathbb{P}\big(Q(t + \delta) = n, \text{at least two events during } (t, t + \delta]\big).$$

Observe that

$$R_n(t, \delta) \le \mathbb{P}(\text{at least two events during } (t, t + \delta]) \le C \cdot \delta^2.$$

For example

$$\mathbb{P}(\text{at least two births during } (t, t + \delta]) = \sum_{k \ge 2} \frac{(\lambda\delta)^k}{k!} e^{-\lambda\delta}$$

$$\le (\lambda\delta)^2 \sum_{j \ge 0} \frac{(\lambda\delta)^j}{j!} e^{-\lambda\delta} = (\lambda\delta)^2.$$

It follows,

$$\begin{aligned} p_n(t + \delta) &= \mathbb{P}\big(Q(t + \delta) = n, \text{at most one event during } (t, t + \delta]\big) + R_n(t, \delta) \\ &= p_n(t)\mathbb{P}(\text{no events during } (t, t + \delta]) + p_{n-1}(t)\mathbb{P}(\text{single birth during } (t, t + \delta]) \\ &\quad + p_{n+1}(t)\mathbb{P}(\text{single death during } (t, t + \delta]) + o(\delta) \\ &= p_n(t)(1 - (\lambda + \mu)\delta) + p_{n-1}(t)\lambda\delta + p_{n+1}(t)\mu\delta + o(\delta). \end{aligned}$$

Similarly,

$$p_0(t+\delta) = p_0(t)\mathbb{P}(\text{no birth during } (t, t+\delta]) + p_1(t)\mathbb{P}(\text{single death during } (t, t+\delta]) + o(\delta)$$
$$= p_0(t)(1-\lambda\delta) + p_1(t)\mu\delta + o(\delta).$$

Exercise 7.25 In terms of the Laplace transforms $\hat{p}_n(\theta) := \int_0^\infty e^{-\theta t}p_n(t)dt$ we obtain

$$\begin{cases} \mu\hat{p}_{n+1}(\theta) - (\lambda + \mu + \theta)\hat{p}_n(\theta) + \lambda\hat{p}_{n-1}(\theta) &= 0, \qquad \text{for } n \geq 1, \\ \mu\hat{p}_1(\theta) - (\lambda + \theta)\hat{p}_0(\theta) &= 1. \end{cases}$$

Its solution is given by

$$\hat{p}_n(\theta) = \theta^{-1}(1 - \alpha(\theta))\alpha(\theta)^n, \qquad \alpha(\theta) := \frac{(\lambda + \mu + \theta) - \sqrt{(\lambda + \mu + \theta)^2 - 4\lambda\mu}}{2\mu}.$$

Theorem 7.26 *Stationarity. Let $\rho = \lambda/\mu$ be the traffic intensity. As $t \to \infty$ for all $n \geq 0$*

$$\mathbb{P}(Q(t) = n) \to \pi_n := \begin{cases} (1-\rho)\rho^n & \text{if } \rho < 1, \\ 0 & \text{if } \rho \geq 1. \end{cases}$$

The result asserts that the queue settles down into equilibrium if and only if the service times are shorter than the inter-arrival times on average. Observe that if $\rho \geq 1$, we have $\sum_n \pi_n = 0$ despite $\sum_n p_n(t) = 1$.

Proof. As $\theta \to 0$

$$\alpha(\theta) = \frac{(\lambda + \mu + \theta) - \sqrt{(\lambda - \mu)^2 + 2(\lambda + \mu)\theta + \theta^2}}{2\mu} \to \frac{(\lambda + \mu + \theta) - |\lambda - \mu|}{2\mu} = \begin{cases} \rho & \text{if } \lambda < \mu, \\ 1 & \text{if } \lambda \geq \mu. \end{cases}$$

Thus

$$\theta\hat{p}_n(\theta) = (1 - \alpha(\theta))\alpha(\theta)^n \to \begin{cases} (1-\rho)\rho^n & \text{if } \rho < 1, \\ 0 & \text{if } \rho \geq 1. \end{cases}$$

On the other hand, the process $Q(t)$ is an irreducible Markov chain and the ergodic theorem there are limits $p_n(t) \to \pi_n$ as $t \to \infty$. Therefore, the statement follows from

$$\theta\hat{p}_n(\theta) = \int_0^\infty e^{-u}p_n(u/\theta)du \to \pi_n, \qquad \theta \to 0.$$

Exercise 7.27 Consider the M/M/1 queue with $\rho < 1$. According to Little's Law the time V spent by a customer in the stationary regime queue has mean $\frac{L}{\lambda} = \frac{\rho}{(1-\rho)\lambda} = \frac{1}{\mu-\lambda}$. Verify this by showing that V has an exponential distribution with parameter $\mu - \lambda$.
Hint. Show that $V = S_1 + \ldots + S_{1+Q}$, where S_i are independent $\text{Exp}(\mu)$ random variables and Q is geometric with parameter $1 - \rho$, and then compute the Laplace transform $\mathbb{E}(e^{-uV})$.

7.8 M/G/1 queues

In the M/G/1 queueing system customers arrive according to a Poisson process with intensity λ and the service times S_i have a fixed but unspecified distribution (G for General). Let $\rho = \lambda\mathbb{E}(S)$ be the traffic intensity.

Theorem 7.28 *If the first customer arrives at time X_1, define a typical busy period of the server as $B = \inf\{t > 0 : Q(t + X_1) = 0\}$. We have that $\mathbb{P}(B < \infty) = 1$ if $\rho \leq 1$, and $\mathbb{P}(B < \infty) < 1$ if $\rho > 1$. The Laplace transform $\phi(u) = \mathbb{E}(e^{-uB})$ satisfies the functional equation*

$$\phi(u) = \Psi(u + \lambda - \lambda\phi(u)), \qquad \text{where } \Psi(x) := \mathbb{E}(e^{-xS}).$$

Proof. Imbedded branching process (recall Section 5.5). Call customer C_2 an offspring of customer C_1, if C_2 joins the queue while C_1 is being served. Given $S = t$, the offspring number Z for a single customer has conditional $\text{Pois}(\lambda t)$ distribution. Therefore, the conditional probability generating function of Z is

$$\mathbb{E}(u^Z|S) = \sum_{j=0}^\infty u^j \frac{(\lambda S)^j}{j!}e^{-\lambda S} = e^{S\lambda(u-1)}.$$

This yields
$$h(u) := \mathbb{E}(u^Z) = \mathbb{E}(e^{S\lambda(u-1)}) = \Psi(\lambda(1-u)).$$

The mean offspring number is
$$\mathbb{E}(Z) = h'(1) = -\lambda\Psi'(0) = \lambda\mathbb{E}(S) = \rho.$$

Observe that the event $(B < \infty)$ is equivalent to the extinction of the branching process, and the first assertion follows.

The functional equation follows from the representation $B = S + B_1 + \ldots + B_Z$, where B_i are iid busy time of the offspring customers and Z is the number of offspring with generating function $h(u)$. Indeed, we observe first that
$$\mathbb{E}(e^{-uB_1}\ldots e^{-uB_Z}|S) = \mathbb{E}(\phi(u)^Z|S) = e^{(\lambda\phi(u)-\lambda)S}.$$

This implies
$$\phi(u) = \mathbb{E}(e^{-uS}e^{-uB_1}\ldots e^{-uB_Z}) = \mathbb{E}(e^{-uS}e^{(\lambda\phi(u)-\lambda)S}) = \Psi(u+\lambda-\lambda\phi(u)).$$

Exercise 7.29 Show that for the M/M/1 queue, Z is geometric with parameter $\frac{1}{1+\rho}$. Compute the generating function $\phi(u)$ for the busy period B.

Theorem 7.30 *Stationarity. As* $t \to \infty$ *for all* $n \geq 0$
$$\mathbb{P}(Q(t) = n) \to \begin{cases} \pi_n & \text{if } \rho < 1, \text{ where } \sum_{j=0}^{\infty} \pi_j u^j = (1-\rho)(1-u)\frac{h(u)}{h(u)-u}, \\ 0 & \text{if } \rho \geq 1. \end{cases}$$

Partial proof. Let D_n be the number of customers in the system right after the n-th customer left the system. Denoting Z_n the offspring number of the n-th customer, we get
$$D_{n+1} = D_n + Z_{n+1} - 1_{\{D_n>0\}}.$$

Clearly, D_n forms a Markov chain with the state space $\{0,1,2,\ldots\}$ and transition probabilities
$$\begin{pmatrix} \delta_0 & \delta_1 & \delta_2 & \ldots \\ \delta_0 & \delta_1 & \delta_2 & \ldots \\ 0 & \delta_0 & \delta_1 & \ldots \\ 0 & 0 & \delta_0 & \ldots \\ \ldots & \ldots & \ldots & \ldots \end{pmatrix}, \qquad \delta_j = \mathbb{P}(Z=j) = \mathbb{E}\Big(\frac{(\lambda S)^j}{j!}e^{-\lambda S}\Big).$$

The stationary distribution of this chain can be found from the equation
$$D \stackrel{d}{=} D - 1_{\{D>0\}} + Z.$$

For the probability generating function $\psi(u) = \mathbb{E}(u^D)$, we have
$$\psi(u) = \mathbb{E}(u^{D-1_{\{D>0\}}})h(u) = \Big(\frac{\psi(u)-\psi(0)}{u} + \psi(0)\Big)h(u).$$

Thus
$$\psi(u) = \frac{\psi(0)(u-1)h(u)}{u-h(u)}.$$

If $\rho < 1$, we get
$$1 = \psi(1) = \frac{\psi(0)}{1-h'(1)} = \frac{\psi(0)}{1-\rho}$$

and therefore
$$\mathbb{E}(u^D) = (1-\rho)(1-u)\frac{h(u)}{h(u)-u}. \tag{12}$$

gives the same stationary distribution as in the theorem statement (only for the embedded chain).

From this analysis of the chain (D_n) one can be derive (not shown) the stated convergence for the queue size distribution.

Theorem 7.31 *Suppose a customer joins the queue after some large time has elapsed. She will wait a period W of time before her service begins. If $\rho < 1$, then*

$$\mathbb{E}(e^{-uW}) = \frac{(1-\rho)u}{u - \lambda + \lambda\mathbb{E}(e^{-uS})}.$$

Proof. Suppose that a customer waits for a period of length W and then is served for a period of length S. In the stationary regime the length D of the queue on the departure of this customer is distributed according (12). On the other hand, D conditionally on (W, S) has a Poisson distribution with parameter $\lambda(W + S)$:

$$(1-\rho)(1-s)\frac{h(s)}{h(s) - s} = \mathbb{E}(s^D) = \mathbb{E}(e^{\lambda(W+S)(s-1)}) = \mathbb{E}(e^{\lambda W(s-1)})h(s).$$

Thus, with $u = \lambda(1-s)$,

$$\mathbb{E}(e^{-uW}) = \mathbb{E}(e^{\lambda(s-1)W}) = \frac{(1-\rho)(s-1)}{s - \mathbb{E}(e^{\lambda(s-1)S})} = \frac{(1-\rho)u}{u - \lambda + \lambda\mathbb{E}(e^{-uS})}.$$

Exercise 7.32 Using the previous theorem show that for the M/M/1 queue we have

$$\mathbb{E}(e^{-uW}) = 1 - \rho + \rho\frac{\mu - \lambda}{\mu - \lambda + u}.$$

Theorem 7.33 *Heavy traffic. Let $\rho = \lambda d$ be the traffic intensity of the M/G/1 queue with $M = M(\lambda)$ and $G = D(d)$ concentrated at value d. For $\rho < 1$, let Q_ρ be a r.v. with the equilibrium queue length distribution. Then $(1-\rho)Q_\rho$ converges in distribution as $\rho \nearrow 1$ to the exponential distribution with parameter 2.*

Proof. Applying Theorem 7.30 with $u = e^{-s(1-\rho)}$ and $h(u) = e^{\rho(u-1)}$, we find the Laplace transform of the scaled queue length

$$\mathbb{E}(e^{-s(1-\rho)Q_\rho}) = \sum_{j=0}^{\infty} \pi_j u^j = (1-\rho)(1-u)\frac{e^{\rho(u-1)}}{e^{\rho(u-1)} - u} = \frac{(1-\rho)(1 - e^{-s(1-\rho)})}{1 - e^{\rho(1-u)}e^{-s(1-\rho)}}$$

which converges to $\frac{2}{2+s}$ as $\rho \nearrow 1$.

Exercise 7.34 For $\rho = 0.99$ find $\mathbb{P}(Q_\rho > 60)$ using the previous theorem. Answer 30%.

7.9   G/M/1 queues

Customers' arrival times form a renewal process with inter-arrival times (X_n), and the service times are exponentially distributed with parameter μ. The traffic intensity is $\rho = (\mu\mathbb{E}(X))^{-1}$.

An imbedded Markov chain. Consider the moment T_n at which the n-th customer joins the queue, and let A_n be the number of customers who are ahead of him in the system. Define V_n as the number of departures from the system during the interval $[T_n, T_{n+1})$. Conditionally on A_n and $X_{n+1} = T_{n+1} - T_n$, the r.v. V_n has a truncated Poisson distribution

$$\mathbb{P}(V_n = v | A_n = a, X_{n+1} = x) = \begin{cases} \frac{(\mu x)^v}{v!}e^{-\mu x} & \text{if } v \leq a, \\ \sum_{i \geq a+1}\frac{(\mu x)^i}{i!}e^{-\mu x} & \text{if } v = a + 1. \end{cases}$$

The sequence (A_n) satisfies

$$A_{n+1} = A_n + 1 - V_n$$

and forms a Markov chain with transition probabilities

$$\begin{pmatrix} 1 - \alpha_0 & \alpha_0 & 0 & 0 & \ldots \\ 1 - \alpha_0 - \alpha_1 & \alpha_1 & \alpha_0 & 0 & \ldots \\ 1 - \alpha_0 - \alpha_1 - \alpha_2 & \alpha_2 & \alpha_1 & \alpha_0 & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots \end{pmatrix}, \qquad \alpha_j = \mathbb{E}\left(\frac{(\mu X)^j}{j!}e^{-\mu X}\right).$$

Theorem 7.35 *If $\rho < 1$, then the chain (A_n) is ergodic with a unique stationary distribution $\pi_j = (1-\eta)\eta^j$, where η is the smallest positive root of $\eta = \mathbb{E}(e^{X\mu(\eta-1)})$. If $\rho = 1$, then the chain (A_n) is null recurrent. If $\rho > 1$, then the chain (A_n) is transient.*

Without proof.

Example 7.36 Consider a deterministic arrival process with $\mathbb{P}(X = 1) = 1$ so that $A_n = Q(n) - 1$. Then $\rho = 1/\mu$ and for $\mu > 1$ we have the stationary distribution

$$\mathbb{P}(A = j) = (1-\eta)\eta^j,$$

where $\eta = e^{\mu(\eta-1)}$. For $t \in (0,1)$, we have $Q(n+t) = A_n + 1 - V_n(t)$, where

$$p_t(a) := \mathbb{P}(V_n(t) = v | A_n = a) = \begin{cases} \frac{(\mu t)^v}{v!} e^{-\mu t} & \text{if } v \le a, \\ \sum_{i \ge a+1} \frac{(\mu t)^i}{i!} e^{-\mu t} & \text{if } v = a+1. \end{cases}$$

Using the stationary distribution for A_n we get

$$\mathbb{P}(Q(n+t) = k) = \sum_{j=k-1}^{\infty} (1-\eta)\eta^j p_t(j).$$

This example demonstrates that unlike (D_n) in the case of M/G/1, the stationary distribution of (A_n) need not be the limiting distribution of the queue size $Q(t)$.

Theorem 7.37 *Let $\rho < 1$, and assume that the chain (A_n) is in equilibrium. Then the waiting time W of an arriving customer has an atom of size $1 - \eta$ at zero and for $x \ge 0$*

$$\mathbb{P}(W > x) = \eta e^{-\mu(1-\eta)x}.$$

Proof. If $A_n > 0$, then the waiting time of the n-th customer is

$$W_n = S_1^* + S_2 + S_3 + \ldots + S_{A_n},$$

where S_1^* is the residual service time of the customer under service, and $S_2, S_3, \ldots, S_{A_n}$ are the service times of the others in the queue. By the lack-of-memory property, this is a sum of A_n iid exponentials. Use the equilibrium distribution of A_n to find that

$$\mathbb{E}(e^{-uW}) = \mathbb{E}((\mathbb{E}(e^{-uS}))^A) = \mathbb{E}\left(\left(\frac{\mu}{\mu+u}\right)^A\right) = \frac{(\mu+u)(1-\eta)}{\mu+u-\mu\eta} = (1-\eta) + \eta\frac{\mu(1-\eta)}{\mu(1-\eta)+u}.$$

Exercise 7.38 Using the previous theorem show that for the M/M/1 queue we have

$$\mathbb{E}(e^{-uW}) = 1 - \rho + \rho\frac{\mu-\lambda}{\mu-\lambda+u}.$$

7.10   G/G/1 queues

Now the arrivals of customers form a renewal process with inter-arrival times X_n having an arbitrary common distribution. The service times S_n have another fixed distribution. The traffic intensity is given by $\rho = \mathbb{E}(S)/\mathbb{E}(X)$.

Lemma 7.39 *Lindley's equation. Let W_n be the waiting time of the n-th customer. Then*

$$W_{n+1} = \max\{0, W_n + S_n - X_{n+1}\}.$$

Proof. The n-th customer is in the system for a length $W_n + S_n$ of time. If $X_{n+1} > W_n + S_n$, then the queue is empty at the $(n+1)$-th arrival, and so $W_{n+1} = 0$. If $X_{n+1} \le W_n + S_n$, then the $(n+1)$-th customer arrives while the n-th is still present. In the second case the new customer waits for a period of $W_{n+1} = W_n + S_n - X_{n+1}$.

Theorem 7.40 *Note that $U_n = S_n - X_{n+1}$ is a collection of iid r.v. Denote by $G(x) = \mathbb{P}(U_n \leq x)$ their common distribution function. Let $F_n(x) = \mathbb{P}(W_n \leq x)$. Then for $x \geq 0$*

$$F_{n+1}(x) = \int_{-\infty}^x F_n(x-y)dG(y).$$

There exists a limit $F(x) = \lim_{n\to\infty} F_n(x)$ which satisfies the Wiener-Hopf equation

$$F(x) = \int_{-\infty}^x F(x-y)dG(y).$$

Proof. If $x \geq 0$ then due to the Lindley equation and independence between W_n and $U_n = S_n - X_{n+1}$

$$\mathbb{P}(W_{n+1} \leq x) = \mathbb{P}(W_n + U_n \leq x) = \int_{-\infty}^x \mathbb{P}(W_n \leq x-y)dG(y)$$

and the first part is proved. We claim that

$$F_{n+1}(x) \leq F_n(x) \text{ for all } x \geq 0 \text{ and } n \geq 1. \tag{13}$$

If (13) holds, then the second result follows immediately. We prove (13) by induction. Observe that $F_2(x) \leq 1 = F_1(x)$, and suppose that (13) holds for $n = k - 1$. Then

$$F_{k+1}(x) - F_k(x) = \int_{-\infty}^x (F_k(x-y) - F_{k-1}(x-y))dG(y) \leq 0.$$

Theorem 7.41 *If $\rho < 1$, then F is a non-defective distribution function. If $\rho = 1$ and G is not concentrated at one point or if $\rho > 1$, then $F(x) = 0$ for all x.*

Proof. Imbedded random walk. Define an imbedded random walk by

$$\Sigma_0 = 0, \quad \Sigma_n = U_1 + \ldots + U_n.$$

Lemma 7.39 implies

$$W_{n+1} = \max\{0, U_n, U_n + U_{n-1}, \ldots, U_n + U_{n-1} + \ldots + U_1\} \overset{d}{=} \max\{\Sigma_0, \ldots, \Sigma_n\}.$$

Note that $\mathbb{E}(U) = \mathbb{E}(S) - \mathbb{E}(X)$, and $\mathbb{E}(U) < 0$ is equivalent to $\rho < 1$. It follows that

$$F(x) = \mathbb{P}(\Sigma_n \leq x \text{ for all } n), \qquad x \geq 0.$$

(A) If $\mathbb{E}(U) < 0$, then

$$\mathbb{P}(\Sigma_n > 0 \text{ for infinitely many } n) = \mathbb{P}(n^{-1}\Sigma_n > 0 \text{ i.o.}) = \mathbb{P}(n^{-1}\Sigma_n - \mathbb{E}(U) > |\mathbb{E}(U)| \text{ i.o.}) = 0$$

due to the LLN. Thus $\max\{\Sigma_0, \Sigma_1, \ldots\}$ is either zero or the maximum of only finitely many positive terms, and F is a non-defective distribution.

(B) Next suppose that $\mathbb{E}(U) > 0$ and pick any $x > 0$. For $n \geq 2x/\mathbb{E}(U)$

$$\mathbb{P}(\Sigma_n \geq x) = \mathbb{P}(n^{-1}\Sigma_n - \mathbb{E}(U) \geq n^{-1}x - \mathbb{E}(U)) \geq \mathbb{P}(n^{-1}\Sigma_n - \mathbb{E}(U) \geq -\mathbb{E}(U)/2) = \mathbb{P}(n^{-1}\Sigma_n \geq \mathbb{E}(U)/2).$$

Since $1 - F(x) \geq \mathbb{P}(\Sigma_n \geq x)$, the weak LLN implies $F(x) = 0$.

(C) In the case when $\mathbb{E}(U) = 0$ we need a more precise measure of the fluctuations of Σ_n. According to the law of the iterated logarithm the fluctuations of Σ_n are of order $O(\sqrt{n \log \log n})$ in both positive and negative directions with probability 1, and so for any given $x \geq 0$,

$$1 - F(x) = \mathbb{P}(\Sigma_n > x \text{ for some } n) = 1.$$

Definition 7.42 Define an increasing sequence of r.v. by

$$L(0) = 0, \quad L(n+1) = \min\{k > L(n) : \Sigma_k > \Sigma_{L(n)}\}.$$

The $L(n)$ are called ladder points of the random walk Σ, these are the times when the random walk Σ reaches its new maximal values.

Lemma 7.43 *Let* $\eta = \mathbb{P}(\Sigma_n > 0$ *for some* $n)$. *The total number* Λ *of ladder points has a geometric distribution*

$$\mathbb{P}(\Lambda = k) = (1 - \eta)\eta^k, \quad k = 0, 1, 2, \dots.$$

Theorem 7.44 *If* $\rho < 1$, *the equilibrium waiting time* $W \overset{d}{=} \max\{\Sigma_0, \Sigma_1, \dots\}$ *has the Laplace transform*

$$\mathbb{E}(e^{-\theta W}) = \frac{1 - \eta}{1 - \eta \mathbb{E}(e^{-\theta Y})},$$

where $Y = \Sigma_{L(1)}$ *is the first ladder hight of the imbedded random walk.*

Proof. Follows from the representation in terms of $Y_j = \Sigma_{L(j)} - \Sigma_{L(j-1)}$ which are iid copies of $Y = Y_1$:

$$\max\{\Sigma_0, \Sigma_1, \dots\} = \Sigma_{L(\Lambda)} = Y_1 + \dots + Y_\Lambda.$$

Exercise 7.45 Using the previous theorem show that for the M/M/1 queue we have

$$\mathbb{E}(e^{-uW}) = 1 - \rho + \rho \frac{\mu - \lambda}{\mu - \lambda + u}.$$

8 Martingales

8.1 Definitions and examples

Example 8.1 Martingale: a betting strategy. After tossing a fair coin a gambler wins for heads and looses for tails. Let X_n be the gain of a gambler doubling the bet after each loss. The game stops after the first win. The distribution of X_n evolves as follows

$X_0 = 0$,

$X_1 = 1$ with probability $1/2$ and $X_1 = -1$ with probability $1/2$,

$X_2 = 1$ with probability $3/4$ and $X_2 = -3$ with probability $1/4$,

$X_3 = 1$ with probability $7/8$ and $X_3 = -7$ with probability $1/8$,

and so on so that

$X_n = 1$ with probability $1 - 2^{-n}$ and $X_n = -2^n + 1$ with probability 2^{-n}.

If H_n and T_n are the heads and tails outcomes for the n-th tossing, then

$$X_{n+1} = (2X_n - 1)1_{T_{n+1}} + 1_{H_{n+1}}.$$

This yields in terms of the conditional expectations

$$\mathbb{E}(X_{n+1}|X_n) = (2X_n - 1) \cdot \frac{1}{2} + \frac{1}{2} = X_n.$$

If N is the number of coin tosses until one gets heads, then $\mathbb{P}(N = n) = 2^{-n}$, $n = 1, 2, \dots$ with $\mathbb{E}(N) = 2$. Interestingly, the state X_{N-1} prior to the winning toss has the mean
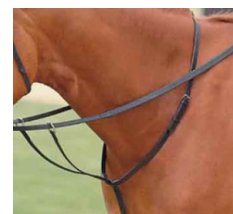
$$\mathbb{E}(X_{N-1}) = \mathbb{E}(1 - 2^{N-1}) = 1 - \sum_{n=1}^{\infty} 2^{n-1} 2^{-n} = -\infty.$$

Definition 8.2 A sequence of sigma-fields $(\mathcal{F}_n)$ such that $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \dots \subset \mathcal{F}_n \subset \dots \subset \mathcal{F}$ is called a filtration. A sequence of r.v. (Y_n) is called adapted to $(\mathcal{F}_n)$ if Y_n is $\mathcal{F}_n$-measurable for all n. In this case the sequence $(Y_n, \mathcal{F}_n)$ is called a martingale if, for all $n \geq 0$,

- $\mathbb{E}(|Y_n|) < \infty$,

- $\mathbb{E}(Y_{n+1}|\mathcal{F}_n) = Y_n$.

Definition 8.3 Let (Y_n) be adapted to a filtration $(\mathcal{F}_n)$. Then $(Y_n, \mathcal{F}_n)$ is called a submartingale if

- $\mathbb{E}(Y_n^+) < \infty$,

The origin of the term "martingale".

- $\mathbb{E}(Y_{n+1}|\mathcal{F}_n) \geq Y_n$.

Definition 8.4 Let (Y_n) be adapted to a filtration $(\mathcal{F}_n)$. Then $(Y_n, \mathcal{F}_n)$ is called a supermartingale if

- $\mathbb{E}(Y_n^-) < \infty$,
- $\mathbb{E}(Y_{n+1}|\mathcal{F}_n) \leq Y_n$.

Definition 8.5 The sequence $(Y_n)_{n\geq 1}$ is a martingale (submartingale, supermartingale) with respect to the sequence $(X_n)_{n\geq 1}$, if $(Y_n, \mathcal{F}_n)$ is a martingale (submartingale, supermartingale) for $\mathcal{F}_n = \sigma\{X_1, \ldots, X_n\}$, the σ-algebras generated by $(X_1, \ldots, X_n)$. It follows that $Y_n = \psi_n(X_1, \ldots, X_n)$. We sometimes just say that (Y_n) is a martingale.

Exercise 8.6 Consider the sequence of means $m_n = \mathbb{E}(Y_n)$. We have $m_{n+1} \geq m_n$ for submartingales, $m_{n+1} \leq m_n$ for supermartingales, and $m_{n+1} = m_n$ for martingales. A martingale is both a sub- and supermartingale. If (Y_n) is a submartingale, then $(-Y_n)$ is a supermartingale.

Example 8.7 Consider a simple random walk $S_n = S_0 + X_1 + \ldots + X_n$ with $\mathbb{P}(X_i = 1) = p$, $\mathbb{P}(X_i = -1) = q$, and $S_0 = k$. The centered $S_n - n(p-q)$ is a martingale:

$$\mathbb{E}(S_{n+1} - (n+1)(p-q)|X_1, \ldots, X_n) = S_n + \mathbb{E}(X_{n+1}) - (n+1)(p-q) = S_n - n(p-q).$$

Another martingale is $Y_n = (q/p)^{S_n}$:

$$\mathbb{E}(Y_{n+1}|X_1, \ldots, X_n) = p(q/p)^{S_n+1} + q(q/p)^{S_n-1} = (q/p)^{S_n} = Y_n$$

with $\mathbb{E}(Y_n) = \mathbb{E}(Y_0) = (q/p)^k$. It is called De Moivre's martingale.

Example 8.8 Let $S_n = X_1 + \ldots + X_n$, where X_i are iid r.v. with zero means and finite variances σ^2. Then $S_n^2 - n\sigma^2$ is a martingale

$$\mathbb{E}(S_{n+1}^2 - (n+1)\sigma^2|X_1, \ldots, X_n) = S_n^2 + 2S_n\mathbb{E}(X_{n+1}) + \mathbb{E}(X_{n+1}^2) - (n+1)\sigma^2 = S_n^2 - n\sigma^2.$$

Example 8.9 Application of the optional stopping theorem. Consider a simple random walk $S_n = S_0 + X_1 + \ldots + X_n$ with $\mathbb{P}(X_i = 1) = p$, $\mathbb{P}(X_i = -1) = q$, and $S_0 = k$. Let T be the first time when it hits 0 or N, where is $0 \leq k \leq N$. Next we show that for $p \neq q$,

$$\mathbb{P}(S_T = 0) = \frac{(p/q)^{N-k} - 1}{(p/q)^N - 1}.$$

Put $p_k := \mathbb{P}(S_T = 0)$ so that $\mathbb{P}(S_T = N) = 1 - p_k$. According to Definition 7.11, T is a stopping time. Applying Theorem 8.41 to De Moivre's martingale we get $\mathbb{E}(Y_T) = \mathbb{E}(Y_0)$. Since $\mathbb{P}(Y_T = 1) = p_k$ and $\mathbb{P}(Y_T = (q/p)^N) = 1 - p_k$, it follows

$$(q/p)^0 p_k + (q/p)^N (1 - p_k) = (q/p)^k,$$

and we immediately derive the stated equality.

In terms of the simple random walk starting at zero and stopped either at $-k$ or at N we find

$$\mathbb{P}(S_T = -k) = \frac{(p/q)^N - 1}{(p/q)^{N+k} - 1}.$$

If $p > q$ and $N \to \infty$, we obtain $\mathbb{P}(S_T = -1) \to \frac{q}{p}$. Thus, if $(-X)$ is the lowest level visited, then X has a shifted geometric distribution distribution with parameter $1 - \frac{q}{p}$.

Example 8.10 Stopped de Moivre's martingale. Consider the same simple random walk and let T be a stopping time of the random walk. Denote $D_n = Y_{T \wedge n}$, where (Y_n) is the de Moivre martingale. It is easy to see that D_n is also a martingale:

$$\mathbb{E}(D_{n+1}|X_1, \ldots, X_n) = \mathbb{E}(Y_{n+1}1_{\{T>n\}} + \sum_{j \leq n} Y_j 1_{\{T=j\}}|X_1, \ldots, X_n)$$

$$= Y_n 1_{\{T>n\}} + \sum_{j \leq n} Y_j 1_{\{T=j\}} = D_n.$$

Example 8.11 Branching processes (recall Section 5.5). Let Z_n be a branching process with $Z_0 = 1$ and the mean offspring number μ. In the supercritical case, $\mu > 1$, the extinction probability $\eta \in [0, 1)$ of Z_n is identified as a solution of the equation $\eta = h(\eta)$, where $h(s) = \mathbb{E}(s^X)$ is the generating function of the offspring number. The process $V_n = \eta^{Z_n}$ is a martingale

$$\mathbb{E}(V_{n+1}|Z_1, \ldots, Z_n) = \mathbb{E}(\eta^{X_1 + \ldots + X_{Z_n}}|Z_1, \ldots, Z_n) = h(\eta)^{Z_n} = V_n.$$

8.2 Convergence in L^2

Lemma 8.12 If (Y_n) is a martingale with $\mathbb{E}(Y_n^2) < \infty$, then $Y_{n+1} - Y_n$ and Y_n are uncorrelated. It follows that (Y_n^2) is a submartingale.

Proof. The first assertion follows from

$$\mathbb{E}(Y_n(Y_{n+1} - Y_n)|\mathcal{F}_n) = Y_n(\mathbb{E}(Y_{n+1}|\mathcal{F}_n) - Y_n) = 0.$$

The second claim is derived as follows

$$\mathbb{E}(Y_{n+1}^2|\mathcal{F}_n) = \mathbb{E}((Y_{n+1} - Y_n)^2 + 2Y_n(Y_{n+1} - Y_n) + Y_n^2|\mathcal{F}_n)$$
$$= \mathbb{E}((Y_{n+1} - Y_n)^2|\mathcal{F}_n) + Y_n^2 \geq Y_n^2.$$

Notice that

$$\mathbb{E}(Y_{n+1}^2) = \mathbb{E}(Y_n^2) + \mathbb{E}((Y_{n+1} - Y_n)^2)$$

so that $\mathbb{E}(Y_n^2)$ is non-decreasing and there always exists a finite or infinite limit

$$M = \lim_{n \to \infty} \mathbb{E}(Y_n^2). \tag{14}$$

Exercise 8.13 Let $J(x)$ be a convex function. If (Y_n) is a martingale with $\mathbb{E}[J(Y_n)] < \infty$, then $J(Y_n)$ is a submartingale. (Hint: use Jensen's inequality.)

Lemma 8.14 Doob-Kolmogorov's inequality. If (Y_n) is a martingale with $\mathbb{E}(Y_n^2) < \infty$, then for any $\epsilon > 0$

$$\mathbb{P}(\max_{1 \leq i \leq n} |Y_i| \geq \epsilon) \leq \frac{\mathbb{E}(Y_n^2)}{\epsilon^2}.$$

Proof. Let $B_1 = \{|Y_1| \geq \epsilon\}$ and $B_k = \{|Y_1| < \epsilon, \ldots, |Y_{k-1}| < \epsilon, |Y_k| \geq \epsilon\}$. Then using a submartingale property (in the second inequality) we get

$$\mathbb{E}(Y_n^2) \geq \sum_{i=1}^n \mathbb{E}(Y_n^2 1_{B_i}) \geq \sum_{i=1}^n \mathbb{E}(Y_i^2 1_{B_i}) \geq \epsilon^2 \sum_{i=1}^n \mathbb{P}(B_i) = \epsilon^2 \mathbb{P}(\max_{1 \leq i \leq n} |Y_i| \geq \epsilon).$$

Theorem 8.15 If (Y_n) is a martingale with finite M defined by (14), then there exists a random variable Y such that $Y_n \to Y$ a.s. and in mean square.

Proof. Step 1. For

$$A_m(\epsilon) = \bigcup_{i \geq 1} \{|Y_{m+i} - Y_m| \geq \epsilon\}$$

we will show that

$$\mathbb{P}(A_m(\epsilon)) \to 0, \quad m \to \infty \text{ for any } \epsilon > 0.$$

For a given m, put $S_n = Y_{m+n} - Y_m$. It is also a martingale, since

$$\mathbb{E}(S_{n+1}|S_1, \ldots, S_n) = \mathbb{E}(\mathbb{E}(S_{n+1}|\mathcal{F}_{m+n})|S_1, \ldots, S_n) = \mathbb{E}(S_n|S_1, \ldots, S_n) = S_n.$$

Apply the Doob-Kolmogorov inequality to this martingale to find that

$$\mathbb{P}(|Y_{m+i} - Y_m| \geq \epsilon \text{ for some } i \in [1, n]) \leq \epsilon^{-2} \mathbb{E}((Y_{m+n} - Y_m)^2) = \epsilon^{-2}(\mathbb{E}(Y_{m+n}^2) - \mathbb{E}(Y_m^2)).$$

Letting $n \to \infty$ we obtain

$$\mathbb{P}(A_m(\epsilon)) \leq \epsilon^{-2}(M - \mathbb{E}(Y_m^2))$$

and hence the step 1 statement follows.

Step 2. Since $A_m(\epsilon_1) \subset A_m(\epsilon_2)$ for $\epsilon_1 > \epsilon_2$, we have, by step 1,

$$\mathbb{P}\Big(\bigcup_{\epsilon > 0}\bigcap_{m \geq 1} A_m(\epsilon)\Big) = \lim_{\epsilon \to 0}\mathbb{P}\Big(\bigcap_{m \geq 1} A_m(\epsilon)\Big) \leq \lim_{\epsilon \to 0}\lim_{m \to \infty}\mathbb{P}(A_m(\epsilon)) = 0.$$

Therefore, the sequence (Y_n) is a.s. Cauchy convergent:

$$\mathbb{P}\Big(\bigcap_{\epsilon > 0}\bigcup_{m \geq 1} A_m^c(\epsilon)\Big) = 1,$$

which implies the existence of Y such that $Y_n \to Y$ a.s.

Step 3. Finally, we prove the convergence in mean square using the Fatou lemma

$$\mathbb{E}((Y_n - Y)^2) = \mathbb{E}(\liminf_{m \to \infty}(Y_n - Y_m)^2) \leq \liminf_{m \to \infty}\mathbb{E}((Y_n - Y_m)^2)$$
$$= \liminf_{m \to \infty}\mathbb{E}(Y_m^2) - \mathbb{E}(Y_n^2) = M - \mathbb{E}(Y_n^2) \to 0, \quad n \to \infty.$$

Example 8.16 Branching processes (recall Section 5.5). Let Z_n be a branching process with $Z_0 = 1$ and the offspring numbers having mean μ and variance σ^2. Since $\mathbb{E}(Z_{n+1}|Z_n) = \mu Z_n$, the ratio $W_n = \mu^{-n} Z_n$ is a martingale with

$$\mathbb{E}(W_n^2) = 1 + (\sigma/\mu)^2(1 + \mu^{-1} + \ldots + \mu^{-n+1}).$$

In the supercritical case, $\mu > 1$, we have $\mathbb{E}(W_n^2) \to 1 + \frac{\sigma^2}{\mu(\mu-1)}$, and there is a r.v. W such that $W_n \to W$ a.s. and in L^2.

The Laplace transform of the limit $\phi(\theta) = \mathbb{E}(e^{-\theta W})$ satisfies the functional equation $\phi(\mu\theta) = h(\phi(\theta))$.

8.3 Doob's decomposition

Definition 8.17 The sequence $(S_n, \mathcal{F}_n)$ is called predictable if $S_0 = 0$, and S_n is $\mathcal{F}_{n-1}$-measurable for all $n \geq 1$. It is also called increasing if $\mathbb{P}(S_n \leq S_{n+1}) = 1$ for all $n \geq 0$.

Theorem 8.18 *Doob's decomposition. A submartingale $(Y_n, \mathcal{F}_n)$ with finite means can be expressed in the form $Y_n = M_n + S_n$, where $(M_n, \mathcal{F}_n)$ is a martingale and $(S_n, \mathcal{F}_n)$ is an increasing predictable process (called the compensator of the submartingale). This decomposition is unique.*

Proof. We define M and S explicitly: $M_0 = Y_0$, $S_0 = 0$, and for $n \geq 0$

$$M_{n+1} - M_n = Y_{n+1} - \mathbb{E}(Y_{n+1}|\mathcal{F}_n), \qquad S_{n+1} - S_n = \mathbb{E}(Y_{n+1}|\mathcal{F}_n) - Y_n.$$

To see uniqueness suppose another such decomposition $Y_n = M_n' + S_n'$. Then

$$M_{n+1}' - M_n' + S_{n+1}' - S_n' = M_{n+1} - M_n + S_{n+1} - S_n.$$

Taking conditional expectations given $\mathcal{F}_n$ we get $S_{n+1}' - S_n' = S_{n+1} - S_n$. This in view of $S_0' = S_0 = 0$ implies $S_n' = S_n$.

Definition 8.19 Let (Y_n) be adapted to $(\mathcal{F}_n)$, and (S_n) be predictable. The sequence

$$Z_n = Y_0 + \sum_{i=1}^{n} S_i(Y_i - Y_{i-1})$$

is called the transform of (Y_n) by (S_n).

Example 8.20 Such transforms with $S_n \geq 0$ are usually interpreted as gambling systems with (Y_n) being a supermartingale (the capital after n gambles each involving a unit stake). Optional skipping is one such strategy. Here the gambler either wagers a unit stake or skip the round: S_n equals either 1 or 0. The following result says that you can not systematically beat the casino.

Theorem 8.21 *Let (Z_n) be the transform of (Y_n) by (S_n). Then*

(i) If (Y_n) is a martingale, then (Z_n) is a martingale so long as $\mathbb{E}|Z_n| < \infty$ for all n.

(ii) If (Y_n) is a submartingale and in addition $S_n \geq 0$ for all n, then (Z_n) is a submartingale so long as $\mathbb{E}(Z_n^+) < \infty$ for all n.

(iii) If (Y_n) is a supermartingale and in addition $S_n \geq 0$ for all n, then (Z_n) is a supermartingale so long as $\mathbb{E}(Z_n^-) < \infty$ for all n.

Proof. All three assertions follow immediately from

$$\mathbb{E}(Z_{n+1}|\mathcal{F}_n) - Z_n = \mathbb{E}(S_{n+1}(Y_{n+1} - Y_n)|\mathcal{F}_n) = S_{n+1}(\mathbb{E}(Y_{n+1}|\mathcal{F}_n) - Y_n).$$

Example 8.22 Optional stopping. The gambler wagers a unit stake on each play until a random time T. In this case $S_n = 1_{\{n \leq T\}}$ and $Z_n = Y_{T \wedge n}$. If S_n is predictable, then $\{T = n\} = \{S_n = 1, S_{n+1} = 0\} \in \mathcal{F}_n$, so that T is a stopping time.

Example 8.23 Optional starting. The gambler does not play until the $(T+1)$-th round, where T is a stopping time. In this case $S_n = 1_{\{T \leq n-1\}}$ is a predictable sequence.

8.4 Hoeffding's inequality

Definition 8.24 Let $(Y_n, \mathcal{F}_n)$ be a martingale. The sequence of martingale differences is defined by $D_n = Y_n - Y_{n-1}$, so that D_n is $\mathcal{F}_n$-measurable and

$$\mathbb{E}|D_n| < \infty, \qquad \mathbb{E}(D_{n+1}|\mathcal{F}_n) = 0, \qquad Y_n = Y_0 + D_1 + \ldots + D_n.$$

Theorem 8.25 *Hoeffding's inequality. Let $(Y_n, \mathcal{F}_n)$ be a martingale, and suppose $\mathbb{P}(|D_n| \leq K_n) = 1$ for a sequence of real numbers K_n. Then for any $x > 0$*

$$\mathbb{P}(|Y_n - Y_0| \geq x) \leq 2 \exp\Big(-\frac{x^2}{2(K_1^2 + \ldots + K_n^2)}\Big).$$

Proof. Let $\theta > 0$.

Step 1. The function $e^{\theta x}$ is convex, therefore

$$e^{\theta d} \leq \frac{1}{2}(1-d)e^{-\theta} + \frac{1}{2}(1+d)e^{\theta} \text{ for all } |d| \leq 1.$$

Hence if D is a r.v. with mean 0 such that $\mathbb{P}(|D| \leq 1) = 1$, then

$$\mathbb{E}(e^{\theta D}) \leq \frac{e^{-\theta} + e^{\theta}}{2} = \sum_{k=0}^{\infty} \frac{\theta^{2k}}{(2k)!} < \sum_{k=0}^{\infty} \frac{\theta^{2k}}{2^k(k)!} = e^{\theta^2/2}.$$

Step 2. Using the martingale differences we obtain

$$\mathbb{E}(e^{\theta(Y_n - Y_0)}|\mathcal{F}_{n-1}) = e^{\theta(Y_{n-1} - Y_0)}\mathbb{E}(e^{\theta D_n}|\mathcal{F}_{n-1}) \leq e^{\theta(Y_{n-1} - Y_0)}e^{\theta^2 K_n^2/2}.$$

Take expectations and iterate to find

$$\mathbb{E}(e^{\theta(Y_n - Y_0)}) \leq \mathbb{E}(e^{\theta(Y_{n-1} - Y_0)})e^{\theta^2 K_n^2/2} \leq \exp\Big(\frac{\theta^2}{2}\sum_{i=1}^{n} K_i^2\Big).$$

Step 3. Due to the Markov inequality we have for any $x > 0$

$$\mathbb{P}(Y_n - Y_0 \geq x) \leq e^{-\theta x}\mathbb{E}(e^{\theta(Y_n - Y_0)}) \leq \exp\Big(-\theta x + \frac{\theta^2}{2}\sum_{i=1}^{n} K_i^2\Big).$$

Set $\theta = x/\sum_{i=1}^{n} K_i^2$ to minimize the exponent. Then

$$\mathbb{P}(Y_n - Y_0 \geq x) \leq \exp\Big(-\frac{x^2}{2(K_1^2 + \ldots + K_n^2)}\Big).$$

Since $(-Y_n)$ is also a martingale, we get

$$\mathbb{P}(Y_n - Y_0 \leq -x) = \mathbb{P}(-Y_n + Y_0 \geq x) \leq \exp\Big(-\frac{x^2}{2(K_1^2 + \ldots + K_n^2)}\Big).$$

Example 8.26 Large deviations. Let X_n be iid $\mathrm{Ber}(p)$ random variables. If $S_n = X_1 + \ldots + X_n$, then $Y_n = S_n - np$ is a martingale with

$$|Y_n - Y_{n-1}| = |X_n - p| \leq \max(p, 1-p).$$

Due to the Hoeffding's inequality for any $x > 0$

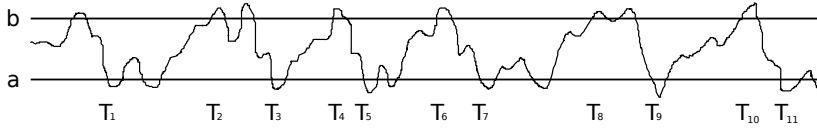$$\mathbb{P}(|S_n - np| \geq x\sqrt{n}) \leq 2\exp\left(-\frac{x^2}{2(\max(p, 1-p))^2}\right).$$

In particular, if $p = 1/2$,

$$\mathbb{P}(|S_n - n/2| \geq x\sqrt{n}) \leq 2e^{-2x^2}.$$

Putting here $x = 3$ we get $\mathbb{P}(|S_n - n/2| \geq 3\sqrt{n}) \leq 3 \cdot 10^{-8}$.

8.5   Convergence in L^1

On the figure below five time intervals are shown: $(T_1, T_2], (T_3, T_4], \ldots, (T_9, T_{10}]$ for the uppcrossings of the interval (a, b). If for all rational intervals (a, b) the number of uppcrossings $U(a, b)$ is finite, then the corresponding trajectory has a (possibly infinite) limit.

All T_k are stopping times. Definition 7.11 for the stopping times can be extended as follows.

Definition 8.27 A random variable T taking values in the set $\{0, 1, 2, \ldots\} \cup \{\infty\}$ is called a stopping time with respect to the filtration $(\mathcal{F}_n)$, if

$$\{T = n\} \in \mathcal{F}_n, \quad \text{for all } n = 0, 1, 2, \ldots.$$

Exercise 8.28 Define more exactly the crossing times T_k. Let $(Y_n, \mathcal{F}_n)_{n \geq 0}$ be an adapted process, and take $a < b$. Put $T_1 = n$ for the smallest n such that $Y_n \leq a$. In particular, if $Y_0 \leq a$, then $T_1 = 0$.

For an even natural number $k \geq 2$ put $T_k = n$ if $Y_i < b$ for $T_{k-1} \leq i \leq n-1$, and $Y_n \geq b$. For an odd natural number $k \geq 3$ put $T_k = n$ if $Y_i > a$ for $T_{k-1} \leq i \leq n-1$, and $Y_n \leq a$. Show that all crossing times T_k are stopping times.

Lemma 8.29 *Snell's uppcrossing inequality. Let $a < b$ and $U_n(a, b)$ is the number of uppcrossings of the interval (a, b) for a submartingale $(Y_0, \ldots, Y_n)$. Then*

$$\mathbb{E}[U_n(a, b)] \leq \frac{\mathbb{E}[(Y_n - a)^+]}{b - a}.$$

Proof. Let I_i be the indicator of the event that $i \in (T_{2k-1}, T_{2k}]$ for some k. Note that I_i is $\mathcal{F}_{i-1}$-measurable, since

$$\{I_i = 1\} = \bigcup_k \{T_{2k-1} \leq i-1 < T_{2k}\} = \bigcup_k \{T_{2k-1} \leq i-1\} \cap \{T_{2k} \leq i-1\}^c$$

is an event that depends on $(Y_0, \ldots, Y_{i-1})$ only. Now, since $Z_n = (Y_n - a)^+$ forms a submartingale, we get

$$\mathbb{E}((Z_i - Z_{i-1}) \cdot I_i) = \mathbb{E}[\mathbb{E}(I_i \cdot (Z_i - Z_{i-1})|\mathcal{F}_{i-1})] = \mathbb{E}[I_i \cdot (\mathbb{E}(Z_i|\mathcal{F}_{i-1}) - Z_{i-1})]$$
$$\leq \mathbb{E}[\mathbb{E}(Z_i|\mathcal{F}_{i-1}) - Z_{i-1}] = \mathbb{E}(Z_i) - \mathbb{E}(Z_{i-1}).$$

Now, observe that (inspect the figure above)

$$(b - a) \cdot U_n(a, b) \leq \sum_{i=1}^{n} (Z_i - Z_{i-1}) I_i.$$

Thus

$$(b - a) \cdot \mathbb{E}(U_n(a, b)) \leq \mathbb{E}(Z_n) - \mathbb{E}(Z_0) \leq \mathbb{E}(Z_n).$$

Theorem 8.30 *Suppose $(Y_n, \mathcal{F}_n)$ is a submartingale such that for some finite constant M,*

$$\sup_{n \geq 1} \mathbb{E}(Y_n^+) \leq M.$$

(i) There exists a r.v. Y such that $Y_n \to Y$ almost surely. In addition: (ii) if $\mathbb{E}|Y_0| < \infty$, the limit Y has a finite mean, and (iii) if (Y_n) is uniformly integrable, see Definition 3.22, then $Y_n \to Y$ in L^1.

Proof. (i) Using Snell's inequality we obtain that $U(a, b) = \lim U_n(a, b)$ satisfies

$$\mathbb{E}(U(a, b)) \leq \frac{M + |a|}{b - a}.$$

Therefore, $\mathbb{P}(U(a, b) < \infty) = 1$. Since there are only countably many rationals, it follows that with probability 1, $U(a, b) < \infty$ for all rational (a, b), and $Y_n \to Y$ almost surely.

(ii) We have to check that $\mathbb{E}|Y| < \infty$ given $\mathbb{E}|Y_0| < \infty$. Indeed, since $|Y_n| = 2Y_n^+ - Y_n$ and $\mathbb{E}(Y_n | \mathcal{F}_0) \geq Y_0$, we get

$$\mathbb{E}(|Y_n| \big| \mathcal{F}_0) \leq 2\mathbb{E}(Y_n^+ \big| \mathcal{F}_0) - Y_0.$$

By Fatou's lemma

$$\mathbb{E}(|Y| \big| \mathcal{F}_0) = \mathbb{E}(\liminf_{n \to \infty} |Y_n| \big| \mathcal{F}_0) \leq \liminf_{n \to \infty} \mathbb{E}(|Y_n| \big| \mathcal{F}_0) \leq 2 \liminf_{n \to \infty} \mathbb{E}(Y_n^+ \big| \mathcal{F}_0) - Y_0,$$

and it remains to observe that $\mathbb{E}(\liminf_{n \to \infty} \mathbb{E}(Y_n^+ \big| \mathcal{F}_0)) \leq M$, again due to Fatou's lemma.

(iii) Finally, recall that according to Theorem 3.25, given $Y_n \xrightarrow{P} Y$, the uniform integrability of (Y_n) is equivalent to $\mathbb{E}|Y_n| < \infty$ for all n, $\mathbb{E}|Y| < \infty$, and $Y_n \xrightarrow{L^1} Y$.

Corollary 8.31 *Any martingale, submartingale or supermartingale $(Y_n, \mathcal{F}_n)$ satisfying $\sup_n \mathbb{E}|Y_n| \leq M$ converges almost surely to a r.v. with a finite limit.*

Corollary 8.32 *A non-negative supermartingale converges almost surely. A non-positive submartingale converges almost surely.*

Example 8.33 De Moivre martingale $Y_n = (q/p)^{S_n}$ is non-negative and hence converges a.s. to some limit Y. Let $p \neq q$. Since $S_n \to \infty$ for $p > q$ and $S_n \to -\infty$ for $p < q$, in both cases we get $\mathbb{P}(Y = 0) = 1$. Note that Y_n does not converge in mean, since $\mathbb{E}(Y_n) = \mathbb{E}(Y_0) \neq 0$.

Example 8.34 Doob's martingale. Let Z be a r.v. on $(\Omega, \mathcal{F}, \mathbb{P})$ such that $\mathbb{E}(|Z|) < \infty$. For a filtration $(\mathcal{F}_n)$ define $Y_n = \mathbb{E}(Z | \mathcal{F}_n)$. Then $(Y_n, \mathcal{F}_n)$ is a martingale: first, by Jensen's inequality,

$$\mathbb{E}(|Y_n|) = \mathbb{E}|\mathbb{E}(Z | \mathcal{F}_n)| \leq \mathbb{E}(\mathbb{E}(|Z| | \mathcal{F}_n)) = \mathbb{E}(|Z|),$$

and secondly

$$\mathbb{E}(Y_{n+1} | \mathcal{F}_n) = \mathbb{E}(\mathbb{E}(Z | \mathcal{F}_{n+1}) | \mathcal{F}_n) = \mathbb{E}(Z | \mathcal{F}_n) = Y_n.$$

As we show next the Doob martingale is uniformly integrable. Again due to Jensen's inequality,

$$|Y_n| = |\mathbb{E}(Z | \mathcal{F}_n)| \leq \mathbb{E}(|Z| | \mathcal{F}_n) =: Z_n,$$

so that for any a

$$|Y_n| 1_{\{|Y_n| \geq a\}} \leq Z_n 1_{\{Z_n \geq a\}}.$$

By the Definition 2.7 of conditional expectation we have

$$\mathbb{E}\big((|Z| - Z_n)1_{\{Z_n \geq a\}}\big) = 0,$$

which together with the previous inequality imply

$$\mathbb{E}(|Y_n|1_{\{|Y_n|\geq a\}}) \leq \mathbb{E}(|Z|1_{\{Z_n \geq a\}}).$$

To derive from here the uniform integrability

$$\sup_n \mathbb{E}(|Y_n|1_{\{|Y_n|\geq a\}}) \to 0, \quad a \to \infty,$$

it is enough to show that

$$\sup_n \mathbb{E}(|Z|1_{\{Z_n \geq a\}}) \to 0, \quad a \to \infty.$$

The latter is obtained using the Markov inequality:

$$\mathbb{E}(|Z|1_{\{Z_n \geq a\}}) \leq C \cdot \mathbb{P}(Z_n \geq a) + \mathbb{E}(|Z|1_{\{|Z|\geq C\}}) \leq C \cdot \frac{\mathbb{E}(Z_n)}{a} + \mathbb{E}(|Z|1_{\{|Z|\geq C\}}),$$

for arbitrary $C > 0$, implying that

$$\sup_n \mathbb{E}(|Z|1_{\{Z_n \geq a\}}) \leq C \cdot \frac{\mathbb{E}|Z|}{a} + \mathbb{E}(|Z|1_{\{|Z|\geq C\}}).$$

It remains to send first $a \to \infty$ and then $C \to \infty$.

By Theorem 8.30, Doob's martingale converges $\mathbb{E}(Z|\mathcal{F}_n) \to Y$ a.s. and in mean as $n \to \infty$. It is actually the case (without proof) that the limit $Y = \mathbb{E}(Z|\mathcal{F}_\infty)$, where $\mathcal{F}_\infty$ is the smallest σ-algebra containing all $\mathcal{F}_n$. There is an important converse result (without proof): if a martingale $(Y_n, \mathcal{F}_n)$ converges in mean, then there exists a r.v. Z with finite mean such that $Y_n = \mathbb{E}(Z|\mathcal{F}_n)$.

Exercise 8.35 Knapsack problem. It is required to pack a knapsack of volume c to maximum benefit. Suppose you have n objects, the ith object having volume V_i and worth W_i, where $V_1, \ldots, V_n, W_1, \ldots, W_n$ are independent non-negative random variables with finite means. You wish to find a vector $z_1, \ldots, z_n$ of 0's and 1's such that

$$z_1 V_1 + \ldots + z_n V_n \leq c$$

and which maximizes $W = z_1 W_1 + \ldots + z_n W_n$. Let Z be the maximal possible W, and show that given $W_i \leq 1$ for all i,

$$\mathbb{P}(|Z - \mathbb{E}Z| \geq x) \leq 2e^{-\frac{x^2}{2n}}.$$

In particular,

$$\mathbb{P}(|Z - \mathbb{E}Z| \geq 3\sqrt{n}) \leq 0.022.$$

Solution. Consider a filtration $\mathcal{F}_0 = \{\emptyset, \Omega\}$ and $\mathcal{F}_i = \sigma(V_1, W_1, \ldots, V_i, W_i)$ for $i = 1, \ldots, n$. And let $Y_i = \mathbb{E}(Z|\mathcal{F}_i)$, $i = 0, 1, \ldots, n$ be a Doob martingale. The assertion follows from the Hoeffding inequality if we verify that $|Y_i - Y_{i-1}| \leq 1$. To this end define $Z(j)$ as the maximal worth attainable without using the jth object. Clearly,

$$\mathbb{E}(Z(j)|\mathcal{F}_j) = \mathbb{E}(Z(j)|\mathcal{F}_{j-1}).$$

Since obviously $Z(j) \leq Z \leq Z(j) + 1$, we have

$$Y_i - Y_{i-1} = \mathbb{E}(Z|\mathcal{F}_i) - \mathbb{E}(Z|\mathcal{F}_{i-1}) \leq 1 + \mathbb{E}(Z(i)|\mathcal{F}_i) - \mathbb{E}(Z(i)|\mathcal{F}_{i-1}) = 1,$$

and

$$Y_i - Y_{i-1} = \mathbb{E}(Z|\mathcal{F}_i) - \mathbb{E}(Z|\mathcal{F}_{i-1}) \geq \mathbb{E}(Z(i)|\mathcal{F}_i) - \mathbb{E}(Z(i)|\mathcal{F}_{i-1}) - 1 = -1.$$

8.6 Bounded stopping times. Optional sampling theorem

The stopped de Moivre martingale from Example 8.10 is also a martingale. A general statement of this type follows next.

Theorem 8.36 *Let $(Y_n, \mathcal{F}_n)$ be a submartingale and let T be a stopping time. Then $(Y_{T\wedge n}, \mathcal{F}_n)$ is a submartingale. If moreover, $\mathbb{E}|Y_n| < \infty$, then $(Y_n - Y_{T\wedge n}, \mathcal{F}_n)$ is also a submartingale.*

Proof. The random variable $Z_n = Y_{T\wedge n}$ is $\mathcal{F}_n$-measurable:

$$Z_n = \sum_{i=0}^{n-1} Y_i 1_{\{T=i\}} + Y_n 1_{\{T\geq n\}},$$

with

$$\mathbb{E}(Z_n^+) \leq \sum_{i=0}^{n} \mathbb{E}(Y_i^+) < \infty.$$

It remains to see that

$$Z_{n+1} - Z_n = (Y_{n+1} - Y_n)1_{\{T>n\}}$$

implies one hand,

$$\mathbb{E}(Z_{n+1} - Z_n|\mathcal{F}_n) = \mathbb{E}(Y_{n+1} - Y_n|\mathcal{F}_n)1_{\{T>n\}} \geq 0,$$

and on the other hand,

$$\mathbb{E}(Y_{n+1} - Z_{n+1} - Y_n + Z_n|\mathcal{F}_n) = \mathbb{E}(Y_{n+1} - Y_n|\mathcal{F}_n)1_{\{T\leq n\}} \geq 0.$$

Corollary 8.37 *If $(Y_n, \mathcal{F}_n)$ is a martingale, then it is both a submartingale and a supermartingale, and therefore, for a given stopping time T, both $(Y_{T\wedge n}, \mathcal{F}_n)$ and $(Y_n - Y_{T\wedge n}, \mathcal{F}_n)$ are martingales.*

Definition 8.38 For a stopping time T with respect to filtration $(\mathcal{F}_n)$, denote by $\mathcal{F}_T$ the σ-algebra of all events A such that

$$A \cap \{T = n\} \in \mathcal{F}_n \quad \text{for all } n.$$

The stopping time T is called bounded if $\mathbb{P}(T \leq N) = 1$ for some finite constant N.

Theorem 8.39 *Optional sampling. Let $(Y_n, \mathcal{F}_n)$ be a submartingale.*
(i) If T is a bounded stopping time, then $\mathbb{E}(Y_T^+) < \infty$ and $\mathbb{E}(Y_T|\mathcal{F}_0) \geq Y_0$.
(ii) If $0 = T_0 \leq T_1 \leq T_2 \leq \ldots$ is a sequence of bounded stopping times, then $(Y_{T_j}, \mathcal{F}_{T_j})$ is a submartingale.

Proof. (i) Let $\mathbb{P}(T \leq N) = 1$ where N is a positive constant. Since $Z_n = Y_{T\wedge n}$ is a submartingale and $Z_N = Y_T$, we have $\mathbb{E}(Y_T^+) < \infty$ and $\mathbb{E}(Y_T|\mathcal{F}_0) \geq Y_0$.

(ii) It is easy to see that for an arbitrary stopping time S, the random variable Y_S is $\mathcal{F}_S$-measurable:

$$\{Y_S \leq x\} \cap \{S = n\} = \{Y_n \leq x\} \cap \{S = n\} \in \mathcal{F}_n.$$

Consider two bounded stopping times $S \leq T \leq N$. We have to show that $W := \mathbb{E}(Y_T|\mathcal{F}_S) \geq Y_S$. For this, observe that for $A \in \mathcal{F}_S$ we have

$$\mathbb{E}(W1_A) = \mathbb{E}(Y_T 1_A) = \sum_{k\leq N} \mathbb{E}(Y_T 1_{A\cap\{S=k\}}) = \sum_{k\leq N} \mathbb{E}\Big(1_{A\cap\{S=k\}}\mathbb{E}(Y_T|\mathcal{F}_k)\Big),$$

and since in view of Theorem 8.36,

$$\mathbb{E}(Y_T|\mathcal{F}_k) = \mathbb{E}(Y_{T\wedge N}|\mathcal{F}_k) \geq Y_{T\wedge k} \quad \text{for all } k \leq N,$$

we conclude

$$\mathbb{E}(W1_A) \geq \mathbb{E}\Big(\sum_{k\leq N} 1_{A\cap\{S=k\}}Y_{T\wedge k}\Big) = \mathbb{E}\Big(\sum_{k\leq N} 1_{A\cap\{S=k\}}Y_k\Big) = \mathbb{E}(Y_S 1_A),$$

so that $\mathbb{E}((W - Y_S)1_A) \geq 0$ for all $A \in \mathcal{F}_S$. In particular, for any $\epsilon > 0$ and $A_\epsilon = \{W \leq Y_S - \epsilon\}$, we have

$$0 \leq \mathbb{E}((W - Y_S)1_{A_\epsilon}) \leq -\epsilon\mathbb{P}(A_\epsilon)$$

implying $\mathbb{P}(A_\epsilon) = 0$. Thus $W \geq Y_S$ with probability 1.

Exercise 8.40 Let the process (Y_n) be a martingale and T be a bounded stopping time. Show that according to Theorem 8.39 we have $\mathbb{E}(Y_T|\mathcal{F}_0) = Y_0$ and $\mathbb{E}(Y_T) = \mathbb{E}(Y_0)$.

8.7 Unbounded stopping times

The martingale property in Exercise 8.40 is not enough for Example 8.9 because the corresponding stopping time T is not bounded. However, according to Section 5.1,

$$\mathbb{E}(T) = \frac{1}{q-p}\left[k - N\cdot\frac{1-(q/p)^k}{1-(q/p)^N}\right]$$

is finite and the following theorems work.

Theorem 8.41 *Optional stopping. Let $(Y_n,\mathcal{F}_n)$ be a martingale and T be a stopping time. Then $\mathbb{E}(Y_T) = \mathbb{E}(Y_0)$ if (a) $\mathbb{P}(T < \infty) = 1$, (b) $\mathbb{E}|Y_T| < \infty$, and (c) $\mathbb{E}(Y_n 1_{\{T>n\}}) \to 0$ as $n \to \infty$.*

Proof. From $Y_T = Y_{T\wedge n} + (Y_T - Y_n)1_{\{T>n\}}$ using that $\mathbb{E}(Y_{T\wedge n}) = \mathbb{E}(Y_0)$ we obtain

$$\mathbb{E}(Y_T) = \mathbb{E}(Y_0) + \mathbb{E}(Y_T 1_{\{T>n\}}) - \mathbb{E}(Y_n 1_{\{T>n\}}).$$

It remains to apply (c) and observe that due to the dominated convergence $\mathbb{E}(Y_T 1_{\{T>n\}}) \to 0$.

Theorem 8.42 *Let $(Y_n, \mathcal{F}_n)$ be a martingale and T be a stopping time. Then $\mathbb{E}(Y_T) = \mathbb{E}(Y_0)$ if (a) $\mathbb{E}(T) < \infty$ and (b) there exists a constant c such that for any n*

$$\mathbb{E}(|Y_{n+1} - Y_n|\big|\mathcal{F}_n)1_{\{T>n\}} \le c1_{\{T>n\}}.$$

Proof. Since $T \wedge n \to T$, we have $Y_{T\wedge n} \to Y_T$ a.s. It follows that

$$\mathbb{E}(Y_0) = \mathbb{E}(Y_{T\wedge n}) \to \mathbb{E}(Y_T)$$

as long as $(Y_{T\wedge n})$ is uniformly integrable. By Exercise 3.23, to prove the uniform integrability it is enough to observe that

$$\sup_n |Y_{T\wedge n}| \le |Y_0| + W, \quad W := |Y_1 - Y_0| + \ldots + |Y_T - Y_{T-1}|,$$

and to verify that $\mathbb{E}(W) < \infty$. Indeed, since $\mathbb{E}(|Y_i - Y_{i-1}|1_{\{T\ge i\}}\big|\mathcal{F}_{i-1}) \le c1_{\{T\ge i\}}$, we have $\mathbb{E}(|Y_i - Y_{i-1}|1_{\{T\ge i\}}) \le c\mathbb{P}(T \ge i)$ and therefore

$$\mathbb{E}(W) = \mathbb{E}\Big(\sum_{i=1}^T |Y_i - Y_{i-1}|\Big) = \sum_{i=1}^\infty \mathbb{E}(|Y_i - Y_{i-1}|1_{\{T\ge i\}}) = \sum_{i=1}^\infty \mathbb{E}\Big(\mathbb{E}(|Y_i - Y_{i-1}|1_{\{T\ge i\}}\big|\mathcal{F}_{i-1})\Big)$$

$$\le c\sum_{i=1}^\infty \mathbb{P}(T \ge i) \le c\mathbb{E}(T) < \infty.$$

Example 8.43 Wald's equality. Let (X_n) be iid r.v. with finite mean μ and $S_n = X_1 + \ldots + X_n$, then $Y_n = S_n - n\mu$ is a martingale with respect to $\mathcal{F}_n = \sigma\{X_1, \ldots, X_n\}$. Now

$$\mathbb{E}(|Y_{n+1} - Y_n|\big|\mathcal{F}_n) = \mathbb{E}|X_{n+1} - \mu| = \mathbb{E}|X_1 - \mu| < \infty.$$

We deduce from Theorem 8.42 that $\mathbb{E}(Y_T) = \mathbb{E}(Y_0)$ for any stopping time T with finite mean, implying that $\mathbb{E}(S_T) = \mu\mathbb{E}(T)$.

Lemma 8.44 *Wald's identity. Let (X_n) be iid r.v. with a finite $M(t) = \mathbb{E}(e^{tX})$. Put $S_n = X_1 + \ldots + X_n$. If T is a stopping time with finite mean such that $|S_n|1_{\{T>n\}} \le c1_{\{T>n\}}$, then*

$$\mathbb{E}(e^{tS_T}M(t)^{-T}) = 1 \text{ whenever } M(t) \ge 1.$$

Proof. Define $Y_0 = 1$, $Y_n = e^{tS_n}M(t)^{-n}$, and let $\mathcal{F}_n = \sigma\{X_1, \ldots, X_n\}$. It is clear that (Y_n) is a martingale and thus the claim follows from Theorem 8.42 if verify condition (b). Indeed,

$$\mathbb{E}(|Y_{n+1} - Y_n|\big|\mathcal{F}_n) = Y_n\mathbb{E}|e^{tX}M(t)^{-1} - 1| \le Y_n\mathbb{E}(e^{tX}M(t)^{-1} + 1) = 2Y_n.$$

Furthermore, given $M(t) \ge 1$,

$$Y_n 1_{\{T>n\}} = e^{tS_n}M(t)^{-n}1_{\{T>n\}} \le e^{c|t|}1_{\{T>n\}}.$$

Example 8.45 Simple random walk S_n with $\mathbb{P}(X_i = 1) = p$ and $\mathbb{P}(X_i = -1) = q$. Let T be the first exit time of $(-a, b)$. By Lemma 8.44 with $M(t) = pe^t + qe^{-t}$,

$$e^{-at}\mathbb{E}(M(t)^{-T}1_{\{S_T=-a\}}) + e^{bt}\mathbb{E}(M(t)^{-T}1_{\{S_T=b\}}) = 1 \text{ whenever } M(t) \geq 1.$$

Setting $M(t) = s^{-1}$ in $M(t) = pe^t + qe^{-t}$ we obtain a quadratic equation for e^t having two solutions $\lambda_i = \lambda_i(s)$:

$$\lambda_1 = \frac{1 + \sqrt{1 - 4pqs^2}}{2ps}, \qquad \lambda_2 = \frac{1 - \sqrt{1 - 4pqs^2}}{2ps}, \quad s \in [0, 1].$$

They give us two linear equations

$$\lambda_i^{-a}\mathbb{E}(s^T 1_{\{S_T=-a\}}) + \lambda_i^b \mathbb{E}(s^T 1_{\{S_T=b\}}) = 1, \quad i = 1, 2,$$

resulting in

$$\mathbb{E}(s^T 1_{\{S_T=-a\}}) = \frac{\lambda_1^a \lambda_2^a(\lambda_1^b - \lambda_2^b)}{\lambda_1^{a+b} - \lambda_2^{a+b}}, \qquad \mathbb{E}(s^T 1_{\{S_T=b\}}) = \frac{\lambda_1^a - \lambda_2^a}{\lambda_1^{a+b} - \lambda_2^{a+b}}.$$

Summing up these two relations we get the probability generating function

$$\mathbb{E}(s^T) = \frac{\lambda_1^a(1 - \lambda_2^{a+b}) + \lambda_2^a(\lambda_1^{a+b} - 1)}{\lambda_1^{a+b} - \lambda_2^{a+b}}.$$

8.8 Maximal inequality

Theorem 8.46 *Maximal inequality.*

(i) If (Y_n) is a submartingale, then for any $\epsilon > 0$

$$\mathbb{P}(\max_{0 \leq i \leq n} Y_i \geq \epsilon) \leq \frac{\mathbb{E}(Y_n^+)}{\epsilon}.$$

(ii) If (Y_n) is a supermartingale and $\mathbb{E}|Y_0| < \infty$, then for any $\epsilon > 0$

$$\mathbb{P}(\max_{0 \leq i \leq n} Y_i \geq \epsilon) \leq \frac{\mathbb{E}(Y_0) + \mathbb{E}(Y_n^-)}{\epsilon}.$$

Proof. (i) If (Y_n) is a submartingale, then (Y_n^+) is a non-negative submartingale with finite means. Introduce a stopping time

$$T := \min\{n : Y_n \geq \epsilon\} = \min\{n : Y_n^+ \geq \epsilon\}$$

and notice that

$$\{T \leq n\} = \{\max_{0 \leq i \leq n} Y_i \geq \epsilon\}.$$

By the second part of Theorem 8.36, $\mathbb{E}(Y_n^+ - Y_{T \wedge n}^+) \geq 0$. Therefore,

$$\mathbb{E}(Y_n^+) \geq \mathbb{E}(Y_{T \wedge n}^+) = \mathbb{E}(Y_T^+ 1_{\{T \leq n\}}) + \mathbb{E}(Y_n^+ 1_{\{T > n\}}) \geq \mathbb{E}(Y_T^+ 1_{\{T \leq n\}}) \geq \epsilon \mathbb{P}(T \leq n),$$

implying the first stated inequality.

Furthermore, since $\mathbb{E}(Y_{T \wedge n}^+ 1_{\{T > n\}}) = \mathbb{E}(Y_n^+ 1_{\{T > n\}})$, we have

$$\mathbb{E}(Y_n^+ 1_{\{T \leq n\}}) \geq E(Y_{T \wedge n}^+ 1_{\{T \leq n\}}) = \mathbb{E}(Y_T^+ 1_{\{T \leq n\}}) \geq \epsilon \mathbb{P}(T \leq n).$$

Using this we get even a stronger inequality

$$\mathbb{P}(A) \leq \frac{\mathbb{E}(Y_n^+ 1_A)}{\epsilon}, \text{ where } A = \{\max_{0 \leq i \leq n} Y_i \geq \epsilon\}. \tag{15}$$

(ii) If (Y_n) is a supermartingale, then by the first part of Theorem 8.36, $\mathbb{E}(-Y_{T \wedge n}) \geq \mathbb{E}(-Y_0)$. Thus

$$\mathbb{E}(Y_0) \geq \mathbb{E}(Y_{T \wedge n}) = \mathbb{E}(Y_T I_{\{T \leq n\}}) + \mathbb{E}(Y_n I_{\{T > n\}}) \geq \epsilon \mathbb{P}(T \leq n) - \mathbb{E}(Y_n^-)$$

giving the second assertion.

Corollary 8.47 *If (Y_n) is a submartingale and $\mathbb{E}|Y_0| < \infty$, then for any $\epsilon > 0$*

$$\mathbb{P}(\max_{0 \leq i \leq n} |Y_i| \geq \epsilon) \leq \frac{2\mathbb{E}(Y_n^+) - \mathbb{E}(Y_0)}{\epsilon}.$$

Proof. Let $\epsilon > 0$. If (Y_n) is a submartingale, then $(-Y_n)$ is a supermartingale so that according to (ii),

$$\mathbb{P}(\min_{0 \leq i \leq n} Y_i \leq -\epsilon) = \mathbb{P}(\max_{0 \leq i \leq n} (-Y_i) \geq \epsilon) \leq \frac{\mathbb{E}(-Y_0) + \mathbb{E}[(-Y_n)^-]}{\epsilon} = \frac{\mathbb{E}(Y_n^+) - \mathbb{E}(Y_0)}{\epsilon}.$$

Adding this to (i) we arrive at the asserted inequality.

Exercise 8.48 Show that if (Y_n) is a martingale, then

$$\mathbb{P}(\max_{0 \leq i \leq n} |Y_i| \geq \epsilon) \leq \frac{\mathbb{E}|Y_n|}{\epsilon}.$$

Corollary 8.49 *Doob-Kolmogorov's inequality. If (Y_n) is a martingale with finite second moments, then (Y_n^2) is a submartingale for any $\epsilon > 0$*

$$\mathbb{P}(\max_{1 \leq i \leq n} |Y_i| \geq \epsilon) = \mathbb{P}(\max_{1 \leq i \leq n} Y_i^2 \geq \epsilon^2) \leq \frac{\mathbb{E}(Y_n^2)}{\epsilon^2}.$$

Corollary 8.50 *Kolmogorov's inequality. Let (X_n) are independent r.v. with zero means and finite variances (σ_n^2), then for any $\epsilon > 0$*

$$\mathbb{P}(\max_{1 \leq i \leq n} |X_1 + \ldots + X_i| \geq \epsilon) \leq \frac{\sigma_1^2 + \ldots + \sigma_n^2}{\epsilon^2}.$$

Theorem 8.51 *Convergence in L^r. Let $r > 1$. Suppose $(Y_n, \mathcal{F}_n)$ is a martingale such that $\mathbb{E}(|Y_n|^r) \leq M$ for some constant M and all n. Then $Y_n \to Y$ in L^r, where Y is the a.s. limit of Y_n as $n \to \infty$.*

Proof. Combining Corollary 8.31 and Lyapunov's inequality we get the a.s. convergence $Y_n \to Y$. We prove $Y_n \xrightarrow{L^r} Y$ using Theorem 3.26. For this we need to verify that (Y_n^r) is uniformly integrable. Observe first that

$$\bar{Y}_n := \max_{0 \leq i \leq n} |Y_i|$$

have finite r-th moment

$$\mathbb{E}(\bar{Y}_n^r) \leq \mathbb{E}((|Y_0| + \ldots + |Y_n|)^r) < \infty.$$

By Exercise 3.20,

$$\mathbb{E}(\bar{Y}_n^r) = \int_0^\infty r x^{r-1} \mathbb{P}(\bar{Y}_n > x) dx.$$

Applying (15) to $A(x) = \{\bar{Y}_n \geq x\}$, we obtain

$$\mathbb{E}(\bar{Y}_n^r) \leq \int_0^\infty r x^{r-2} \mathbb{E}(|Y_n| 1_{A(x)}) dx = \mathbb{E}\left(|Y_n| \int_0^\infty r x^{r-2} 1_{A(x)} dx\right) = \frac{r}{r-1} \mathbb{E}(|Y_n| \bar{Y}_n^{r-1}).$$

By Hölder's inequality,

$$\mathbb{E}(|Y_n| \bar{Y}_n^{r-1}) \leq (\mathbb{E}(|Y_n|^r))^{1/r} (\mathbb{E}(\bar{Y}_n^r))^{(r-1)/r}$$

and we conclude

$$\mathbb{E}(\max_{0 \leq i \leq n} |Y_i|^r) = \mathbb{E}(\bar{Y}_n^r) \leq \left(\frac{r}{r-1}\right)^r \mathbb{E}(|Y_n|^r) \leq \left(\frac{r}{r-1}\right)^r M.$$

Thus by monotone convergence $\mathbb{E}(\sup_n |Y_n|^r) < \infty$ implying that (Y_n^r) is uniformly integrable, see Exercise 3.23.

8.9 Backward martingales. Strong LLN

Definition 8.52 Let $(\mathcal{G}_n)$ be a decreasing sequence of σ-algebras:

$$\mathcal{G}_0 \supset \mathcal{G}_1 \supset \mathcal{G}_2 \supset \dots,$$

and (Y_n) be a sequence of adapted r.v. The sequence $(Y_n, \mathcal{G}_n)$ is called a backward (or reversed) martingale if, for all $n \geq 0$,

$$\mathbb{E}(|Y_n|) < \infty,$$
$$\mathbb{E}(Y_n|\mathcal{G}_{n+1}) = Y_{n+1}.$$

Theorem 8.53 *Let $(Y_n, \mathcal{G}_n)$ be a backward martingale. Then Y_n converges to a limit Y almost surely and in mean.*

Proof. The sequence $Y_n = \mathbb{E}(Y_0|\mathcal{G}_n)$ is uniformly integrable, see the proof in Example 8.34. Therefore, it suffices to prove a.s. convergence. Applying Lemma 8.29 to the martingale $(Y_n, \mathcal{G}_n), \dots, (Y_0, \mathcal{G}_0)$ we obtain $\mathbb{E}(U_n(a,b)) \leq \frac{\mathbb{E}((Y_0-a)^+)}{b-a}$ for the number $U_n(a,b)$ of $[a,b]$ uppcrossings by $(Y_n, \dots, Y_0)$. Now let $n \to \infty$ and repeat the proof of Theorem 8.30 to get the required a.s. convergence.

Theorem 8.54 *Strong LLN. Let $X_1, X_2, \dots$ be iid random variables defined on the same probability space. If $\mathbb{E}|X_1| < \infty$, then*

$$\frac{X_1 + \dots + X_n}{n} \to \mathbb{E}X_1$$

almost surely and in L^1. On the other hand, if

$$\frac{X_1 + \dots + X_n}{n} \overset{a.s.}{\to} \mu$$

for some constant μ, then $\mathbb{E}|X_1| < \infty$.

Proof. Let $\mathbb{E}|X_1| < \infty$. Set $S_n = X_1 + \dots + X_n$ and let $\mathcal{G}_n = \sigma(S_n, S_{n+1}, \dots)$, then by symmetry,

$$\mathbb{E}(S_n|\mathcal{G}_{n+1}) = \mathbb{E}(S_n|S_{n+1}) = \sum_{i=1}^{n} \mathbb{E}(X_i|S_{n+1}) = n\mathbb{E}(X_1|S_{n+1}).$$

On the other hand,

$$S_{n+1} = \mathbb{E}(S_{n+1}|S_{n+1}) = (n+1)\mathbb{E}(X_1|S_{n+1}).$$

We conclude that S_n/n is a backward martingale, and according to the Backward Martingale Convergence Theorem there exists Y such that $S_n/n \to Y$ a.s. and in mean. By Kolmogorov's zero-one law, Y is almost surely constant, and hence a.s.

$$Y = \mathbb{E}(S_n/n) = \mathbb{E}(X_1).$$

To prove the second part assume that $S_n/n \overset{a.s.}{\to} \mu$ for some constant μ. Then $X_n/n \overset{a.s.}{\to} 0$ by the theory of convergent real series. Indeed, from $(a_1 + \dots + a_n)/n \to \mu$ it follows that

$$\frac{a_n}{n} = \frac{a_1 + \dots + a_{n-1}}{n(n-1)} + \frac{a_1 + \dots + a_n}{n} - \frac{a_1 + \dots + a_{n-1}}{n-1} \to 0$$

Now, in view of $X_n/n \overset{a.s.}{\to} 0$, the second Borell-Cantelli lemma gives

$$\sum_n \mathbb{P}(|X_n| \geq n) < \infty,$$

since otherwise $\mathbb{P}(n^{-1}|X_n| \geq 1 \text{ i.o.}) = 1$. Thus

$$\mathbb{E}|X_1| = \int_0^\infty \mathbb{P}(|X_1| > x)dx \leq \sum_n \mathbb{P}(|X_1| \geq n) = \sum_n \mathbb{P}(|X_n| \geq n) < \infty.$$

9 Diffusion processes

9.1 The Wiener process

Definition 9.1 The standard Wiener process $W(t) = W_t$ is a continuous time analogue of a simple symmetric random walk. It is characterized by three properties:

- $W_0 = 0$,

- W_t has independent increments with $W_t - W_s \sim N(0, t - s)$ for $0 \le s < t$,

- the path $t \to W_t$ is continuous with probability 1.

Next we sketch a construction of W_t for $t \in [0, 1]$. First observe that the following theorem is not enough.

Theorem 9.2 *Kolmogorov's extension theorem. Assume that for any vector $(t_1, \ldots, t_n)$ with $t_i \in [0, 1]$ there given a joint distribution function $F_{(t_1, \ldots, t_n)}(x_1, \ldots, x_n)$. Suppose that these distribution functions satisfy two consistency conditions*
(i) $F_{(t_1, \ldots, t_n, t_{n+1})}(x_1, \ldots, x_n, \infty) = F_{(t_1, \ldots, t_n)}(x_1, \ldots, x_n)$,
(ii) if π is a permutation of $(1, \ldots, n)$, then $F_{(t_{\pi(1)}, \ldots, t_{\pi(n)})}(x_{\pi(1)}, \ldots, x_{\pi(n)}) = F_{(t_1, \ldots, t_n)}(x_1, \ldots, x_n)$.
Put $\Omega = \{functions\ \omega : [0, 1] \to \mathbb{R}\}$ and $\mathcal{F}$ is the σ-algebra generated by the finite-dimensional sets $\{\omega : \omega(t_i) \in B_i, i = 1, \ldots, n\}$, where B_i are Borel subsets of $\mathbb{R}$. Then there is a unique probability measure $\mathbb{P}$ on $(\Omega, \mathcal{F})$ such that a stochastic process defined by $X(t, \omega) = \omega(t)$ has the finite-dimensional distributions $F_{(t_1, \ldots, t_n)}(x_1, \ldots, x_n)$.

The problem is that the set $\{\omega : t \to \omega(t)$ is continuous$\}$ does not belong to $\mathcal{F}$, since all events in $\mathcal{F}$ may depend on only countably many coordinates.

The above problem can be fixed if we focus of the subset $\mathcal{Q}$ be the set of dyadic rationals $\{t = m2^{-n}$ for some $0 \le m \le 2^n, n \ge 1\}$.

Step 1. Let $(X_{m,n})$ be a collection of Gaussian r.v. such that if we put $X(t) = X_{m,n}$ for $t = m2^{-n}$, then

- $X(0) = 0$,

- $X(t)$ has independent increments with $X(t) - X(s) \sim N(0, t - s)$ for $0 \le s < t$, $s, t \in \mathcal{Q}$.

According to Theorem 9.2 the process $X(t)$, $t \in \mathcal{Q}$ can be defined on $(\Omega_q, \mathcal{F}_q, \mathbb{P}_q)$ where index q means the restriction $t \in \mathcal{Q}$.

Step 2. For $m2^{-n} \le t < (m+1)2^{-n}$ define

$$X_n(t) = X_{m,n} + 2^n(t - m2^{-n})(X_{m+1,n} - X_{m,n}).$$

For each n the process $X_n(t)$ has continuous paths for $t \in [0, 1]$. Think of $X_{n+1}(t)$ as being obtained from $X_n(t)$ by repositioning the centers of the line segments by iid normal amounts. If $t \in \mathcal{Q}$, then $X_n(t) = X(t)$ for all large n. Thus $X_n(t) \to X(t)$ for all $t \in \mathcal{Q}$.

Step 3. Show that $X(t)$ is a.s. uniformly continuous over $t \in \mathcal{Q}$. It follows from $X_n(t) \to X(t)$ a.s. uniformly over $t \in \mathcal{Q}$. To prove the latter we use the Weierstrass M-test by observing that $X_n(t) = Z_1(t) + \ldots + Z_n(t)$, where $Z_i(t) = X_i(t) - X_{i-1}(t)$, and showing that

$$\sum_{i=1}^{\infty} \sup_{t \in \mathcal{Q}} |Z_i(t)| < \infty. \tag{16}$$

Using independence and normality of the increments one can show for $x_i = c\sqrt{i 2^{-i} \log 2}$ that

$$\mathbb{P}(\sup_{t \in \mathcal{Q}} |Z_i(t)| > x_i) \le 2^{i-1} \frac{2^{-ic^2}}{c\sqrt{i \log 2}}.$$

The first Borel-Cantelli lemma implies that for $c > 1$ the events $\sup_{t \in \mathcal{Q}} |Z_i(t)| > x_i$ occur finitely many times and (16) follows.

Step 4. Define $W(t)$ for any $t \in [0, 1]$ by moving our probability measure to $(C, \mathcal{C})$, where $C = $ continuous $\omega : [0, 1) \to \mathbb{R}$ and $\mathcal{C}$ is the σ-algebra generated by the coordinate maps $t \to \omega(t)$. To do this, we observe that the map ψ that takes a uniformly continuous point in Ω_q to its unique continuous extension in C is measurable, and we set $\mathbb{P}(A) = \mathbb{P}_q(\psi^{-1}(A))$.

9.2 Properties of the Wiener process

We will prove some of the following properties of the standard Wiener process.

(i) The vector $(W(t_1), \ldots, W(t_n))$ has the multivariate normal distribution with zero means and covariances $\mathbb{C}ov(W(t_i), W(t_j)) = \min(t_i, t_j)$.

(ii) For any positive s the shifted process $\tilde{W}_t = W_{t+s} - W_s$ is a standard Wiener process. This implies the (weak) Markov property.

(iii) For any non-negative stopping time T the shifted process $W_{t+T} - W_T$ is a standard Wiener process. This implies the strong Markov property.

(iv) Let $T(x) = \inf\{t : W(t) = x\}$ be the first passage time. It is a stopping time for the Wiener process.

(v) The r.v. $M(t) = \max\{W(s) : 0 \leq s \leq t\}$ has the same distribution as $|W(t)|$ and has density function $f(x) = \frac{2}{\sqrt{2\pi t}} e^{-\frac{x^2}{2t}}$ for $x \geq 0$.

(vi) The r.v. $T(x) \stackrel{d}{=} (x/Z)^2$, where $Z \sim \mathrm{N}(0,1)$, has density function $f(t) = \frac{|x|}{\sqrt{2\pi t^3}} e^{-\frac{x^2}{2t}}$ for $t \geq 0$.

(vii) If $\mathcal{F}_t$ is the filtration generated by $(W(u), u \leq t)$, then $(e^{\theta W(t) - \theta^2 t/2}, \mathcal{F}_t)$ is a martingale.

(viii) Consider the Wiener process on $t \in [0, \infty)$ with a negative drift $W(t) - mt$ with $m > 0$. Its maximum is exponentially distributed with parameter $2m$.

Proof of (i) If $0 \leq s < t$, then
$$\mathbb{E}(W(s)W(t)) = \mathbb{E}[W(s)^2 + W(s)(W(t) - W(s))] = \mathbb{E}[W(s)^2] = s.$$

Proof of (v) For $x > 0$ we have $\{T(x) \leq t\} = \{M(t) \geq x\}$. This and
$$\mathbb{P}(M(t) \geq x) = \mathbb{P}(M(t) \geq x, W(t) - x \geq 0) + \mathbb{P}(M(t) \geq x, W(t) - x < 0)$$
imply
$$\mathbb{P}(M(t) \geq x, W(t) < x) = \mathbb{P}(M(t) \geq x, W(t) - W(T(x)) < 0, T(x) \leq t)$$
$$= \mathbb{P}(M(t) \geq x, W(t) - W(T(x)) \geq 0, T(x) \leq t) = \mathbb{P}(M(t) \geq x, W(t) \geq x).$$
Thus
$$\mathbb{P}(M(t) \geq x) = 2\mathbb{P}(M(t) \geq x, W(t) \geq x) = 2\mathbb{P}(W(t) \geq x)$$
$$= \mathbb{P}(W(t) \geq x) + \mathbb{P}(W(t) \leq -x) = \mathbb{P}(|W(t)| \geq x).$$

Proof of (vi) We have
$$\mathbb{P}(T(x) \leq t) = \mathbb{P}(M(t) \geq x) = \mathbb{P}(|W(t)| \geq x) = \frac{2}{\sqrt{2\pi t}} \int_x^\infty e^{-\frac{y^2}{2t}} dy = \int_0^t \frac{|x|}{\sqrt{2\pi u^3}} e^{-\frac{x^2}{2u}} du.$$

Proof of (vii) Bringing $e^{\theta W(s)}$ outside
$$\mathbb{E}(e^{\theta W(t)}|\mathcal{F}_s) = e^{\theta W(s)}\mathbb{E}(e^{\theta(W(t) - W(s))}|\mathcal{F}_s) = e^{\theta W(s)}e^{\theta^2(t-s)/2}.$$

Proof of (viii) It suffices to prove that for any $x > 0$
$$\mathbb{P}(W(t) - mt = x \text{ for some } t) = e^{-2mx}.$$

Let $T(a, b)$ be the first exit time from the interval (a, b). Applying a continuous version of the optional stopping theorem to the martingale $U(t) = e^{2mW(t) - 2m^2 t}$ we obtain $\mathbb{E}(U(T(a, x))) = \mathbb{E}(U(0)) = 1$. Thus
$$1 = e^{2mx}\mathbb{P}(U(T(a, x)) = x) + e^{2ma}\mathbb{P}(U(T(a, x)) = a).$$

Letting $a \to -\infty$ we obtain the desired relation.

9.3 Examples of diffusion processes

Definition 9.3 An Ito diffusion process $X(t) = X_t$ is a Markov process with continuous sample paths characterized by the standard Wiener process W_t in terms of a stochastic differential equation

$$dX_t = \mu(t, X_t)dt + \sigma(t, X_t)dW_t. \tag{17}$$

Here $\mu(t, x)$ and $\sigma^2(t, x)$ are the instantaneous mean and variance for the increments of the diffusion process.

The second term of the stochastic differential equation is defined in terms of the Ito integral leading to the integrated form of the equation

$$X_t - X_0 = \int_0^t \mu(s, X_s)ds + \int_0^t \sigma(s, X_s)dW_s.$$

The Ito integral $J_t = \int_0^t Y_s dW_s$ is defined for a certain class of adapted processes Y_t. The process J_t is a martingale (cf Theorem 8.21).

Example 9.4 The Wiener process with a drift $W_t + mt$ corresponds to $\mu(t, x) = m$ and $\sigma(t, x) = \sigma^2$.

Example 9.5 The Ornstein-Uhlenbeck process: $\mu(t, x) = -\alpha(x - \theta)$ and $\sigma(t, x) = \sigma^2$. Given the initial value X_0 the process is described by the stochastic differential equation

$$dX_t = -\alpha(X_t - \theta)dt + \sigma dW_t, \tag{18}$$

which is a continuous version of an AR(1) process $X_n = aX_{n-1} + Z_n$. This process can be interpreted as the evolution of a phenotypic trait value (like logarithm of the body size) along a lineage of species in terms of the adaptation rate $\alpha > 0$, the optimal trait value θ, and the noise size $\sigma > 0$.

Let $f(t, y|s, x)$ be the density of the distribution of X_t given the position at an earlier time $X_s = x$. Then

$$\text{forward equation} \quad \frac{\partial f}{\partial t} = -\frac{\partial}{\partial y}[\mu(t, y)f] + \frac{1}{2}\frac{\partial^2}{\partial y^2}[\sigma^2(t, y)f],$$

$$\text{backward equation} \quad \frac{\partial f}{\partial s} = -\mu(s, x)\frac{\partial f}{\partial x} - \frac{1}{2}\sigma^2(s, x)\frac{\partial^2 f}{\partial x^2}.$$

Example 9.6 The Wiener process W_t corresponds to $\mu(t, x) = 0$ and $\sigma(t, x) = 1$. The forward and backward equations for the density $f(t, y|s, x) = \frac{1}{\sqrt{2\pi(t-s)}}e^{-\frac{(y-x)^2}{2(t-s)}}$ are

$$\frac{\partial f}{\partial t} = \frac{1}{2}\frac{\partial^2}{\partial y^2}, \quad \frac{\partial f}{\partial s} = -\frac{1}{2}\frac{\partial^2 f}{\partial x^2}.$$

Example 9.7 For $dX_t = \sigma(t)dW_t$ the equations

$$\frac{\partial f}{\partial t} = \frac{\sigma^2(t)}{2}\frac{\partial^2}{\partial y^2}, \quad \frac{\partial f}{\partial s} = -\frac{\sigma^2(t)}{2}\frac{\partial^2 f}{\partial x^2}$$

imply that X_t has a normal distribution with zero mean and variance $\int_0^t \sigma^2(u)du$.

9.4 The Ito formula

Main rule: $(dW_t)^2$ should be replaced by dt.

Theorem 9.8 *Let $f(t, x)$ be twice continuously differentiable on $[0, \infty) \times \mathbb{R}$ and X_t is given by (17). Then $Y_t = f(t, B_t)$ is also an Ito process given by*

$$dY_t = \{f_t(t, X_t) + f_x(t, X_t)\mu(t, X_t) + \frac{1}{2}f_{xx}(t, X_t)\sigma^2(t, X_t)\}dt + f_x(t, X_t)\sigma(t, X_t)dW_t,$$

where

$$f_x(t, X_t) = \frac{\partial}{\partial x}f(t, x)|_{x=X_t}, \quad f_t(t, X_t) = \frac{\partial}{\partial t}f(t, x)|_{x=X_t}, \quad f_{xx}(t, X_t) = \frac{\partial^2}{\partial x^2}f(t, x)|_{x=X_t}.$$

Example 9.9 The distribution of the Ornstein-Uhlenbeck process X_t is normal with

$$\mathbb{E}(X_t) = \theta + e^{-\alpha t}(X_0 - \theta), \quad \mathbb{V}ar(X_t) = \sigma^2(1 - e^{-2\alpha t})/2\alpha \tag{19}$$

implying that X_t looses the effect of the ancestral state X_0 at an exponential rate. In the long run X_0 is forgotten, and the OU–process acquires a stationary normal distribution with mean θ and variance $\sigma^2/2\alpha$.

To verify these formula for the mean and variance we apply the following simple version of Ito's lemma: if $dX_t = \mu_t dt + \sigma_t dW_t$, then for any nice function $f(t, x)$

$$df(t, X_t) = \frac{\partial f}{\partial t}dt + \frac{\partial f}{\partial x}(\mu_t dt + \sigma_t dW_t) + \frac{1}{2}\frac{\partial^2 f}{\partial x^2}\sigma_t^2 dt.$$

Let $f(t, x) = xe^{\alpha t}$. Then using (18) we obtain

$$d(X_t e^{\alpha t}) = \alpha X_t e^{\alpha t}dt + e^{\alpha t}\Big(\alpha(\theta - X_t)dt + \sigma dW_t\Big) = \theta e^{\alpha t}\alpha dt + \sigma e^{\alpha t}dW_t.$$

Integration gives

$$X_t e^{\alpha t} - X_0 = \theta(e^{\alpha t} - 1) + \sigma \int_0^t e^{\alpha u}dW_u,$$

implying (19), since in view of Example 9.7 (see also (8) with $F_t = t$) we have

$$\mathbb{E}\Big(\int_0^t e^{\alpha u}dW_u\Big)^2 = \int_0^t e^{2\alpha u}du = \frac{e^{2\alpha t} - 1}{2\alpha}.$$

Observe that the correlation coefficient between $X(s)$ and $X(s + t)$ equals

$$\rho(s, s + t) = e^{-\alpha t}\sqrt{\frac{1 - e^{-2\alpha s}}{1 - e^{-2\alpha(s+t)}}} \to e^{-\alpha t}, \qquad s \to \infty.$$

Example 9.10 Geometric Brownian motion $Y_t = e^{\mu t + \sigma W_t}$. Due to the Ito formula

$$dY_t = (\mu + \frac{1}{2}\sigma^2)Y_t dt + \sigma Y_t dW_t,$$

so that $\mu(t, x) = (\mu + \frac{1}{2}\sigma^2)x$ and $\sigma^2(t, x) = \sigma^2 x^2$. The process Y_t is a martingale iff $\mu = -\frac{1}{2}\sigma^2$.

Example 9.11 Take $f(t, x) = x^2$. Then the Ito formula gives

$$dX_t^2 = 2X_t dX_t + \sigma(t, X_t)^2 dt.$$

In particular, $dW_t^2 = 2W_t dW_t + dt$ so that $\int_0^t W_t dW_t = (W_t^2 - t)/2$.

Example 9.12 Product rule. If

$$dX_t = \mu_1(t, X_t)dt + \sigma_1(t, X_t)dW_t, \quad dY_t = \mu_2(t, Y_t)dt + \sigma_2(t, Y_t)dW_t,$$

then

$$d(X_t Y_t) = X_t dY_t + Y_t dX_t + \sigma_1(t, X_t)\sigma_2(t, Y_t)dt.$$

This follows from $2XY = (X + Y)^2 - X^2 - Y^2$ and

$$dX_t^2 = 2X_t dX_t + \sigma_1(t, X_t)^2 dt, \quad dY_t^2 = 2Y_t dY_t + \sigma_2(t, X_t)^2 dt,$$
$$d(X_t + Y_t)^2 = 2(X_t + Y_t)(dX_t + dY_t) + (\sigma_1(t, X_t) + \sigma_2(t, X_t))^2 dt.$$

9.5 The Black-Scholes formula

Lemma 9.13 *Let $(W_t, 0 \le t \le T)$ be the standard Wiener process on $(\Omega, \mathcal{F}, \mathbb{P})$ and let $\nu \in \mathbb{R}$. Define another measure by $\mathbb{Q}(A) = \mathbb{E}(e^{\nu W_T - \nu^2 T/2} 1_{\{A\}})$. Then $\mathbb{Q}$ is a probability measure and $\tilde{W}_t = W_t - \nu t$, regarded as a process on the probability space $(\Omega, \mathcal{F}, \mathbb{Q})$, is the standard Wiener process.*

Proof. By definition, $\mathbb{Q}(\Omega) = e^{-\nu^2 T/2} \mathbb{E}(e^{\nu W_T}) = 1$. For the finite-dimensional distributions let $0 = t_0 < t_1 < \ldots < t_n = T$ and $x_0, x_1, \ldots, x_n \in \mathbb{R}$. Writing $\{W(t_i) \in dx_i\}$ for the event $\{x_i < W(t_i) \le x_i + dx_i\}$ we have that

$$\mathbb{Q}(W(t_1) \in dx_1, \ldots, W(t_n) \in dx_n) = \mathbb{E}\big(e^{\nu W_T - \nu^2 T/2} 1_{\{W(t_1) \in dx_1, \ldots, W(t_n) \in dx_n\}}\big)$$

$$= e^{\nu x_n - \nu^2 T/2} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi(t_i - t_{i-1})}} \exp\Big(-\frac{(x_i - x_{i-1})^2}{2(t_i - t_{i-1})}\Big) dx_i$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi(t_i - t_{i-1})}} \exp\Big(-\frac{(x_i - x_{i-1} - \nu(t_i - t_{i-1}))^2}{2(t_i - t_{i-1})}\Big) dx_i.$$

Black-Scholes model. Writing B_t for the cost of one unit of a risk-free bond (so that $B_0 = 1$) at time t we have that

$$dB_t = rB_t dt \text{ or } B_t = e^{rt}.$$

The price (per unit) S_t of a stock at time t satisfies the stochastic differential equation

$$dS_t = S_t(\mu dt + \sigma dW_t) \text{ with solution } S_t = \exp\{(\mu - \sigma^2/2)t + \sigma W_t\}.$$

This is a geometric Brownian motion, and parameter σ is called volatility of the price process.

European call option. The buyer of the option may purchase one unit of stock at the exercise date T (fixed time) and the strike price K:

- if $S_T > K$, an immediate profit will be $S_T - K$,

- if $S_T \le K$, the call option will not be exercised.

The value of the option at time $t < T$ is $V_t = e^{-r(T-t)}(S_T - K)^+$, where S_T is not known.

Theorem 9.14 *Black-Scholes formula. Let $t < T$. The value of the European call option at time t is*

$$V_t = S_t \Phi(d_1(t, S_t)) - K e^{-r(T-t)} \Phi(d_2(t, S_t)),$$

where $\Phi(x)$ is the standard normal distribution function and

$$d_1(t, x) = \frac{\log(x/K) + (r + \sigma^2/2)(T-t)}{\sigma\sqrt{T-t}}, \qquad d_2(t, x) = \frac{\log(x/K) + (r - \sigma^2/2)(T-t)}{\sigma\sqrt{T-t}}.$$

Definition 9.15 Let $\mathcal{F}_t$ be the σ-algebra generated by $(S_u, 0 \le u \le t)$. A portfolio is a pair (α_t, β_t) of $\mathcal{F}_t$-adapted processes. The value of the portfolio is $V_t(\alpha, \beta) = \alpha_t S_t + \beta_t B_t$. The portfolio is called self-financing if

$$dV_t(\alpha, \beta) = \alpha_t dS_t + \beta_t dB_t.$$

We say that a self-financing portfolio (α_t, β_t) replicates the given European call if $V_T(\alpha, \beta) = (S_T - K)^+$ almost surely.

Proof. We are going to apply Lemma 9.13 with $\nu = \frac{r-\mu}{\sigma}$. Note also that under $\mathbb{Q}$ the process $e^{-rt} S_t = \exp\{(\nu\sigma - \sigma^2/2)t + \sigma W_t\} = e^{\sigma\tilde{W}_t - \sigma^2 t/2}$ is a martingale.

Take without a proof that there exists a self-financing portfolio (α_t, β_t) replicating the European call option in question. If the market contains no arbitrage opportunities, we have $V_t = V_t(\alpha, \beta)$ and therefore

$$d(e^{-rt} V_t) = e^{-rt} dV_t - re^{-rt} V_t dt = e^{-rt} \alpha_t(dS_t - rS_t dt) + e^{-rt} \beta_t(dB_t - rB_t dt)$$

$$= \alpha_t e^{-rt} S_t((\mu - r)dt + \sigma dW_t) = \alpha_t e^{-rt} S_t \sigma d\tilde{W}_t.$$

This defines a martingale under $\mathbb{Q}$:

$$e^{-rt}V_t = V_0 + \int_0^t \alpha_u e^{-ru} S_u \sigma d\tilde{W}_u.$$

Thus

$$V_t = e^{rt}\mathbb{E}_{\mathbb{Q}}(e^{-rT}V_T|\mathcal{F}_t) = e^{-r(T-t)}\mathbb{E}_{\mathbb{Q}}((S_T - K)^+|\mathcal{F}_t) = e^{-r(T-t)}\mathbb{E}_{\mathbb{Q}}((ae^Z - K)^+|\mathcal{F}_t),$$

where $a = S_t$ with

$$Z = \exp\{(\mu - \sigma^2/2)(T - t) + \sigma(W_T - W_t)\} = \exp\{(r - \sigma^2/2)(T - t) + \sigma(\tilde{W}_T - \tilde{W}_t)\}.$$

Since $(Z|\mathcal{F}_t)_{\mathbb{Q}} \sim \mathrm{N}(\gamma, \tau^2)$ with $\gamma = (r - \sigma^2/2)(T - t)$ and $\tau^2 = (T - t)\sigma^2$. It remains to observe that for any constant a and $Z \sim \mathrm{N}(\gamma, \tau^2)$

$$\mathbb{E}((ae^Z - K)^+) = ae^{\gamma + \tau^2/2}\Phi\Big(\frac{\log(a/K) + \gamma}{\tau} + \tau\Big) - K\Phi\Big(\frac{\log(a/K) + \gamma}{\tau}\Big).$$