

# 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. In this text the formula labels (\*), (\*\*) operate locally.

Last updated May 23, 2018.

## Contents

<b>Abstract</b>	<b>1</b>
<b>1 Random events and random variables</b>	<b>2</b>
1.1 Probability space . . . . .	2
1.2 Conditional probability and independence . . . . .	3
1.3 Random variables . . . . .	4
1.4 Random vectors . . . . .	5
1.5 Filtration . . . . .	6
<b>2 Expectation and conditional expectation</b>	<b>7</b>
2.1 Expectation . . . . .	7
2.2 Conditional expectation and prediction . . . . .	8
2.3 Multinomial distribution . . . . .	9
2.4 Multivariate normal distribution . . . . .	10
2.5 Sampling from a distribution . . . . .	11
2.6 Probability generating, moment generating, and characteristic functions . . . . .	11
<b>3 Convergence of random variables</b>	<b>12</b>
3.1 Borel-Cantelli lemmas . . . . .	12
3.2 Inequalities . . . . .	13
3.3 Modes of convergence . . . . .	14
3.4 Continuity of expectation . . . . .	16
<b>4 Limit theorems for the sums of IID random variables</b>	<b>18</b>
4.1 Weak law of large numbers . . . . .	18
4.2 Central limit theorem . . . . .	19
4.3 Strong LLN . . . . .	19
4.4 Large deviations . . . . .	20
<b>5 Markov chains</b>	<b>23</b>
5.1 Simple random walks . . . . .	23
5.2 Markov chains . . . . .	25
5.3 Stationary distributions . . . . .	26
5.4 Reversibility . . . . .	27
5.5 Branching process . . . . .	28
5.6 Poisson process and continuous-time Markov chains . . . . .	29
5.7 The generator of a continuous-time Markov chain . . . . .	29

<b>6</b>	<b>Stationary processes</b>	<b>30</b>
6.1	Weakly and strongly stationary processes . . . . .	30
6.2	Linear prediction . . . . .	31
6.3	Linear combination of sinusoids . . . . .	32
6.4	Spectral representation . . . . .	33
6.5	Stochastic integral . . . . .	35
6.6	Ergodic theorem for the weakly stationary processes . . . . .	37
6.7	Ergodic theorem for the strongly stationary processes . . . . .	38
6.8	Gaussian processes . . . . .	39
<b>7</b>	<b>Renewal theory and Queues</b>	<b>40</b>
7.1	Renewal function . . . . .	40
7.2	Renewal equation and excess life . . . . .	41
7.3	LLN and CLT for the renewal process . . . . .	42
7.4	Stopping times and Wald's equation . . . . .	43
7.5	Renewal theorems and stationarity . . . . .	44
7.6	Renewal-reward processes . . . . .	46
7.7	Regeneration technique for queues . . . . .	46
7.8	M/M/1 queues . . . . .	48
7.9	M/G/1 queues . . . . .	49
7.10	G/M/1 queues . . . . .	52
7.11	G/G/1 queues . . . . .	53
<b>8</b>	<b>Martingales</b>	<b>55</b>
8.1	Definitions and examples . . . . .	55
8.2	Convergence in $L^2$ . . . . .	56
8.3	Doob decomposition . . . . .	58
8.4	Hoeffding inequality . . . . .	58
8.5	Convergence in $L^1$ . . . . .	60
8.6	Doob's martingale . . . . .	62
8.7	Bounded stopping times. Optional sampling theorem . . . . .	63
8.8	Unbounded stopping times . . . . .	64
8.9	Maximal inequality and convergence in $L^r$ . . . . .	66
8.10	Backward martingales. Martingale proof of the strong LLN . . . . .	68
<b>9</b>	<b>Diffusion processes</b>	<b>69</b>
9.1	The Wiener process . . . . .	69
9.2	Properties of the Wiener process . . . . .	70
9.3	Examples of diffusion processes . . . . .	72
9.4	The Ito formula . . . . .	72
9.5	The Black-Scholes formula . . . . .	74

# 1 Random events and random variables

## 1.1 Probability space

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

- the *sample space*  $\Omega$  is the set of all possible outcomes of the experiment,
- the  $\sigma$ -*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  $A_i \in \mathcal{F}$ ,  $0 = 1, 2, \dots$ , 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  $P$  is a function on  $\mathcal{F}$  satisfying three probability axioms

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

De Morgan laws

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

Properties derived from the axioms

$$\begin{aligned} P(\emptyset) &= 0, \\ P(A^c) &= 1 - P(A), \\ P(A \cup B) &= P(A) + P(B) - P(A \cap B), \\ P(A \Delta B) &= P(A) + P(B) - 2P(A \cap B), \end{aligned}$$

where  $A \Delta B = (A \cup B) \setminus (A \cap B)$  is a symmetric difference of two events.

Inclusion-exclusion rule

$$\begin{aligned} P(A_1 \cup \dots \cup A_n) &= \sum_i P(A_i) - \sum_{i < j} P(A_i \cap A_j) + \sum_{i < j < k} P(A_i \cap A_j \cap A_k) - \dots \\ &\quad + (-1)^{n+1} P(A_1 \cap \dots \cap A_n). \end{aligned}$$

Continuity of the probability measure

if  $A_1 \subset A_2 \subset \dots$  and  $A = \cup_{i=1}^{\infty} A_i = \lim_{i \rightarrow \infty} A_i$ , then  $P(A) = \lim_{i \rightarrow \infty} P(A_i)$ ,

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

## 1.2 Conditional probability and independence

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

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

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

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

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

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

**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,  $P(A|C) = P(A)$  and  $P(B|C) = P(B)$  but  $P(A \cap B|C) = 0$ .

### 1.3 Random variables

A real random variable is a measurable function  $X : \Omega \rightarrow \mathbb{R}$  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  $P_X(B) = P(X \in B)$  defines a new probability space  $(\mathbb{R}, \mathcal{B}, P_X)$ , where  $\mathcal{B} = \sigma(\text{all open intervals})$  is the Borel sigma-algebra.

**Definition 1.3** Distribution function (cumulative distribution function)

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

In terms of the distribution function we get

$$\begin{aligned} P(a < X \leq b) &= F(b) - F(a), \\ P(X < x) &= F(x-), \\ P(X = x) &= F(x) - F(x-). \end{aligned}$$

Any monotone right-continuous function with

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

can be a distribution function.

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

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

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

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  $p = P(A)$  has a Bernoulli distribution

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

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

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

**Example 1.6** Consider a sequence of Bernoulli trials, that is independent experiments with two possible outcomes: success with probability  $p$  and failure with probability  $q = 1 - p$ . Let  $X$  be the number of failures until the first success. Then  $X$  has a geometric distribution  $\text{Geom}(p)$ :

$$P(X = k) = q^k p, \quad k = 0, 1, \dots$$

The number of trials  $Y = X + 1$  until the first success has a shifted geometric distribution  $\text{Geom}_1(p)$ :

$$P(Y = k) = q^{k-1} p, \quad k = 1, 2, \dots$$

**Example 1.7** (Cantor distribution) Consider  $(\Omega, \mathcal{F}, P)$  with  $\Omega = [0, 1]$ ,  $\mathcal{F} = \mathcal{B}_{[0,1]}$ , and

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

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 1.8** The joint distribution of a random vector  $\mathbf{X} = (X_1, \dots, X_n)$  is the function

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

Marginal distributions

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

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

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

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

**Definition 1.9** Random variables  $(X_1, \dots, X_n)$  are called independent if for any  $(x_1, \dots, x_n)$

$$P(X_1 \leq x_1, \dots, X_n \leq x_n) = P(X_1 \leq x_1) \dots P(X_n \leq x_n).$$

In the jointly continuous case this equivalent to

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

**Example 1.10** In general, the joint distribution can not be recovered from the marginal distributions. If

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

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

**Example 1.11** Consider

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

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.  $F(x, y) \rightarrow 0$  as  $x \rightarrow -\infty$  and  $y \rightarrow -\infty$ ,
3.  $F(x, y) \rightarrow 1$  as  $x \rightarrow \infty$  and  $y \rightarrow \infty$ .

Observe that

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

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

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

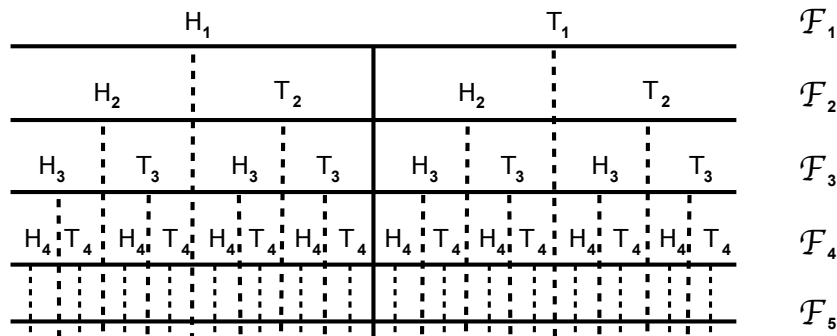


Figure 1: Filtration for four consecutive coin tossings.

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

$$\int_{-\infty}^x \int_{-\infty}^y f(u, v) du dv = \int_0^x \left( \int_u^y e^{-v} dv \right) du = 1 - e^{-x} - xe^{-y},$$

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

$$\int_{-\infty}^x \int_{-\infty}^y f(u, v) du dv = \int_0^y \left( \int_u^x e^{-v} dv \right) du = 1 - e^{-y} - ye^{-y}.$$

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

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

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

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

## 1.5 Filtration

**Definition 1.12** A sequence of sigma-fields  $\{\mathcal{F}_n\}_{n=1}^{\infty}$  such that

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

is called a filtration.

**Example 1.13** To illustrate the definition of filtration, we 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 embedded 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

$$E(S_4 | S_3) = S_3 + 0.5$$

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.

## 2 Expectation and conditional expectation

### 2.1 Expectation

The expected value of  $X$  is

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

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  $A_1, \dots, A_n$  of  $\Omega$ . In this case the meaning of the expectation is obvious

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

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  $E(X) = \lim_{n \rightarrow \infty} E(X_n)$ .

Any r.v.  $X$  can be written as a difference of two non-negative r.v.  $X_1 = X \wedge 0$  and  $X_2 = -X \wedge 0$ . If at least one of  $E(X_1)$  and  $E(X_2)$  is finite, then  $E(X) = E(X_1) - E(X_2)$ , otherwise  $E(X)$  does not exist.

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

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

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

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

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

In general

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

**Example 2.2** Turn to the example of  $(\Omega, \mathcal{F}, P)$  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  $\omega \in [(k-1)3^{-n}, k3^{-n})$ ,  $k = 1, \dots, 3^n$  and  $X_n(1) = 1$ .

$$X_1(\omega) = 0, \quad E(X_1) = 0,$$

$$X_2(\omega) = (1/3) \cdot 1_{\{\omega \in [1/3, 2/3)\}} + (2/3) \cdot 1_{\{\omega \in [2/3, 1]\}}, \quad E(X_2) = (2/3) \cdot (1/2) = 1/3,$$

$$E(X_3) = (2/9) \cdot (1/4) + (2/3) \cdot (1/4) + (8/9) \cdot (1/4) = 4/9,$$

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

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

$$(E(XY))^2 \leq E(X^2)E(Y^2)$$

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

**Definition 2.4** Variance, standard deviation, covariance and correlation

$$\begin{aligned} \text{Var}(X) &= E(X - EX)^2 = E(X^2) - (EX)^2, \quad \sigma_X = \sqrt{\text{Var}(X)}, \\ \text{Cov}(X, Y) &= E(X - EX)(Y - EY) = E(XY) - (EX)(EY), \\ \rho(X, Y) &= \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}. \end{aligned}$$

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

$$\mathbf{V} = \mathbb{E}(\mathbf{X} - \boldsymbol{\mu})^t (\mathbf{X} - \boldsymbol{\mu}) = \|\text{Cov}(X_i, X_j)\|$$

is symmetric and nonnegative-definite. For any vector  $\mathbf{a} = (a_1, \dots, a_n)$  the r.v.  $a_1X_1 + \dots + a_nX_n$  has mean  $\mathbf{a}\boldsymbol{\mu}^t$  and variance

$$\text{Var}(a_1X_1 + \dots + a_nX_n) = \mathbb{E}(\mathbf{a}\mathbf{X}^t - \mathbf{a}\boldsymbol{\mu}^t)(\mathbf{X}\mathbf{a}^t - \boldsymbol{\mu}\mathbf{a}^t) = \mathbf{a}\mathbf{V}\mathbf{a}^t.$$

If  $(X_1, \dots, X_n)$  are independent, then they are uncorrelated:  $\text{Cov}(X_i, X_j) = 0$ .

## 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

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

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

$$\psi(x) = \int_{-\infty}^{\infty} yf_2(y|x)dy.$$

Similarly, one can define a conditional expectation  $\mathbb{E}(Y|X, Z)$  of a random variable  $Y$  conditioned on a random vector  $(X, Z)$ .

### 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)  $\mathbb{E}(Y1_G) = \mathbb{E}(\psi(X)1_G)$  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:  $\text{Var}(Y) = \text{Var}(\mathbb{E}(Y|X)) + \mathbb{E}(\text{Var}(Y|X))$ .

PROOF of (ii) in the discrete case:

$$\begin{aligned} \mathbb{E}(Yg(X)|X) &= \sum_{x,y} yg(x)\mathbb{P}(Y = y, X = x) = \sum_{x,y} g(x)\mathbb{P}(X = x)y\mathbb{P}(Y = y|X = x) \\ &= \sum_x g(x)\psi(x)\mathbb{P}(X = x) = g(X)\mathbb{E}(Y|X). \end{aligned}$$

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

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

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

PROOF of (viii). Consider the probability space  $(\Omega, \mathcal{G}, P)$  and define a finite signed measure  $P_1(G) = E(Y1_G) = \int_G Y(\omega)P(d\omega)$  which is absolutely continuous with respect to  $P$ . Thus  $P_1(G) = \int_G Z(\omega)P(d\omega)$  with  $Z = \partial P_1/\partial P$  being the Radon-Nikodym derivative.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

we have

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

### 2.3 Multinomial distribution

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

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

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

$$(X_1 + X_2, X_3, \dots, X_r) \sim Mn(n, p_1 + p_2, p_3, \dots, p_r).$$

Conditionally on  $X_1$

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

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

$$\begin{aligned} E(X_i X_j) &= E(E(X_i X_j | X_j)) \\ &= E(X_j E(X_i | X_j)) = E(n X_j - X_j^2) \frac{p_i}{1-p_j} \\ &= (n^2 p_j - n p_j (1-p_j) + n^2 p_j^2) \frac{p_i}{1-p_j} = n(n-1)p_i p_j \end{aligned}$$

and  $\text{Cov}(X_i, X_j) = -n p_i p_j$  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

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

Exercise: check the total variance formula for this example.

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

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

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

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

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  $\bar{X} \stackrel{d}{=} X$ . 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  $U_1, U_2, \dots$  which we consider as IID r.v. with  $U_{[0,1]}$  distribution.

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

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

**Example 2.10** Examples of the inverse transform sampling.

- (i) Bernoulli distribution  $X = 1_{\{U \leq p\}}$ ,
- (ii) Binomial sampling:  $S_n = X_1 + \dots + X_n$ ,  $X_k = 1_{\{U_k \leq p\}}$ ,
- (iii) Exponential distribution  $X = -\log(U)/\lambda$ ,
- (iv) Gamma sampling:  $S_n = X_1 + \dots + X_n$ ,  $X_k = -\log(U_k)/\lambda$ .

**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  $U \sim U_{[0,1]}$ . Then

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

## 2.6 Probability generating, moment generating, and characteristic functions

**Definition 2.12** If  $X$  takes values  $k = 0, 1, 2, \dots$  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 probability generating functions:

- (i)  $p_0 = G(0)$ ,  $p_k = \frac{1}{k!} \frac{d^k G(s)}{ds^k} \Big|_{s=0}$ ,
- (ii)  $\mathbb{E}(X) = G'(1)$ ,  $\mathbb{E}(X(X-1)) = G''(1)$ .
- (iii) 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  $G(s) = \frac{p}{1-(1-p)s}$ . Shifted geometric distribution  $G(s) = \frac{ps}{1-(1-p)s}$ .
- (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( $\alpha, \lambda$ ) 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  $\lambda$  has a Gamma( $k, \lambda$ ) 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  $\mathbf{X} = (X_1, \dots, X_n)$  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  $\mathbf{X} = (X_1, \dots, X_n)$  with a multivariate normal distribution any linear combination  $\mathbf{a}\mathbf{X}^t = a_1X_1 + \dots + a_nX_n$  is normally distributed since

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

### 3 Convergence of random variables

#### 3.1 Borel-Cantelli lemmas

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

$$\begin{aligned} \sup_n A_n &= \bigcup_n A_n, & \inf_n A_n &= \bigcap_n A_n, \\ \limsup_{n \rightarrow \infty} A_n &= \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} A_m, & \liminf_{n \rightarrow \infty} A_n &= \bigcup_{n=1}^{\infty} \bigcap_{m=n}^{\infty} A_m. \end{aligned}$$

These relations are explained in terms of the indicator random variables. For example,

$$\sup_n 1_{\{\omega \in A_n\}} = \begin{cases} 1, & \text{if } \omega \in A_i \text{ for some } i \\ 0, & \text{otherwise} \end{cases} = 1_{\{\omega \in \bigcup_n A_n\}}.$$

We will write  $\{A_n \text{ i.o.}\} = \{\text{infinitely many of events } A_1, A_2, \dots \text{ occur}\}$ . If  $A = \{A_n \text{ i.o.}\}$ , then

$$\begin{aligned} \limsup_{n \rightarrow \infty} A_n &= \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} A_m = \{\forall n \exists m \geq n \text{ such that } A_m \text{ occurs}\} = A, \\ A^c &= \liminf_{n \rightarrow \infty} A_n^c = \bigcup_{n=1}^{\infty} \bigcap_{m=n}^{\infty} A_m^c = \{\exists n \text{ such that } A_m^c \text{ occur } \forall m \geq n\} = \{\text{events } A_n \text{ occur finitely often}\}. \end{aligned}$$

**Theorem 3.1** *Borel-Cantelli lemmas.*

1. If  $\sum_{n=1}^{\infty} P(A_n) < \infty$ , then  $P(A_n \text{ i.o.}) = 0$ ,
2. If  $\sum_{n=1}^{\infty} P(A_n) = \infty$  and events  $A_1, A_2, \dots$  are independent, then  $P(A_n \text{ i.o.}) = 1$ .

PROOF. (1) Put  $A = \{A_n \text{ i.o.}\}$ . We have  $A \subset \bigcup_{m \geq n} A_m$ , and so  $P(A) \leq \sum_{m \geq n} P(A_m) \rightarrow 0$  as  $n \rightarrow \infty$ .

(2) By independence

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

It follows  $P(A^c) \leq \sum_n P(\bigcap_{m \geq n} A_m^c) = 0$  which gives  $P(A) = 1$ .

If events  $A_1, A_2, \dots$  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  $X_1, X_2, \dots$  be a sequence of random variables defined on the same probability space and  $\mathcal{H}_n = \sigma(X_{n+1}, X_{n+2}, \dots)$ . Then  $\mathcal{H}_n \supset \mathcal{H}_{n+1} \supset \dots$ , and we define the tail  $\sigma$ -algebra as  $\mathcal{T} = \bigcap_n \mathcal{H}_n$ .

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

**Theorem 3.3** *Kolmogorov zero-one law.* Let  $X_1, X_2, \dots$  be independent random variables defined on the same probability space. For all events  $H \in \mathcal{T}$  from the tail  $\sigma$ -algebra we have either  $P(H) = 0$  or  $P(H) = 1$ .

PROOF. A standard result of measure theory asserts that for any event  $H \in \mathcal{H}_1 = \sigma(X_1, X_2, \dots)$  there exists a sequence of events  $C_n = \sigma(X_1, \dots, X_n)$  such that  $P(H \Delta C_n) \rightarrow 0$  as  $n \rightarrow \infty$ . If  $H \in \mathcal{T}$ , then by independence

$$P(H \cap C_n) = P(H)P(C_n) \rightarrow P(H)^2$$

implying  $P(H) = P(H)^2$ .

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

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

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

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

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

Visiting the origin infinitely often is a tail event with respect to the sequence  $(S_n)$ , but  $S_n$  are not independent and therefore the Kolmogorov zero-one law is not directly applicable here.

On the other hand, whether  $A$  occurs or not depends on the value of  $X_1$ . Indeed, if  $X_2 = X_4 = \dots = 1$  and  $X_3 = X_5 = \dots = -1$ , then  $A$  occurs if  $X_1 = -1$  and does not occur if  $X_1 = 1$ . Thus  $A$  is not a tail event with respect to the sequence  $(X_n)$ . However, if  $(X_n)$  are independent and identically distributed, then by the so-called Hewitt-Savage zero-one law, the event  $A$ , being invariant under finite permutations, has probability either 0 or 1.

## 3.2 Inequalities

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

$$J(E(X)) \leq E(J(X)).$$

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

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

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

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

PROOF. Using truncation,

$$E|X| \geq E(|X|1_{\{|X| \geq a\}}) \geq aE(1_{\{|X| \geq a\}}) = aP(|X| \geq a).$$

**Chebyshev inequality.** Given a random variable  $X$  with mean  $\mu$  and variance  $\sigma^2$  for any  $\epsilon > 0$  we have

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

PROOF. By Markov inequality,

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

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

$$(E(XY))^2 \leq E(X^2)E(Y^2).$$

**Exercise 3.6** For a random variable  $X$  define its cumulant generating function by  $\Lambda(t) = \log M(t)$ , where  $M(t) = E(e^{tX})$  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

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

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

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

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

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

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

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

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

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

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

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

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

### 3.3 Modes of convergence

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

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

if we generalise the definition of random variable by allowing taking the value  $\infty$  or  $-\infty$ .

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

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

**Definition 3.8** Let  $X, X_1, X_2, \dots$  be random variables on some probability space  $(\Omega, \mathcal{F}, P)$ . Define four modes of convergence of random variables

- (i) almost sure convergence  $X_n \xrightarrow{\text{a.s.}} X$ , if  $P(\omega : \lim X_n = X) = 1$ ,
- (ii) convergence in  $r$ -th mean  $X_n \xrightarrow{L^r} X$  for a given  $r \geq 1$ , if  $E|X_n^r| < \infty$  for all  $n$  and

$$E(|X_n - X|^r) \rightarrow 0,$$

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

$$P(X_n \leq x) \rightarrow P(X \leq x) \text{ for all } x \text{ such that } P(X = x) = 0,$$

or equivalently,  $Ef(X_n) \rightarrow Ef(X)$  for all bounded uniformly continuous  $f : R \rightarrow R$ .

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

$$X_n \xrightarrow{L^q} X \Rightarrow_{(i)} X_n \xrightarrow{\text{a.s.}} X \Rightarrow_{(ii)} X_n \xrightarrow{P} X \Rightarrow_{(iv)} X_n \xrightarrow{d} X.$$

$$X_n \xrightarrow{L^q} X \Rightarrow_{(i)} X_n \xrightarrow{L^r} X \Rightarrow_{(iii)}$$

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  $P(\cup_{n \geq m} \{|X_n - X| > \epsilon\}) \rightarrow 0$ ,  $m \rightarrow \infty$  for any  $\epsilon > 0$ . To prove (iv) use

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

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

**Theorem 3.10** Reverse implications:

- (i)  $X_n \xrightarrow{d} c \Rightarrow X_n \xrightarrow{P} c$  for a constant limit,
- (ii)  $X_n \xrightarrow{P} X \Rightarrow X_n \xrightarrow{L^r} X$ , if  $P(|X_n| \leq a) = 1$  for all  $n$  and some positive constant  $a$ ,
- (iii)  $X_n \xrightarrow{P} X \Rightarrow X_n \xrightarrow{\text{a.s.}} X$  if  $P(|X_n - X| > \epsilon) \rightarrow 0$  so fast that

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

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

PROOF. To prove (i) use

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

To prove (ii) use  $P(|X| \leq a) = 1$  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  $B_m = \{|X_n - X| > m^{-1} \text{ i.o.}\}$ . Due to the Borel-Cantelli lemma, condition (\*) implies  $P(B_m) = 0$  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  $1 + 1/2 + \dots + 1/n$ .

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

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

- The random variables  $n \cdot 1_{B_n}$  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  $2 \cdot 1_{B_2}$ .

**Exercise 3.13** Suppose  $X_n \xrightarrow{L^r} X$ , where  $r \geq 1$ . Show that

$$\mathbb{E}(|X_n|^r) \rightarrow \mathbb{E}(|X|^r), n \rightarrow \infty.$$

Hint: use Minkowski inequality two times as you need two estimate  $\overline{\lim} \mathbb{E}(|X_n|^r)$  from above and  $\underline{\lim} \mathbb{E}(|X_n|^r)$  from below.

SOLUTION. By Minkowski inequality,

$$(\mathbb{E}|X_n^r|)^{1/r} \leq (\mathbb{E}|X^r|)^{1/r} + (\mathbb{E}|X_n - X|^r)^{1/r},$$

implying  $\overline{\lim} \mathbb{E}(|X_n|^r) \leq \mathbb{E}(|X|^r)$ . On the other hand,  $\underline{\lim} \mathbb{E}(|X_n|^r) \geq \mathbb{E}(|X|^r)$ , since

$$(\mathbb{E}|X^r|)^{1/r} \leq (\mathbb{E}|X_n^r|)^{1/r} + (\mathbb{E}|X_n - X|^r)^{1/r}.$$

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

**Exercise 3.15** Suppose  $X_n \xrightarrow{L^2} X$ . Show that

$$\text{Var}(X_n) \rightarrow \text{Var}(X), n \rightarrow \infty.$$

Hint: use the previous two exercises.

### 3.4 Continuity of expectation

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

PROOF. Let  $\epsilon > 0$  and use

$$\begin{aligned} |\mathbb{E}(X_n) - \mathbb{E}(X)| &\leq \mathbb{E}|X_n - X| = \mathbb{E}(|X_n - X| \cdot 1_{\{|X_n - X| \leq \epsilon\}}) + \mathbb{E}(|X_n - X| \cdot 1_{\{|X_n - X| > \epsilon\}}) \\ &\leq \epsilon + MP(|X_n - X| > \epsilon). \end{aligned}$$

**Lemma 3.17** *Fatou 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

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

The convergence  $\mathbb{E}(Y \wedge M) \rightarrow \mathbb{E}(Y)$  as  $M \rightarrow \infty$  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  $\omega$ , then, clearly, for all  $\omega$ , 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 lemma.



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

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

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

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

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

$$n(1 - F(n)) \leq E(X1_{X>n}) \rightarrow 0.$$

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

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

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

**Theorem 3.21** *Dominated convergence.* If  $X_n \xrightarrow{a.s.} X$  and almost surely  $|X_n| \leq Y$  and  $E(Y) < \infty$ , then  $E|X| < \infty$  and  $E(X_n) \rightarrow E(X)$ .

PROOF. Apply Fatou lemma twice.

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

$$\sup_n E(|X_n|; |X_n| > a) \rightarrow 0, \quad a \rightarrow \infty.$$

**Exercise 3.23** Put  $Z = \sup_n |X_n|$ . If  $E(Z) < \infty$ , then the sequence  $X_n$  is uniformly integrable. Hint:  $E(|X_n|1_A) \leq E(Z1_A)$  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)  $\sup_n E|X_n| < \infty$ ,
- (ii) for all  $\epsilon > 0$ , there is  $\delta > 0$  such that, for any event  $A$  such that  $P(A) < \delta$ ,

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

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

$$\sup_n E|X_n| = \sup_n \left( E(|X_n|; |X_n| \leq a) + E(|X_n|; |X_n| > a) \right) \leq a + \sup_n E(|X_n|; |X_n| > a).$$

Thus  $\sup_n E|X_n| < \infty$ .

Step 2. Assume that  $(X_n)$  is uniformly integrable. Pick a positive  $a$  such that

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

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

$$E(|X_n|1_A) = E(|X_n|1_{\{A \cap B_n\}}) + E(|X_n|1_{\{A \cap B_n^c\}}) \leq E(|X_n|1_{\{B_n\}}) + aP(A) < \epsilon.$$

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

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

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

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

**Theorem 3.25** Let  $X_n \xrightarrow{P} X$ . The following three statements are equivalent to one another.

- (a) The sequence  $X_n$  is uniformly integrable.
- (b)  $E|X_n| < \infty$  for all  $n$ ,  $E|X| < \infty$ , and  $X_n \xrightarrow{L^1} X$ .
- (c)  $E|X_n| < \infty$  for all  $n$ ,  $E|X| < \infty$ , and  $E|X_n| \rightarrow E|X|$ .

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

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

$$E|X^r| = E\left(\liminf_{n' \rightarrow \infty} |X_{n'}^r|\right) \leq \liminf_{n' \rightarrow \infty} E|X_{n'}^r| \leq \sup_n E|X_n^r| < \infty.$$

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

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

By the Minkowski inequality

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

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

## 4 Limit theorems for the sums of IID random variables

### 4.1 Weak law of large numbers

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

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

This is equivalent to the weak convergence  $F_n \xrightarrow{d} F$  of distribution functions when  $F_n(x) \rightarrow F(x)$  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  $\phi_n(t) \rightarrow \phi(t)$  for all  $t$ ,
- (iii) conversely, if  $\phi(t) = \lim \phi_n(t)$  exists and is continuous at  $t = 0$ , then  $\phi$  is cf of some  $F$ , and  $F_n \xrightarrow{d} F$ .

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

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

PROOF. Let  $F_n$  and  $\phi_n$  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  $\phi_n(t) \rightarrow e^{it\mu}$  which is obtained using a Taylor expansion

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

**Example 4.4** Statistical application: the sample mean is a consistent estimate of the population mean. Counterexample: if  $X_1, X_2, \dots$  are iid with the Cauchy distribution, then  $S_n/n \stackrel{d}{=} X_1$  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  $X_1, X_2, \dots$  are iid with finite mean  $\mu$  and positive finite variance  $\sigma^2$ , then for any  $x$*

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

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

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

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

$$\mathbf{V} = \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & -p_1p_3 & \dots & -p_1p_r \\ -p_2p_1 & p_2(1-p_2) & -p_2p_3 & \dots & -p_2p_r \\ -p_3p_1 & -p_3p_2 & p_3(1-p_3) & \dots & -p_3p_r \\ \dots & \dots & \dots & \dots & \dots \\ -p_r p_1 & -p_r p_2 & -p_r p_3 & \dots & p_r(1-p_r) \end{pmatrix}.$$

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

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

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

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

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

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

## 4.3 Strong LLN

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

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

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

$$\mathbb{E}\left(\left(\frac{X_1 + \dots + X_n}{n} - \mu\right)^2\right) = \text{Var}\left(\frac{X_1 + \dots + X_n}{n}\right) = \frac{n\sigma^2}{n^2} \rightarrow 0.$$

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

$$\frac{X_1 + \dots + X_n}{n} \xrightarrow{\text{a.s.}} \mu$$

for some constant  $\mu$  iff  $E|X_1| < \infty$ . In this case  $\mu = EX_1$  and  $\frac{X_1 + \dots + X_n}{n} \xrightarrow{L^1} \mu$ .

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

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

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

and

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

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

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

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  $X_1, X_2, \dots$  be iid random variables with mean  $\mu$  and variance  $\sigma^2$ . 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

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

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  $S_n = X_1 + \dots + X_n$ , where  $X_1, X_2, \dots$  are iid random variables with mean  $\mu$ , and suppose that their moment generating function  $M(t) = E(e^{tX})$  is finite in some neighborhood of the origin  $t = 0$ .

(i) If  $a > \mu$  and the following regularity condition hold

$$\Lambda^*(a) = a\tau - \Lambda(\tau), \quad \text{for some } \tau > 0, \quad M(\tau) < \infty,$$

then  $\Lambda^*(a) > 0$  and

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

(ii) If  $a < \mu$  and the following regularity condition hold

$$\Lambda^*(a) = a\tau - \Lambda(\tau), \quad \text{for some } \tau < 0, \quad M(\tau) < \infty,$$

then  $\Lambda^*(a) > 0$  and

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

PROOF. Without loss of generality assume  $\mu = 0$  and take  $a > 0$ . The upper bound for (i) is obtained in the form

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

Indeed, for  $t > 0$ ,

$$e^{tS_n} > e^{nat} \mathbf{1}_{\{S_n > na\}},$$

so that

$$\mathbb{P}(S_n > na) \leq e^{-nat} \mathbb{E} e^{tS_n} = (e^{-at} M(t))^n = e^{-n(at - \Lambda(t))}.$$

With  $\mu = 0$  and  $a > 0$  we have

$$\Lambda^*(a) = \sup_{t>0} \{at - \Lambda(t)\},$$

implying

$$\frac{1}{n} \log \mathbb{P}(S_n > na) \leq -\sup_{t>0} \{at - \Lambda(t)\} = -\Lambda^*(a).$$

Next we obtain the lower bound for (i)

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{P}(S_n > na) \geq -\Lambda^*(a).$$

Let  $F$  be the common distribution function of the  $X_i$ . For a "tilted distribution"

$$\tilde{F}(y) = \frac{1}{M(\tau)} \int_{-\infty}^y e^{\tau u} dF(u),$$

we have

$$\tilde{M}(t) = \int_{-\infty}^{\infty} e^{ty} d\tilde{F}(y) = \frac{1}{M(\tau)} \int_{-\infty}^{\infty} e^{ty} e^{\tau y} dF(y) = \frac{M(t + \tau)}{M(\tau)}.$$

Consider  $\tilde{S}_n = \tilde{X}_1 + \dots + \tilde{X}_n$ , where iid r.v.  $(\tilde{X}_n)$  have common distribution  $\tilde{F}$ . Denote

$$F_n(x) = \mathbb{P}(S_n \leq x), \quad \tilde{F}_n(x) = \mathbb{P}(\tilde{S}_n \leq x).$$

From

$$\int_{-\infty}^{\infty} e^{ty} d\tilde{F}_n(y) = \left( \frac{M(t + \tau)}{M(\tau)} \right)^n = \frac{1}{M^n(\tau)} \int_{-\infty}^{\infty} e^{(t+\tau)y} dF_n(y)$$

we see that

$$d\tilde{F}_n(y) = M^{-n}(\tau) e^{\tau y} dF_n(y).$$

Let  $b > a$ . It follows,

$$\begin{aligned} \mathbb{P}(S_n > na) &= \int_{na}^{\infty} dF_n(y) = M^n(\tau) \int_{na}^{\infty} e^{-\tau y} d\tilde{F}_n(y) \\ &\geq M^n(\tau) e^{-\tau nb} \int_{na}^{nb} d\tilde{F}_n(y) = e^{-n(\tau b - \Lambda(\tau))} \mathbb{P}(na < \tilde{S}_n \leq nb). \end{aligned}$$

Since

$$\mathbb{E} \tilde{X}_i = \tilde{M}'(0) = \Lambda'(\tau) = a, \quad \text{Var} \tilde{X}_i = \tilde{M}''(0) - \tilde{M}'(0)^2 = \Lambda''(\tau) \in (0, \infty),$$

we can apply the CLT which gives  $\mathbb{P}(na < \tilde{S}_n \leq nb) \rightarrow 1/2$ . We conclude that

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{P}(S_n > na) \geq -(\tau b - \Lambda(\tau)),$$

and to finish the proof for (i), it remains to send  $b \searrow a$ .

To derive (ii) from (i) we replace  $X_i$  by  $-\tilde{X}_i$  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 (i),

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

**Example 4.12** 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 normalised partial sums  $S_n/n$  are also normally distributed with mean  $\mu$  and variance  $\sigma^2/n$  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

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

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,

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

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 4.13** For  $X_i$  with  $\text{Ber}(p)$  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

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

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  $S_n = a + X_1 + \dots + X_n$  where  $X_1, X_2, \dots$  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  $N - k$ . 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  $\tau_b$  be the first hitting time of  $b$ , then for  $b > 0$

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

and

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

(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  $D_k(N) \rightarrow \frac{k}{q-p}$  as  $N \rightarrow \infty$ .

(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

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

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

$$\begin{aligned} N_n^r(a, b) &= N_n(2r - a, b), \quad a \geq r, b \geq r, \\ N_n^r(a, b) &= N_n(a, 2r - b), \quad a < r, b < r. \end{aligned}$$

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

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

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

$$\begin{aligned} P(S_1 > 0, \dots, S_{n-1} > 0 | S_n = b) &= \frac{b}{n}, \quad b > 0, \\ P(S_1 \neq 0, \dots, S_{n-1} \neq 0, S_n = b) &= \frac{|b|}{n} P(S_n = b), \\ P(S_1 \neq 0, \dots, S_n \neq 0) &= n^{-1} E|S_n|. \end{aligned}$$

It follows that

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

Indeed,

$$\begin{aligned} P(S_1 \neq 0, \dots, S_{2n} \neq 0) &= 2 \sum_{k=1}^n \frac{2k}{2n} P(S_{2n} = 2k) = 2 \sum_{k=1}^n \frac{2k}{2n} \binom{2n}{n+k} 2^{-2n} \\ &= 2^{-2n+1} \sum_{k=1}^n \left( \binom{2n-1}{n+k-1} - \binom{2n-1}{n+k} \right) = 2^{-2n+1} \binom{2n-1}{n} = P(S_{2n} = 0). \end{aligned}$$

(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

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

implying for  $b > 0$

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

The obtained equality

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

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  $\tau_b$  has distribution

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

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

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

**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*

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

PROOF SKETCH. Using (\*) we get

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

and it remains to apply Stirling 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

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

and then (\*), we obtain

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

Then by induction over  $n$  one can show that

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

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

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

where  $\tau_0$  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  $p_{ij}$

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

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

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

**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  $S = \{0, 1, 2, \dots\}$ .

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

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

It is not difficult to see that

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

**Definition 5.5** Classification of states

- state  $i$  is called recurrent (persistent), if  $P(T_i < \infty | X_0 = i) = 1$ ,
- a non-recurrent state is called a transient state,
- a recurrent state  $i$  is called null-recurrent, if  $E(T_i | X_0 = i) = \infty$ ,
- state  $i$  is called positive-recurrent, if  $E(T_i | X_0 = i) < \infty$ .

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

PROOF SKETCH. To prove the first claim observe that  $F_{ii}(1) = P(T_i < \infty | X_0 = i)$ . Using the previous lemma we conclude that state  $i$  is recurrent iff the expected number of visits of the state is infinite  $P_{ii}(1) = \infty$ . The second claim in the aperiodic case follows from the ergodic Theorem 5.14. The periodic case requires an extra argument.

**Definition 5.7** 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.

**Example 5.8** For a simple random walk, we have

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

Notice that it is a periodic chain with period 2. Using the Stirling formula  $n! \sim n^n e^{-n} \sqrt{2\pi n}$  we get

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

Criterion 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 5.9** 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  $\mu_i = E(T_i | X_0 = i)$  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

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

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

$$\begin{aligned} \sum_{j \in S} \rho_j(k) &= \sum_{j \in S} \sum_{n=1}^{\infty} P(X_n = j, T_k \geq n | X_0 = k) \\ &= \sum_{n=1}^{\infty} P(T_k \geq n | X_0 = k) = E(T_k | X_0 = k) = \mu_k. \end{aligned}$$

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

$$y_i = \sum_{j \in S \setminus \{s\}} p_{ij} y_j, \quad i \in S \setminus \{s\}. \quad (*)$$

PROOF SKETCH. Main step. Let  $\tau_j$  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  $(*)$ .

**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  $(*)$ . The chain is transient.
- Solve the equation  $\pi \mathbf{P} = \pi$  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

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

More generally, for an aperiodic state  $j$  and any state  $i$  we have that  $p_{ij}^{(n)} \rightarrow \frac{f_{ij}}{\mu_j}$ , 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

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

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

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

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

PROOF. Using  $(*)$  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  $S = \{0, 1, \dots, m\}$  and transition probabilities

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

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  $\pi_i = \binom{m}{i} 2^{-m}$  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  $\mu = E(X)$  and always assume that

$$P(X = 1) < 1.$$

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

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

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

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

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

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

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

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

and

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

Letting  $n \rightarrow \infty$  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

$$Q_1 = P(X = 0) = h(0) < x_0,$$

so that, by induction,

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

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

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

PROOF. The variance of the generation size  $x_n = \text{Var}(Z_n)$  satisfies iteration

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

because

$$\begin{aligned} \text{Var}(Z_{n+1}) &= E(\text{Var}(Z_{n+1}|Z_n)) + \text{Var}(E(Z_{n+1})|Z_n) \\ &= E(Z_n \sigma^2) + \text{Var}(\mu Z_n). \end{aligned}$$

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  $\lambda_i$ ,
- (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  $\lambda$  is the number of events observed up to time  $t$  given that the inter-arrival times are independent exponentials with parameter  $\lambda$ :

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

To check the last formula observe that

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

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

$$\begin{aligned} P(T_1 + \dots + T_k \leq t) &= \int_0^t \frac{\lambda^k}{(k-1)!} x^{k-1} e^{-\lambda x} dx = \frac{(\lambda t)^k}{(k-1)!} \int_0^1 y^{k-1} e^{-\lambda t y} dy \\ &= \frac{(\lambda t)^k}{k!} e^{-\lambda t} + \frac{(\lambda t)^{k+1}}{k!} \int_0^1 y^k e^{-\lambda t y} dy = \frac{(\lambda t)^k}{k!} e^{-\lambda t} + P(T_1 + \dots + T_{k+1} \leq t). \end{aligned}$$

**Explosion:**  $P(X(t) = \infty) > 0$  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

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

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

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

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

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  $(\lambda_i)$ .

**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:  $\pi\mathbf{P}_t = \pi$  for all  $t$  iff  $\pi\mathbf{G} = \mathbf{0}$ .*

PROOF:

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

**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  $\pi$ , then it is unique and for all  $(i, j)$*

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

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

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

**Definition 6.2** A real-valued process  $\{X(t), t \geq 0\}$  is called weakly stationary with mean  $\mu$  and autocovariance function  $c(h)$  if for all  $t \geq 0$  and  $h \geq 0$

$$\mathbb{E}(X(t)) = \mu, \quad \text{Cov}(X(t), X(t+h)) = c(h) \in (-\infty, \infty).$$

Given that  $c(0) > 0$ , the ratio  $\rho(h) = \frac{c(h)}{c(0)}$  is called the autocorrelation function of the weakly stationary process.

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

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

More specifically, look at Example 5.13.

**Example 6.4** A sequence  $\{X_n, n = 1, 2, \dots\}$  of independent random variables with the Cauchy distribution is an example of a strongly stationary process, which is not weakly stationary.

## 6.2 Linear prediction

Knowing the past  $(X_r, X_{r-1}, \dots, X_{r-s})$  we would like to predict a future value  $X_{r+k}$  by a linear combination

$$\hat{X}_{r+k} = a + a_0 X_r + a_1 X_{r-1} + \dots + a_s X_{r-s}.$$

We will call  $\hat{X}_{r+k}$  the best linear predictor if  $a = a(k, s)$ ,  $a_j = a_j(k, s)$ ,  $0 \leq j \leq s$  are chosen to minimise the mean-square error size  $E(X_{r+k} - \hat{X}_{r+k})^2$ . (Compare the best linear predictor to the best predictor  $E(X_{r+k} | X_r, X_{r-1}, \dots, X_{r-s})$ , see Section 2.2.)

**Theorem 6.5** For a weakly stationary sequence with mean  $\mu$  and autocovariance  $c(m)$ , the best linear predictor is given by

$$a = \mu(1 - a_0 - \dots - a_s),$$

and the vector  $(a_0, \dots, a_s)$  satisfying the system of linear equations

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

PROOF. By subtracting the mean  $\mu$  we reduce the problem to the zero-mean case. Geometrically, in the zero-mean case, the best linear predictor  $\hat{X}_{r+k}$  makes an error,  $X_{r+k} - \hat{X}_{r+k}$ , which is *orthogonal* to the past  $(X_r, X_{r-1}, \dots, X_{r-s})$ , in that the covariances are equal to zero:

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

Plugging  $\hat{X}_{r+k} = a_0 X_r + a_1 X_{r-1} + \dots + a_s X_{r-s}$  into the last relation, we arrive at the claimed equations.

To justify the geometric intuition let  $H$  be a linear space generated by  $(X_r, X_{r-1}, \dots, X_{r-s})$ . In the zero-mean case, if  $W \in H$ , then for any real  $a$ ,

$$E(X_{r+k} - a - W)^2 = a^2 + E(X_{r+k} - W)^2 \geq E(X_{r+k} - W)^2,$$

showing that the best linear predictor belongs to  $H$ . We have to show that if  $W \in H$  is such that for all  $Z \in H$ ,

$$E(X_{r+k} - W)^2 \leq E(X_{r+k} - Z)^2,$$

then

$$E((X_{r+k} - W)X_{r-m}) = 0, \quad m = 0, \dots, s.$$

Indeed, suppose to the contrary that for some  $m$ ,

$$E((X_{r+k} - W)X_{r-m}) = c(0)d, \quad d > 0.$$

Then writing  $W' = W + d \cdot X_{r-m}$  we arrive at a contradiction

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

**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  $Y_n = \sum_{m \geq 0} \alpha^m Z_{n-m}$  is weakly stationary with zero mean and autocovariance for  $m \geq 0$ ,

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

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

$$\begin{aligned} a_0 + a_1\alpha + a_2\alpha^2 + \dots + a_s\alpha^s &= \alpha^k, \\ a_0\alpha + a_1 + a_2\alpha + \dots + a_s\alpha^{s-1} &= \alpha^{k+1}. \end{aligned}$$

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.

### 6.3 Linear combination of sinusoids

**Example 6.8** For a sequence of fixed frequencies  $0 \leq \lambda_1 < \dots < \lambda_k < \infty$  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  $A_1, B_1, \dots, A_k, B_k$  are uncorrelated r.v. with zero means and  $\text{Var}(A_j) = \text{Var}(B_j) = \sigma_j^2$ . Then

$$\begin{aligned} \text{Cov}(X(t), X(s)) &= \mathbb{E}(X(t)X(s)) = \sum_{j=1}^k \mathbb{E}(A_j^2 \cos(\lambda_j t) \cos(\lambda_j s) + B_j^2 \sin(\lambda_j t) \sin(\lambda_j s)) \\ &= \sum_{j=1}^k \sigma_j^2 \cos(\lambda_j(s-t)), \\ \text{Var}(X(t)) &= \sum_{j=1}^k \sigma_j^2. \end{aligned}$$

Thus  $X(t)$  is weakly stationary with  $\mu = 0$  and

$$\begin{aligned} c(t) &= \sum_{j=1}^k \sigma_j^2 \cos(\lambda_j t), \quad c(0) = \sum_{j=1}^k \sigma_j^2, \\ \rho(t) &= \frac{c(t)}{c(0)} = \sum_{j=1}^k g_j \cos(\lambda_j t) = \int_0^\infty \cos(\lambda t) dG(\lambda), \end{aligned}$$

where  $G$  is a distribution function over  $[0, \infty)$  defined as

$$g_j = \frac{\sigma_j^2}{\sigma_1^2 + \dots + \sigma_k^2}, \quad G(\lambda) = \sum_{j: \lambda_j \leq \lambda} g_j.$$

We can write

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

where

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

**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

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

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

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

**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  $k = 1$ ,  $\lambda = 1$ ,  $\sigma_1^2 = \frac{1}{2}$ . It is a strongly stationary process, since  $X(t + h) = \cos(t + Y')$ , where  $Y' = Y + h$  is uniformly distributed over  $[0, 2\pi]$  modulo  $2\pi$ .

To find the distribution of  $X(t)$  observe that for an arbitrary bounded measurable function  $\phi(x)$ ,

$$\begin{aligned} E(\phi(X(t))) &= E(\phi(\cos(t + Y))) = \frac{1}{2\pi} \int_0^{2\pi} \phi(\cos(t + y)) dy = \frac{1}{2\pi} \int_t^{t+2\pi} \phi(\cos(z)) dz = \frac{1}{2\pi} \int_0^{2\pi} \phi(\cos(z)) dz \\ &= \frac{1}{2\pi} \left( \int_0^\pi \phi(\cos(z)) dz + \int_\pi^{2\pi} \phi(\cos(z)) dz \right) = \frac{1}{2\pi} \left( \int_0^\pi \phi(\cos(\pi - y)) dy + \int_0^\pi \phi(\cos(\pi + y)) dy \right) \\ &= \frac{1}{\pi} \int_0^\pi \phi(-\cos(y)) dy. \end{aligned}$$

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

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

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

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

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

$$\begin{aligned} E(X_n) &= 0, \quad c(n) = \sum_{j=1}^k \sigma_j^2 \cos(\lambda_j n), \quad \rho(n) = \int_0^\pi \cos(\lambda n) dG(\lambda), \\ X_n &= \int_0^\pi \cos(n\lambda) dU(\lambda) + \int_0^\pi \sin(n\lambda) dV(\lambda). \end{aligned}$$

## 6.4 Spectral representation

It turns out that any weakly stationary process  $\{X(t) : -\infty < t < \infty\}$  with zero mean can be approximated by a linear combination of uncorrelated sinusoids.

The autocovariance  $c(t)$  of a stationary process is a non-negative definite function in that

$$\sum_{j=1}^n \sum_{k=1}^n c(t_k - t_j) z_j z_k = \text{Var} \left( \sum_{k=1}^n z_k X(t_k) \right) \geq 0$$

for any  $t_1, \dots, t_n$  and  $z_1, \dots, z_n$ . 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  $\lambda$  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

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

**Exercise 6.13** Find the autocorrelation function of a stationary process with the spectral density function

- (a)  $f(\lambda) = \sqrt{2/\pi} e^{-\lambda^2/2}$  for  $\lambda \geq 0$ ,
- (b)  $f(\lambda) = e^{-\lambda}$  for  $\lambda \geq 0$ .

Discuss the differences in the dependence patterns for these two cases.

SOLUTION. (a) To compute

$$\rho(t) = \sqrt{2/\pi} \int_0^\infty \cos(\lambda t) e^{-\lambda^2/2} d\lambda = \sqrt{1/(2\pi)} \int_{-\infty}^\infty \cos(\lambda t) e^{-\lambda^2/2} d\lambda,$$

we can use the formula for the characteristic function of the standard normal distribution

$$\sqrt{1/(2\pi)} \int_{-\infty}^\infty e^{it\lambda} e^{-\lambda^2/2} d\lambda = e^{-t^2/2}.$$

We get immediately that  $\rho(t) = e^{-t^2/2}$ . The correlation between two values of the process separate by  $t$  units of time very quickly decreases with increasing  $t$ .

(b) To compute

$$\rho(t) = \int_0^\infty \cos(\lambda t) e^{-\lambda} d\lambda,$$

we can use the formula for the characteristic function of the exponential distribution

$$\int_{-\infty}^\infty e^{it\lambda} e^{-\lambda} d\lambda = \frac{1}{1-it} = \frac{1+it}{1+t^2}.$$

We derive  $\rho(t) = \frac{1}{1+t^2}$ . Correlation decreases much slower than in the previous case.

**Example 6.14** Consider an irreducible continuous time Markov chain  $\{X(t), t \geq 0\}$  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

$$\begin{aligned} p_{11}(t) &= 1 - p_{12}(t) = \frac{\beta}{\alpha + \beta} + \frac{\alpha}{\alpha + \beta} e^{-t(\alpha+\beta)}, \\ p_{22}(t) &= 1 - p_{21}(t) = \frac{\alpha}{\alpha + \beta} + \frac{\beta}{\alpha + \beta} e^{-t(\alpha+\beta)}, \end{aligned}$$

and we find for  $t \geq 0$

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

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

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

**Example 6.15** Discrete white noise: a sequence  $X_0, X_1, \dots$  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 6.16** 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 an orthogonal pair of 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  $\text{Var}(U(\lambda)) = \text{Var}(V(\lambda)) = G(\lambda)$ .

**Theorem 6.17** 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 an orthogonal pair of 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  $\text{Var}(U(\lambda)) = \text{Var}(V(\lambda)) = G(\lambda)$ .

## 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  $t$ ,
- $\mathbb{E}(|S(t+h) - S(t)|^2) \rightarrow 0$  as  $h \searrow 0$  for all  $t$ ,
- orthogonal increments:  $\mathbb{E}([S(v) - S(u)][\bar{S}(t) - \bar{S}(s)]) = 0$  whenever  $u < v \leq s < t$ .

Put

$$F(t) := \begin{cases} \mathbb{E}(|S(t) - S(0)|^2), & \text{if } t \geq 0, \\ -\mathbb{E}(|S(t) - S(0)|^2), & \text{if } t < 0. \end{cases}$$

Since the process has orthogonal increments we obtain

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

implying that  $F$  is monotonic and right-continuous.

Let  $\psi : \mathbb{R} \rightarrow \mathbb{C}$  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  $\psi$  with respect to  $S$ ,

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

possessing the following important property

$$E(I(\psi_1)I(\psi_2)) = \int_{-\infty}^{\infty} \psi_1(t)\psi_2(t)dF(t). \quad (**)$$

1. For an arbitrary step function

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

put

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

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

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

2. There exists a sequence of step functions such that

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

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

PROOF SKETCH of Theorem 6.17 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  $a_1, a_2, \dots \in \mathbb{C}$ ,  $n \in \mathbb{N}$ ,  $m_1, m_2, \dots \in \mathbb{Z}$ . Similarly, let  $H_F$  be the set of linear combinations of sinusoids  $f_n(x) := e^{inx}$ . Define the linear mapping  $\mu : H_F \rightarrow H_X$  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

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

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

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 (\*). 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:  $dS(\lambda) = \overline{dS(-\lambda)}$ .

**Example 6.18** The Wiener process  $W_t$  is an example of a process  $S(t)$  with orthogonal increments. In this case the key property (\*\*\*) holds with  $F(t) = t$ :

$$\mathbb{E} \left( \int_0^\infty \psi_1(t) dW_t \times \int_0^\infty \psi_2(t) dW_t \right) = \int_0^\infty \psi_1(t) \psi_2(t) dt.$$

**Exercise 6.19** Check the properties of  $U(\lambda)$  and  $V(\lambda)$  stated in Theorem 6.17 with  $dG(\lambda) = 2dF(\lambda)$  for  $\lambda > 0$  and  $G(0) - G(0-) = F(0) - F(0-)$ .

## 6.6 Ergodic theorem for the weakly stationary processes

**Theorem 6.20** Let  $\{X_n, n = 1, 2, \dots\}$  be a weakly stationary process with mean  $\mu$  and autocovariance function  $c(m)$ . There exists a r.v.  $Y$  with mean  $\mu$  and variance

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

such that

$$\frac{X_1 + \dots + X_n}{n} \xrightarrow{L^2} Y, \quad n \rightarrow \infty.$$

PROOF SKETCH. 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

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

where

$$g_n(\lambda) = \frac{\cos(\lambda) + \dots + \cos(n\lambda)}{n}, \quad h_n(\lambda) = \frac{\sin(\lambda) + \dots + \sin(n\lambda)}{n}.$$

In terms of complex numbers we have

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

The last expression equals 1 for  $\lambda = 0$ . For  $\lambda \neq 0$ , we get

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

so that

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

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

$$\begin{aligned} \int_0^\pi g_n(\lambda) dU(\lambda) &\xrightarrow{L^2} \int_0^\pi 1_{\{\lambda=0\}} dU(\lambda) = U(0) - U(0-), \\ \int_0^\pi h_n(\lambda) dV(\lambda) &\xrightarrow{L^2} 0. \end{aligned}$$

Thus  $\bar{X}_n \xrightarrow{L^2} Y := U(0) - U(0-)$  and by Theorem 6.17, we have

$$\text{Var}(Y) = \text{Var}(U(0) - U(0-)) = c(0)(G(0) - G(0-)).$$

It remains to observe that

$$\lim_{n \rightarrow \infty} n^{-1} \sum_{j=1}^n c(j) = c(0)(G(0) - G(0-)),$$

which follows from the representation

$$n^{-1} \sum_{j=1}^n c(j) = c(0) \int_0^\pi g_n(\lambda) dG(\lambda).$$

## 6.7 Ergodic theorem for the strongly stationary processes

**Theorem 6.21** *If  $\{X_n, n = 1, 2, \dots\}$  is a strongly stationary sequence with a finite mean  $\mu$ , then there is a random variable  $Y$  with the mean  $\mu$  such that*

$$\bar{X}_n := \frac{X_1 + \dots + X_n}{n} \rightarrow Y \text{ a.s. and in } L^1.$$

In the ergodic case, when  $\text{Var}(Y) = 0$ ,

$$\bar{X}_n \rightarrow \mu \text{ a.s. and in } L^1.$$

WITHOUT A PROOF.

**Example 6.22** Let  $Z_1, \dots, Z_k$  be iid with a finite mean  $\mu$ . Then the following cyclic process

$$\begin{aligned} X_1 &= Z_1, \dots, X_k = Z_k, \\ X_{k+1} &= Z_1, \dots, X_{2k} = Z_k, \\ X_{2k+1} &= Z_1, \dots, X_{3k} = Z_k, \dots, \end{aligned}$$

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

$$\frac{X_1 + \dots + X_n}{n} \rightarrow \frac{Z_1 + \dots + Z_k}{k}.$$

**Example 6.23** Let  $\{X_n, n = 1, 2, \dots\}$  be an irreducible positive-recurrent Markov chain with the state space  $S = \{0, \pm 1, \pm 2, \dots\}$ . Let  $\pi = (\pi_j)_{j \in S}$  be the unique stationary distribution. If  $X_0$  has distribution  $\pi$ , 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 mean  $\mu = \pi_k$  and autocovariance function

$$c(m) = \text{Cov}(I_n, I_{n+m}) = E(I_n I_{n+m}) - \pi_k^2 = \pi_k(p_{kk}^{(m)} - \pi_k).$$

Since  $p_{kk}^{(m)} \rightarrow \pi_k$  as  $m \rightarrow \infty$  we have  $c(m) \rightarrow 0$  and the limit in Theorem 6.20 has zero variance. It follows that  $n^{-1}(I_1 + \dots + I_n)$ , the proportion of  $(X_1, \dots, X_n)$  visiting state  $k$ , converges to  $\pi_k$  as  $n \rightarrow \infty$  in  $L^2$ . It follows that this process is ergodic and the convergence also holds almost surely.

**Example 6.24** 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  $Y_{n+1} = 2Y_n \pmod{1}$ . We see that  $(Y_n)$  is a strictly stationary Markov chain. From  $Y_{n+1} = 2^n X \pmod{1}$  we derive

$$\begin{aligned} E(Y_1 Y_{n+1}) &= \int_0^1 x(2^n x)_{\text{mod}1} dx = \sum_{j=0}^{2^n-1} \int_{j2^{-n}}^{(j+1)2^{-n}} x(2^n x - j) dx \\ &= 2^{-2n} \sum_{j=0}^{2^n-1} \int_0^1 (y+j)y dy = 2^{-2n} \sum_{j=0}^{2^n-1} \left( \frac{1}{3} + \frac{j}{2} \right) = \frac{1}{4} + \frac{2^{-n}}{12}. \end{aligned}$$

Thus  $c(n) = \frac{2^{-n}}{12}$  tends to zero as  $n \rightarrow \infty$ . This implies that  $n^{-1} \sum_{j=1}^n Y_j \rightarrow 1/2$  in  $L^2$  and almost surely.

## 6.8 Gaussian processes

**Definition 6.25** A random process  $\{X(t), t \geq 0\}$  is called Gaussian if for any  $(t_1, \dots, t_n)$  the vector  $(X(t_1), \dots, X(t_n))$  has a multivariate normal distribution, see Section 2.4.

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

**Theorem 6.26** A Gaussian process  $\{X(t), t \geq 0\}$  is Markov iff for any  $0 \leq t_1 < \dots < t_n$

$$E(X(t_n) | X(t_1), \dots, X(t_{n-1})) = E(X(t_n) | X(t_{n-1})). \quad (*)$$

PROOF. Clearly, the Markov property implies (\*). To prove the converse we have to show that in the Gaussian case (\*) gives

$$\text{Var}(X(t_n) | X(t_1), \dots, X(t_{n-1})) = \text{Var}(X(t_n) | X(t_{n-1})),$$

since the conditional distribution is determined by the conditional mean and variance. The last equality is verified using the fact that  $X(t_n) - E\{X(t_n) | X(t_1), \dots, X(t_{n-1})\}$  is orthogonal to  $(X(t_1), \dots, X(t_{n-1}))$ , see Section 2.2, which in the Gaussian case means independence:

$$\begin{aligned} &E\left\{ \left( X(t_n) - E\{X(t_n) | X(t_1), \dots, X(t_{n-1})\} \right)^2 | X(t_1), \dots, X(t_{n-1}) \right\} \\ &= E\left\{ \left( X(t_n) - E\{X(t_n) | X(t_1), \dots, X(t_{n-1})\} \right)^2 \right\} = E\left\{ \left( X(t_n) - E\{X(t_n) | X(t_{n-1})\} \right)^2 \right\} \\ &= E\left\{ \left( X(t_n) - E\{X(t_n) | X(t_{n-1})\} \right)^2 | X(t_{n-1}) \right\}. \end{aligned}$$

**Example 6.27** A stationary Gaussian Markov process is called the Ornstein-Uhlenbeck process. It is characterized by three parameters  $(\theta, \alpha, \sigma^2)$ , where  $\theta \in (-\infty, \infty)$  is the stationary mean of the process,  $\alpha > 0$  is the rate of attraction to the stationary mean, and  $\sigma^2$  is the noise variance. In this case the autocorrelation function has the form  $\rho(t) = e^{-\alpha t}$ ,  $t \geq 0$ . 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

$$E(X(t+s) | X(s)) = \theta + \rho(t)(X(s) - \theta)$$

we derive

$$\begin{aligned}\rho(t+s) &= c(0)^{-1} \mathbf{E}((X(t+s) - \theta)(X(0) - \theta)) = c(0)^{-1} \mathbf{E}\{\mathbf{E}((X(t+s) - \theta)(X(0) - \theta) | X(0), X(s))\} \\ &= \rho(t)c(0)^{-1} \mathbf{E}((X(s) - \theta)(X(0) - \theta)) \\ &= \rho(t)\rho(s).\end{aligned}$$

See also Example 9.5.

## 7 Renewal theory and Queues

### 7.1 Renewal function

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

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

The distribution function of  $P(T_k \leq t) = F^{*k}(t)$  is a convolution of  $F(t) = P(X_1 \leq t)$ :

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

**Exercise 7.1** Show that

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

**Definition 7.2** Put  $m(t) := \mathbf{E}(N(t))$ . 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

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

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  $\lambda = 1/\mu$ .



## 7.2 Renewal equation and excess life

**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

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

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

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

In particular,  $P(C(t) = t) = 1 - F(t)$ .

PROOF. Consider the tail distribution function  $b(t) = P(E(t) > y)$ . The first claim follows from the following renewal equation

$$\begin{aligned} b(t) &= \mathbb{E}\left(\mathbb{E}(1_{\{E(t) > y\}} | X_1)\right) = \mathbb{E}\left(1_{\{\tilde{E}(t-X_1) > y\}} 1_{\{X_1 \leq t\}} + 1_{\{X_1 > t+y\}}\right) \\ &= 1 - F(t+y) + \int_0^t b(t-x)dF(x). \end{aligned}$$

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

$$P(C(t) \geq y) = P(E(t-y) > y) \quad (*)$$

and the second claim follows from the just proven first claim.

**Example 7.8** For the Poisson process with rate  $\lambda$ , the distribution of the excess lifetime is given by

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

so that it is independent of the observation time  $t$ . Therefore, in view of (\*), we have for  $y \leq t$ ,

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

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".

**Exercise 7.9** Let the times between the events of a renewal process  $N$  be uniformly distributed on  $[0, 1]$ .

- For the mean function  $m(t) = \mathbb{E}(N(t))$ , show that  $m(t) = e^t - 1$  for  $0 \leq t \leq 1$ .
- Show that for  $0 \leq t \leq 1$ , the variance is

$$\text{Var}(N(t)) = e^t(1 + 2t - e^t).$$

Hint: start by writing down the renewal equations for  $m(t)$  and the second moment  $m_2(t) = \mathbb{E}(N^2(t))$ .

SOLUTION. (a) Since the inter-arrival times are uniformly distributed on  $[0, 1]$ , we have  $F(t) = t \cdot 1_{\{0 \leq t \leq 1\}}$ . The renewal property gives

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

Taking expectations we see that  $m(t) = \mathbb{E}N(t)$  satisfies the renewal equation

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

For  $0 \leq t \leq 1$ , we have

$$m(t) = t + \int_0^t m(u)du,$$

and therefore  $m'(t) = 1 + m(t)$  with  $m(0) = 0$ . Thus  $m(t) = e^t - 1$  for  $0 \leq t \leq 1$ .

(b) From

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

we get

$$N^2(t) = 1_{\{X_1 \leq t\}}(1 + 2\tilde{N}(t - X_1) + \tilde{N}^2(t - X_1)).$$

Taking expectations we obtain the renewal equation for the second moment  $m_2(t) = \mathbb{E}(N^2(t))$

$$m_2(t) = A(t) + \int_0^t m_2(t-u)dF(u),$$

where

$$A(t) = \int_0^t (1 + 2m(t-u))dF(u).$$

Using the renewal function  $U(t) = 1 + m(t)$  we find the solution of this renewal equation as

$$m_2(t) = \int_0^t A(t-u)dU(u).$$

For  $0 \leq t \leq 1$ , we have  $U(t) = e^t$  and

$$A(t) = \int_0^t (2e^u - 1)du = 2e^t - 2 - t,$$

so that

$$\begin{aligned} m_2(t) &= 2e^t - 2 - t + \int_0^t (2e^{t-u} - 2 - t + u)e^u du \\ &= te^t + \int_0^t ue^u du = 1 - e^t + 2te^t. \end{aligned}$$

Thus for  $0 \leq t \leq 1$ , the variance is

$$\text{Var}(N(t)) = 1 - e^t + 2te^t - (e^t - 1)^2 = e^t + 2te^t - e^{2t} = e^t(1 + 2t - e^t).$$

### 7.3 LLN and CLT for the renewal process

**Theorem 7.10** *Law of large numbers:  $N(t)/t \xrightarrow{\text{a.s.}} 1/\mu$  as  $t \rightarrow \infty$ . 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  $N(t) > 0$ ,

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

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

**Theorem 7.11** *Central limit theorem.* If  $\sigma^2 = \text{Var}(X_1)$  is positive and finite, then

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

PROOF. The usual CLT implies that for any  $x$ ,

$$\mathbb{P}\left(\frac{T_{a(t)} - \mu a(t)}{\sigma\sqrt{a(t)}} \leq x\right) \rightarrow \Phi(x) \text{ as } a(t) \rightarrow \infty.$$

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

$$\mathbb{P}\left(\frac{N(t) - t/\mu}{\sqrt{t\sigma^2/\mu^3}} \geq x\right) = \mathbb{P}(N(t) \geq a(t)) = \mathbb{P}(T_{a(t)} \leq t),$$

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

$$\mathbb{P}\left(\frac{N(t) - t/\mu}{\sqrt{t\sigma^2/\mu^3}} \geq x\right) \rightarrow \Phi(-x) = 1 - \Phi(x).$$

## 7.4 Stopping times and Wald's equation

**Definition 7.12** Let  $M$  be a r.v. taking values in the set  $\{1, 2, \dots\}$ . We call it a stopping time with respect to the sequence  $(X_n)$  of random variables, if

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

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

**Theorem 7.13** *Wald's equation.* Let  $X_1, X_2, \dots$  be iid r.v. with finite mean  $\mu$ , and let  $M$  be a stopping time with respect to the sequence  $(X_n)$  such that  $\mathbb{E}(M) < \infty$ . Then

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

PROOF. Observe that

$$\sum_{i=1}^{\infty} X_i 1_{\{M \geq i\}} = \sum_{i=1}^{\infty} X_i \sum_{j=i}^{\infty} 1_{\{M=j\}} = \sum_{j=1}^{\infty} 1_{\{M=j\}} \sum_{i=1}^j X_i = (X_1 + \dots + X_M) \sum_{j=1}^{\infty} 1_{\{M=j\}} = X_1 + \dots + X_M,$$

so that

$$Y_n := \sum_{i=1}^n X_i 1_{\{M \geq i\}} \xrightarrow{\text{a.s.}} X_1 + \dots + X_M.$$

For all  $n$ , we have

$$|Y_n| \leq Y, \quad Y = \sum_{i=1}^{\infty} |X_i| 1_{\{M \geq i\}},$$

and

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

Here we used independence between  $\{M \geq i\}$  and  $X_i$ , which follows from the fact that  $\{M \geq i\}$  is the complementary event to  $\{M \leq i-1\} \in \sigma\{X_1, \dots, X_{i-1}\}$ . By the dominated convergence Theorem 3.21

$$\mathbb{E}(X_1 + \dots + X_M) = \lim_{n \rightarrow \infty} \mathbb{E}\left(\sum_{i=1}^n X_i 1_{\{M \geq i\}}\right) = \sum_{i=1}^{\infty} \mathbb{E}(X_i) \mathbb{P}(M \geq i) = \mu\mathbb{E}(M).$$

**Example 7.14** Observe that  $M = N(t)$  is not a stopping time with respect to the sequence  $(X_n)$  of inter-arrival times, and in general  $E(T_{N(t)}) \neq \mu m(t)$ . Indeed, for the Poisson process  $\mu m(t) = t$  while  $T_{N(t)} = t - C(t)$ , where  $C(t)$  is the current lifetime.

**Theorem 7.15** *Elementary renewal theorem:  $m(t)/t \rightarrow 1/\mu$  as  $t \rightarrow \infty$ , where  $\mu = E(X_1)$  is finite or infinite.*

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

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

the Wald equation implies

$$E(T_{N(t)+1}) = \mu E(N(t) + 1) = \mu U(t).$$

From  $T_{N(t)+1} = t + E(t)$  we get

$$U(t) = \mu^{-1}(t + E(E(t))),$$

so that  $U(t) \geq \mu^{-1}t$ . Moreover, if  $P(X_1 \leq a) = 1$  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  $\mu_a$  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  $\mu_a \rightarrow \mu$  as  $a \rightarrow \infty$ .

## 7.5 Renewal theorems and stationarity

**Theorem 7.16** *Renewal theorem. If  $X_1$  is not arithmetic, then for any positive  $h$*

$$U(t+h) - U(t) \rightarrow \mu^{-1}h, \quad t \rightarrow \infty.$$

WITHOUT PROOF.

**Example 7.17** 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  $R_X \subset \{1, 2, \dots\}$  and  $R_X \not\subset \{k, 2k, \dots\}$  for any  $k = 2, 3, \dots$ , 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 7.16 claims that  $U(n) - U(n-1) \rightarrow \mu^{-1}$ .

**Theorem 7.18** *Key renewal theorem. If  $X_1$  is not arithmetic,  $\mu < \infty$ , and  $g : [0, \infty) \rightarrow [0, \infty)$  is a monotone function, then*

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

PROOF SKETCH. Using Theorem 7.16, 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 7.19** *If  $X_1$  is not arithmetic and  $\mu < \infty$ , then*

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

PROOF. Apply the key renewal theorem and Lemma 7.7.

**Definition 7.20** Let  $X_1, X_2, \dots$  be independent positive r.v. such that  $X_2, X_3, \dots$  have the same distribution. If as before,  $T_0 = 0$  and  $T_n = X_1 + \dots + X_n$ , then  $N^d(t) = \max\{n : T_n \leq t\}$  is called a *delayed renewal process*. It is described by two distributions  $F(t) = P(X_i \leq t)$ ,  $i \geq 2$  and  $F^d(t) = P(X_1 \leq t)$ .

**Lemma 7.21** *The mean  $m^d(t) = EN^d(t)$  satisfies the renewal equation*

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

PROOF. First observe that conditioning on  $X_1$  gives

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

Now since

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

we have

$$\begin{aligned} m * F^d(t) &:= \int_0^t m(t-u) dF^d(u) = F * F^d(t) + m * F * F^d(t) = F^d * F(t) + m * F^d * F(t) \\ &= (F^d + m * F^d) * F(t) = m^d * F(t) = \int_0^t m^d(t-u) dF(u). \end{aligned}$$

**Theorem 7.22** *The process  $N^d(t)$  has stationary increments:  $N^d(s+t) - N^d(s) \stackrel{d}{=} N^d(t)$ , if and only if*

$$F^d(y) = \mu^{-1} \int_0^y (1 - F(x)) dx. \quad (*)$$

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

PROOF. Necessity. If  $N^d(t)$  has stationary increments, then  $m^d(s+t) = m^d(s) + m^d(t)$ , and we get  $m^d(t) = tm^d(1)$ . Substitute this into the renewal equation of Lemma 7.21 to obtain

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

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

Sufficiency. Given (\*), we have

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

Taking the Laplace-Stieltjes transforms in Lemma 7.21 yields

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

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

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

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

and

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

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

$$\begin{aligned} \int_0^t P(E^d(t-u) > y) dF^d(u) &= g * U * F^d(t) = g * F^d * U(t) = \mu^{-1} \int_0^t (1 - F(t+y-u)) du \\ &= \mu^{-1} \int_y^{t+y} (1 - F(x)) dx = F^d(t+y) - F^d(y). \end{aligned}$$

## 7.6 Renewal-reward processes

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

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

**Theorem 7.23** *Renewal-reward theorem.* Suppose  $(X_i, R_i)$  have finite means  $\mu = E(X)$  and  $E(R)$ . Then

$$\frac{W(t)}{t} \xrightarrow{\text{a.s.}} \frac{E(R)}{\mu}, \quad \frac{EW(t)}{t} \rightarrow \frac{E(R)}{\mu}, \quad t \rightarrow \infty.$$

PROOF. Applying two laws of large numbers, Theorem 4.9 and Theorem 7.10, we get the first claim

$$\frac{W(t)}{t} = \frac{R_1 + \dots + R_{N(t)}}{N(t)} \cdot \frac{N(t)}{t} \xrightarrow{\text{a.s.}} \frac{E(R)}{\mu}.$$

The second claim is an extension of Theorem 7.15. Again, using the Wald equation we find

$$EW(t) = E(R_1 + \dots + R_{N(t)+1}) - ER_{N(t)+1} = U(t)E(R) - r(t), \quad r(t) = E(R_{N(t)+1}).$$

The result will follow once we have shown that  $r(t)/t \rightarrow 0$  as  $t \rightarrow \infty$ . By conditioning on  $X_1$  we arrive at the renewal equation

$$r(t) = H(t) + \int_0^t r(t-u)dF(u), \quad H(t) = E(R_1 1_{\{X_1 > t\}}).$$

Since  $H(t) \rightarrow 0$ , for any  $\epsilon > 0$ , there exists a finite  $M = M(\epsilon)$  such that  $|H(t)| < \epsilon$  for  $t \geq M$ . Therefore, when  $t \geq M$ ,

$$\begin{aligned} t^{-1}|r(t)| &\leq t^{-1} \int_0^{t-M} |H(t-u)|dU(u) + t^{-1} \int_{t-M}^t |H(t-u)|dU(u) \\ &\leq t^{-1} \left( \epsilon U(t) + (U(t) - U(t-M))E|R| \right). \end{aligned}$$

Using the renewal theorems we get  $\limsup |r(t)/t| \leq \epsilon/\mu$  and it remains to let  $\epsilon \rightarrow 0$ .

## 7.7 Regeneration technique for queues

Consider a general queueing system assuming that 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, so that  $Q(T) = 0$ . This can be called a regeneration time, because 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*, that is  $P(T < \infty) = 1$ , we get a renewal process of regeneration times  $0 = T_0 < T_1 < T_2 < T_3 < \dots$ . Write  $N_i$  for the number of customers arriving during the cycle  $[T_{i-1}, T_i)$  and put  $N = N_1, T = T_1$ . To be able to apply Theorem 7.23 we shall assume

$$E(T) < \infty, \quad E(N) < \infty, \quad E(NT) < \infty.$$

(A) Let the reward associated with the inter-arrival time  $X_i = T_i - T_{i-1}$  to be

$$R_i = \int_{T_{i-1}}^{T_i} Q(u)du.$$

We see that

$$\frac{1}{t}(R_1 + \dots + R_{N(t)}) \approx \frac{1}{t} \int_0^t Q(u)du = \sum_{k=1}^{\infty} k \cdot \frac{\text{time with } k \text{ customers during } [0, t]}{t},$$

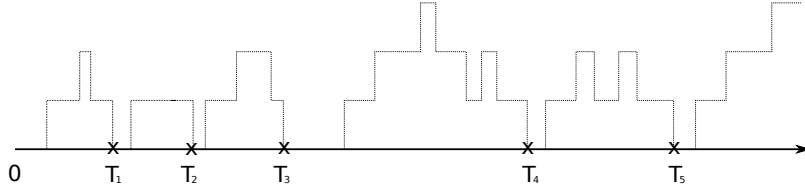


Figure 2: The queue length and the regeneration times. Here  $N_1 = 2$ ,  $N_2 = 1$ ,  $N_3 = 2$ ,  $N_4 = 4$ ,  $N_5 = 3$ .

where the last expression is the time average queue length. Since  $R := R_1 \leq NT$ , we have  $E(R) \leq E(NT) < \infty$ , and by Theorem 7.23,

$$t^{-1} \int_0^t Q(u) du \xrightarrow{\text{a.s.}} E(R)/E(T) =: L, \quad \text{the long run average queue length.}$$

(B) Let the reward associated with the inter-arrival time  $X_i$  to be the number of new customers  $N_i$ . Observe that

$$W(t) = \sum_{i=1}^{N(t)} N_i$$

is approximately the number of customers arrived by time  $t$ . By Theorem 7.23,

$$W(t)/t \xrightarrow{\text{a.s.}} E(N)/E(T) =: \lambda, \quad \text{the long run rate of arrival.}$$

(C) Consider now the renewal process  $\{K_n, n \geq 0\}$  with discrete inter-arrival times  $N_i$ , and the associated reward-renewal process with rewards  $S_i$  defined as the total time spent in the system by the customers arrived during  $[T_{i-1}, T_i)$ . If the  $n$ -th customer spends time  $V_n$ , then putting  $S = S_1$ ,  $N = N_1$ , we get

$$S = V_1 + \dots + V_N.$$

We see that

$$S_1 + \dots + S_{K_n} = \sum_{i=1}^{K_n} (V_1 + \dots + V_{N_i}) \approx \sum_{i=1}^n V_i,$$

where the last sum is the time spent in the system by the first  $n$  customers. Since  $E(S) \leq E(NT) < \infty$ , again by Theorem 7.23, we have

$$n^{-1} \sum_{i=1}^n V_i \xrightarrow{\text{a.s.}} E(S)/E(N) =: \nu, \quad \text{the long run average time spent by a customer in the system.}$$

**Theorem 7.24** *Little's law.* If  $E(T) < \infty$ ,  $E(N) < \infty$ , and  $E(NT) < \infty$ , then  $L = \lambda\nu$ .

Although it looks intuitively reasonable:

average queue length = average time in the system per customer  $\times$  arrival rate of customers,

it is 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 total of customer time spent during the first cycle is

$$S = \sum_{i=1}^N V_i = \int_0^T Q(u) du.$$

Therefore,

$$E(S) = E\left(\sum_{i=1}^N V_i\right) = E\left(\int_0^T Q(u) du\right) = E(R).$$

Thus, combining (A), (B), (C) we get

$$\frac{L}{\lambda\nu} = \frac{E(R)}{E(T)} \cdot \frac{E(T)}{E(N)} \cdot \frac{E(N)}{E(S)} = 1.$$

## 7.8 M/M/1 queues

Consider a queueing system with the inter-arrival times  $(X_i)$  and service times for a single service  $(S_i)$  being two independent iid sequences. The Kendall notation annotates such a queueing system by a triple

distribution of  $X$ /distribution of  $S$ /number of servers.

Here we usually assume that number of servers is 1, and that the FIFO service system is "first in first out". We will be mostly interested in the queue length  $Q(t)$  at time  $t$ . The ratio

$$\rho = \frac{\mathbb{E}(S)}{\mathbb{E}(X)}$$

will be called the traffic intensity. The light traffic case is  $\rho < 1$ .

Consider a queueing system with exponential inter-arrival times, exponential service times, and a single server. It is denoted as M/M/1, where the letter M stands for the Markov discipline. The traffic intensity for the M/M/1 system is simply

$$\rho = \lambda/\mu,$$

where  $\lambda$  is the arrival rate and  $\mu$  is the service rate. In this case, the process  $\{Q(t)\}_{t \geq 0}$  is a birth-death process with the birth rate  $\lambda$  and the death rate  $\mu$  for the positive states, and no deaths at state zero.

**Theorem 7.25** *The probabilities  $p_n(t) = \mathbb{P}(Q(t) = n)$  for a M/M/1 queue satisfy the Kolmogorov forward equations*

$$\begin{cases} p'_n(t) &= -(\lambda + \mu)p_n(t) + \lambda p_{n-1}(t) + \mu p_{n+1}(t) & \text{for } n \geq 1, \\ p'_0(t) &= -\lambda p_0(t) + \mu p_1(t), \end{cases}$$

subject to the boundary condition  $p_n(0) = 1_{\{n=0\}}$ .

PROOF. For  $n \geq 1$  and a small positive  $\delta$ , put

$$R_n(t, \delta) = \mathbb{P}(Q(t + \delta) = n; \text{at least two events during } (t, t + \delta]).$$

Observe that

$$R_n(t, \delta) \leq \mathbb{P}(\text{at least two events during } (t, t + \delta]) \leq C \cdot \delta^2.$$

For example, using Poisson distribution for the number of births, we get

$$\begin{aligned} \mathbb{P}(\text{at least two births during } (t, t + \delta]) &= \sum_{k \geq 2} \frac{(\lambda\delta)^k}{k!} e^{-\lambda\delta} \\ &\leq (\lambda\delta)^2 \sum_{j \geq 0} \frac{(\lambda\delta)^j}{j!} e^{-\lambda\delta} = (\lambda\delta)^2. \end{aligned}$$

It follows, that for  $n \geq 1$ ,

$$\begin{aligned} p_n(t + \delta) &= \mathbb{P}(Q(t + \delta) = n; \text{at most one event during } (t, t + \delta]) + R_n(t, \delta) \\ &= p_n(t)\mathbb{P}(\text{no events during } (t, t + \delta]) + p_{n-1}(t)\mathbb{P}(\text{single birth during } (t, t + \delta]) \\ &\quad + p_{n+1}(t)\mathbb{P}(\text{single death during } (t, t + \delta]) + o(\delta) \\ &= p_n(t)(1 - (\lambda + \mu)\delta) + p_{n-1}(t)\lambda\delta + p_{n+1}(t)\mu\delta + o(\delta). \end{aligned}$$

This yields the first stated equation. Similarly, to get the second equation observe that

$$\begin{aligned} p_0(t + \delta) &= p_0(t)\mathbb{P}(\text{no birth during } (t, t + \delta]) + p_1(t)\mathbb{P}(\text{single death during } (t, t + \delta]) + o(\delta) \\ &= p_0(t)(1 - \lambda\delta) + p_1(t)\mu\delta + o(\delta). \end{aligned}$$



**Exercise 7.26** In terms of the Laplace transforms  $\hat{p}_n(\theta) := \int_0^\infty e^{-\theta t} p_n(t) dt$  we obtain

$$\begin{cases} \mu \hat{p}_{n+1}(\theta) - (\lambda + \mu + \theta) \hat{p}_n(\theta) + \lambda \hat{p}_{n-1}(\theta) &= 0, & \text{for } n \geq 1, \\ \mu \hat{p}_1(\theta) - (\lambda + \theta) \hat{p}_0(\theta) &= 1. \end{cases}$$

Solve this system of linear equations.

SOLUTION.

$$\hat{p}_n(\theta) = \theta^{-1} (1 - \alpha(\theta)) \alpha(\theta)^n, \quad \alpha(\theta) := \frac{(\lambda + \mu + \theta) - \sqrt{(\lambda + \mu + \theta)^2 - 4\lambda\mu}}{2\mu}.$$

**Theorem 7.27 Stationarity.** Consider a M/M/1 queue with traffic intensity  $\rho$ . As  $t \rightarrow \infty$ ,

$$P(Q(t) = n) \rightarrow \pi_n := \begin{cases} (1 - \rho)\rho^n & \text{if } \rho < 1, \\ 0 & \text{if } \rho \geq 1, \end{cases} \quad n = 0, 1, 2, \dots$$

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  $\theta \rightarrow 0$

$$\alpha(\theta) = \frac{(\lambda + \mu + \theta) - \sqrt{(\lambda - \mu)^2 + 2(\lambda + \mu)\theta + \theta^2}}{2\mu} \rightarrow \frac{(\lambda + \mu + \theta) - |\lambda - \mu|}{2\mu} = \begin{cases} \rho & \text{if } \lambda < \mu, \\ 1 & \text{if } \lambda \geq \mu. \end{cases}$$

Thus

$$\theta \hat{p}_n(\theta) = (1 - \alpha(\theta)) \alpha(\theta)^n \rightarrow \begin{cases} (1 - \rho)\rho^n & \text{if } \rho < 1, \\ 0 & \text{if } \rho \geq 1. \end{cases}$$

On the other hand, the process  $Q(t)$  is an irreducible Markov chain and by the ergodic theorem there are limits  $p_n(t) \rightarrow \pi_n$  as  $t \rightarrow \infty$ . Therefore, the statement follows from

$$\theta \hat{p}_n(\theta) = \int_0^\infty e^{-u} p_n(u/\theta) du \rightarrow \pi_n, \quad \theta \rightarrow 0.$$

**Exercise 7.28** Consider the M/M/1 queue with  $\rho < 1$ . In the stationary regime  $p_n(t) \equiv \pi_n$ , the average number of customers in the line  $L$  is the mean of a geometric distribution, so that

$$L = \frac{\rho}{1 - \rho} = \frac{\lambda}{\mu - \lambda}.$$

According to Little's Law the time  $V$  spent by a customer in a stationary queue has mean

$$\nu = \frac{L}{\lambda} = \frac{1}{\mu - \lambda}.$$

Verify this by showing that  $V$  has an exponential distribution with parameter  $\mu - \lambda$ .

Hint. Show that  $V = S_1 + \dots + S_{1+Q}$ , where  $S_i$  are independent  $\text{Exp}(\mu)$  random variables and  $Q$  is geometric with parameter  $1 - \rho$ , and then compute the Laplace transform  $E(e^{-uV})$ .

## 7.9 M/G/1 queues

In the M/G/1 queueing system customers arrive according to a Poisson process with intensity  $\lambda$  and the service times  $S_i$  have a fixed but unspecified distribution (G for General). In this case the traffic intensity is given by

$$\rho = \lambda E(S).$$

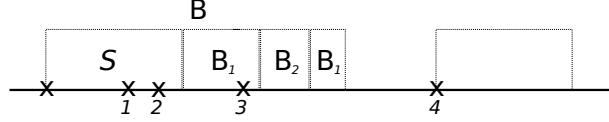


Figure 3: A typical busy period in a M/G/1 queue. Here  $Z = 2$ .

**Theorem 7.29** Define a typical busy period of the server as  $B = \inf\{t > 0 : Q(t + X_1) = 0\}$ , where  $X_1$  is the arrival time of the first customer.

(i) If  $\rho \leq 1$ , then  $P(B < \infty) = 1$ , and if  $\rho > 1$ , then  $P(B < \infty) < 1$ .

(ii) The Laplace transform  $\phi(u) = E(e^{-uB})$  satisfies the functional equation

$$\phi(u) = \Psi(u + \lambda - \lambda\phi(u)), \quad \text{where } \Psi(x) := E(e^{-xS}).$$

PROOF. We use an embedded branching process construction (branching processes are introduced in 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 the value of the service time  $S = t$ , the offspring number  $Z$  for a single customer has conditional  $\text{Pois}(\lambda t)$  distribution. Therefore, the conditional probability generating function of  $Z$  is

$$E(u^Z | S) = \sum_{j=0}^{\infty} u^j \frac{(\lambda S)^j}{j!} e^{-\lambda S} = e^{S\lambda(u-1)}.$$

This yields the following expression for the probability generating function of the offspring number

$$h(u) := E(u^Z) = E(e^{S\lambda(u-1)}),$$

or

$$h(u) = \Psi(\lambda(1 - u)).$$

The mean offspring number is

$$E(Z) = h'(1) = -\lambda\Psi'(0) = \lambda E(S) = \rho.$$

Observe that the event  $(B < \infty)$  is equivalent to the extinction of the embedded branching process, and the first assertion follows.

The functional equation is derived from the representation  $B = S + B_1 + \dots + B_Z$ , where  $B_i$  are iid busy times of the offspring customers and  $Z$  has the generating function  $h(u)$ . Indeed, by independence

$$E(e^{-uB_1} \dots e^{-uB_Z} | S) = E(\phi(u)^Z | S) = e^{(\lambda\phi(u) - \lambda)S}.$$

This implies

$$\phi(u) = E(e^{-uS} e^{-uB_1} \dots e^{-uB_Z}) = E(e^{-uS} e^{(\lambda\phi(u) - \lambda)S}) = \Psi(u + \lambda - \lambda\phi(u)).$$

**Exercise 7.30** 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 7.31 Stationarity.** As  $t \rightarrow \infty$ , for every  $n \geq 0$ ,

$$P(Q(t) = n) \rightarrow \begin{cases} \pi_n & \text{if } \rho < 1, \text{ where } \sum_{j=0}^{\infty} \pi_j u^j = (1 - \rho)(1 - u) \frac{h(u)}{h(u) - u}, \\ 0 & \text{if } \rho \geq 1. \end{cases}$$

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  $\{0, 1, 2, \dots\}$  and transition probabilities

$$\begin{pmatrix} \delta_0 & \delta_1 & \delta_2 & \dots \\ \delta_0 & \delta_1 & \delta_2 & \dots \\ 0 & \delta_0 & \delta_1 & \dots \\ 0 & 0 & \delta_0 & \dots \\ \dots & \dots & \dots & \dots \end{pmatrix}, \quad \delta_j = P(Z = j) = E\left(\frac{(\lambda S)^j}{j!} e^{-\lambda S}\right).$$

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  $\psi(u) = E(u^D)$ , we have

$$\psi(u) = E(u^{D-1_{\{D>0\}}})h(u) = \left(\frac{\psi(u) - \psi(0)}{u} + \psi(0)\right)h(u).$$

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

$$E(u^D) = (1-\rho)(1-u)\frac{h(u)}{h(u)-u}. \quad (*)$$

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 derive (not shown here) the stated convergence for the queue size distribution.

**Theorem 7.32** *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  $P(W = 0) = 1 - \rho$  and*

$$E(e^{-uW}) = \frac{(1-\rho)u}{u-\lambda + \lambda E(e^{-uS})}.$$

PROOF. Suppose that a customer waits for a period of length  $W$  before it 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 to (\*). On the other hand,  $D$  conditionally on  $(W, S)$  has a Poisson distribution with parameter  $\lambda(W + S)$ . Therefore,

$$(1-\rho)(1-s)\frac{h(s)}{h(s)-s} = E(s^D) = E(e^{\lambda(W+S)(s-1)}) = E(e^{\lambda W(s-1)})h(s).$$

Thus, with  $u = \lambda(1-s)$ ,

$$E(e^{-uW}) = E(e^{\lambda(s-1)W}) = \frac{(1-\rho)(s-1)}{s - E(e^{\lambda(s-1)S})} = \frac{(1-\rho)u}{u-\lambda + \lambda E(e^{-uS})}.$$

**Exercise 7.33** Using the previous theorem show that for the M/M/1 queue we have

$$E(e^{-uW}) = 1 - \frac{\rho u}{\mu - \lambda + u}.$$

Find  $P(W = 0)$ . Show that conditionally on  $W > 0$ , this distribution is exponential.

**Theorem 7.34** *Heavy traffic. Consider an M/G/1 queue with  $M = M(\lambda)$  and  $G = D(d)$ , meaning that the service time  $S$  is deterministic in that  $P(S = d) = 1$  for some  $d > 0$ . In this case, the traffic intensity is  $\rho = \lambda d$ . For  $\rho < 1$ , let  $Q_\rho$  be a r.v. with the equilibrium queue length distribution. Then, as  $\rho \nearrow 1$ , the distribution of  $(1-\rho)Q_\rho$  converges to the exponential distribution with parameter 2.*

PROOF. Applying Theorem 7.31 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 7.35** For  $\rho = 0.99$  find  $\mathbb{P}(Q_\rho > 60)$  using the previous theorem. Answer 30%.

## 7.10 G/M/1 queues

Customers' arrival times form a renewal process with inter-arrival times  $(X_n)$  having a general distribution  $G$ , and the service times are exponentially distributed with parameter  $\mu$ . The traffic intensity is

$$\rho = \frac{1}{\mu \mathbb{E}(X)}.$$

**An embedded 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 this customer 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 the state space  $\{0, 1, 2, \dots\}$  and transition probabilities

$$\Pi_A = \begin{pmatrix} 1 - \alpha_0 & \alpha_0 & 0 & 0 & \dots \\ 1 - \alpha_0 - \alpha_1 & \alpha_1 & \alpha_0 & 0 & \dots \\ 1 - \alpha_0 - \alpha_1 - \alpha_2 & \alpha_2 & \alpha_1 & \alpha_0 & \dots \\ \dots & \dots & \dots & \dots & \dots \end{pmatrix}, \quad \alpha_j = \mathbb{E}\left(\frac{(\mu X)^j}{j!} e^{-\mu X}\right).$$

**Theorem 7.36** If  $\rho < 1$ , then the  $A$ -chain is ergodic with a unique stationary distribution

$$\pi_j = (1 - \eta)\eta^j, \quad j = 0, 1, \dots,$$

where  $\eta$  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.

PARTIAL PROOF. Let  $\rho < 1$  and put  $\bar{\pi} = (\pi_0, \pi_1, \dots)$ . To see that  $\eta \in (0, 1)$  observe that  $a(x) = \mathbb{E}(e^{-X\mu(1-x)})$  has a convex graph. Since  $a(0) > 0$ ,  $a(1) = 1$ , and  $a'(1) = \rho^{-1} > 1$ , we conclude that there is an intersection  $\eta \in (0, 1)$  of the lines  $y = x$  and  $y = a(x)$ . The vector  $\bar{\pi}$  gives the stationary distribution of the  $A$ -chain, since we have  $\bar{\pi} = \bar{\pi}\Pi_A$  due to the following equalities

$$\sum_{j=0}^{\infty} \alpha_j \eta^j = \eta, \quad \pi_j + \pi_{j+1} + \dots = \eta^j.$$

**Example 7.37** 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

$$p_t(a) := \mathbb{P}(V_n(t) = v | A_n = a) = \begin{cases} \frac{(\mu t)^v}{v!} e^{-\mu t} & \text{if } v \leq a, \\ \sum_{i \geq a+1} \frac{(\mu t)^i}{i!} e^{-\mu t} & \text{if } v = a + 1. \end{cases}$$

Using the stationary distribution for  $A_n$  we get

$$P(Q(n+t) = k) = \sum_{j=k-1}^{\infty} (1-\eta)\eta^j p_t(j).$$

This example demonstrates that unlike  $(D_n)$  in the case of M/G/1, the stationary distribution of  $(A_n)$  need not to be the limiting distribution of the queue size  $Q(t)$ .

**Theorem 7.38** *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  $x \geq 0$*

$$P(W > x) = \eta e^{-\mu(1-\eta)x}.$$

PROOF. If  $A_n > 0$ , then the waiting time of the  $n$ -th customer is

$$W_n = S_1^* + S_2 + S_3 + \dots + S_{A_n},$$

where  $S_1^*$  is the residual service time of the customer under service, and  $S_2, S_3, \dots, S_{A_n}$  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

$$E(e^{-uW}) = E((E(e^{-uS}))^{A_n}) = E\left(\left(\frac{\mu}{\mu+u}\right)^{A_n}\right) = \frac{(\mu+u)(1-\eta)}{\mu+u-\mu\eta} = (1-\eta) + \eta \frac{\mu(1-\eta)}{\mu(1-\eta)+u}.$$

**Exercise 7.39** Using the previous theorem, show that for the M/M/1 queue we have

$$E(e^{-uW}) = 1 - \rho + \rho \frac{\mu - \lambda}{\mu - \lambda + u}.$$

## 7.11 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. We exclude the trivial case when  $P(S = d) = P(X = d) = 1$  for a positive constant  $d$ .

**Lemma 7.40** *Lindley 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  $X_{n+1} \leq W_n + S_n$ , 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 7.41** *Note that  $U_n = S_n - X_{n+1}$  is a collection of iid r.v. Denote by  $G(x) = P(U \leq x)$  their common distribution function. Let  $F_n(x) = P(W_n \leq x)$ . Then for  $x \geq 0$*

$$F_{n+1}(x) = \int_{-\infty}^x F_n(x-y) dG(y).$$

*There exists a limit  $F(x) = \lim_{n \rightarrow \infty} F_n(x)$  which satisfies the Wiener-Hopf equation*

$$F(x) = \int_{-\infty}^x F(x-y) dG(y) = E(F(x-U); U \leq x).$$

*In terms of an embedded random walk  $\Sigma_n = U_1 + \dots + U_n$ ,*

$$F(x) = P(\Sigma_n \leq x \text{ for all } n), \quad x \geq 0.$$

PROOF. If  $x \geq 0$  then due to the Lindley equation and independence between  $W_n$  and  $U_n = S_n - X_{n+1}$

$$P(W_{n+1} \leq x) = P(W_n + U_n \leq x) = \int_{-\infty}^x P(W_n \leq x - y) dG(y)$$

and the first part is proved. We claim that

$$F_{n+1}(x) \leq F_n(x) \text{ for all } x \geq 0 \text{ and } n \geq 1. \quad (*)$$

If  $(*)$  holds, then the second result follows immediately. We prove  $(*)$  by induction. Observe that  $F_2(x) \leq 1 = F_1(x)$ , and suppose that  $(*)$  holds for  $n = k - 1$ . Then

$$F_{k+1}(x) - F_k(x) = \int_{-\infty}^x (F_k(x - y) - F_{k-1}(x - y)) dG(y) \leq 0.$$

Turning to the embedded random walk observe that Lemma 7.40 implies

$$W_{n+1} = \max\{0, U_1 + \dots + U_n, U_2 + \dots + U_n, \dots, U_{n-1} + U_n, U_n\}.$$

Draw a picture with trajectories of  $(W_j)_{j=0}^{n+1}$  and  $(U_i + \dots + U_j)_{j=i}^n$  for all  $i = 1, \dots, n$ , to be persuaded. Since

$$\max\{0, U_n, U_n + U_{n-1}, \dots, U_n + U_{n-1} + \dots + U_1\} \stackrel{d}{=} \max\{\Sigma_0, \dots, \Sigma_n\},$$

we conclude that for  $x \geq 0$ ,

$$F_{n+1}(x) = P(\Sigma_1 \leq x, \dots, \Sigma_n \leq x),$$

and thus  $F$  is the distribution of  $\max\{\Sigma_0, \Sigma_1, \dots\}$ .

**Theorem 7.42** *If  $\rho < 1$ , then  $F$  from the previous theorem is a non-defective distribution function. On the other hand, if  $\rho \geq 1$ , then  $F(x) = 0$  for all  $x$ .*

PROOF. Note that  $E(U) = E(S) - E(X)$ , and  $E(U) < 0$  is equivalent to  $\rho < 1$ . If  $E(U) < 0$ , then

$$P(\Sigma_n > 0 \text{ for infinitely many } n) = P(n^{-1}\Sigma_n > 0 \text{ i.o.}) = P(n^{-1}\Sigma_n - E(U) > |E(U)| \text{ i.o.}) = 0$$

due to the LLN. Thus  $\max\{\Sigma_0, \Sigma_1, \dots\}$  is either zero or the maximum of only finitely many positive terms, and  $F$  is a non-defective distribution.

Next suppose that  $E(U) > 0$  and pick any  $x > 0$ . For  $n \geq 2x/E(U)$

$$P(\Sigma_n \geq x) = P(n^{-1}\Sigma_n - E(U) \geq n^{-1}x - E(U)) \geq P(n^{-1}\Sigma_n - E(U) \geq -E(U)/2) = P(n^{-1}\Sigma_n \geq E(U)/2).$$

Since  $1 - F(x) \geq P(\Sigma_n \geq x)$ , the weak LLN implies  $F(x) = 0$ . In the case when  $E(U) = 0$  we need a more precise measure of the fluctuations of  $\Sigma_n$ . According to the law of the iterated logarithm the fluctuations of  $\Sigma_n$  are of order  $O(\sqrt{n \log \log n})$  in both positive and negative directions with probability 1, and so for any given  $x \geq 0$ ,

$$1 - F(x) = P(\Sigma_n > x \text{ for some } n) = 1.$$

**Definition 7.43** 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  $L(1), L(2), \dots$  are called ladder points of the random walk  $\Sigma$ , these are the times when the random walk  $\Sigma$  reaches its new maximal values.

**Lemma 7.44** *Let  $\rho < 1$  and put*

$$\eta = P(\Sigma_n > 0 \text{ for some } n) = 1 - F(0).$$

*The total number  $\Lambda$  of ladder points has a geometric distribution*

$$P(\Lambda = k) = (1 - \eta)\eta^k, \quad k = 0, 1, 2, \dots$$

PROOF. The underlying random walk has a special Markov property: going from one record height to another is a sequence of Benoulli trials with the probability of success  $\eta$ .

**Theorem 7.45** *If  $\rho < 1$ , the equilibrium waiting time  $W \stackrel{d}{=} \max\{\Sigma_0, \Sigma_1, \dots\}$  has the Laplace transform*

$$\mathbb{E}(e^{-uW}) = \frac{1 - \eta}{1 - \eta \mathbb{E}(e^{-uY})},$$

where  $Y = \Sigma_{L(1)}$  is the first ladder height of the embedded random walk.

PROOF. In terms of the consecutive increments of the record heights  $Y_j = \Sigma_{L(j)} - \Sigma_{L(j-1)}$ , which are iid copies of  $Y = Y_1$ , we have

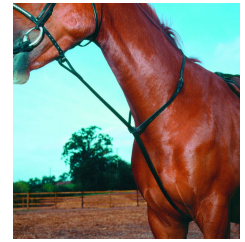
$$\max\{\Sigma_0, \Sigma_1, \dots\} = \Sigma_{L(\Lambda)} = Y_1 + \dots + Y_\Lambda.$$

Therefore,

$$\mathbb{E}(e^{-uW}) = \phi(\mathbb{E}e^{-uY}), \quad \phi(s) = \mathbb{E}(s^\Lambda) = \frac{1 - \eta}{1 - \eta s}.$$

## 8 Martingales

The origin of the term "martingale" →



### 8.1 Definitions and examples

**Definition 8.1** The sequence  $(Y_n)_{n \geq 1}$  is a martingale with respect to the sequence  $(X_n)_{n \geq 1}$ , if  $Y_n = \psi_n(X_1, \dots, X_n)$  and for all  $n \geq 1$ ,

$$\mathbb{E}(Y_{n+1} | X_1, \dots, X_n) = Y_n.$$

We sometimes just say that  $(Y_n)$  is a martingale.

**Example 8.2** After tossing a fair coin a gambler wins for heads and loses for tails. The game starts by a unit bet and the gambler doubles the bet after each loss and the game stops after the first win. Let  $Y_n$  be the gain of the gambler after the  $n$ -th tossing.

The coin tossing outcomes can be modelled by a sequence  $X_1, X_2, \dots$  of independent random variables with the same symmetric Bernoulli distribution. Define  $N$  as the first  $n$  such that  $X_n = 1$ . Then

$$Y_n = 1_{\{n \geq N\}} - (2^n - 1)1_{\{n < N\}} = 1 - 2^n \cdot 1_{\{n < N\}}, \quad n \geq 0.$$

Since  $\mathbb{P}(N > n) = 2^{-n}$ , we find  $\mathbb{E}Y_n = 0$  confirming that the game is fair.

The martingale property of this betting strategy is established as follows

$$\begin{aligned} \mathbb{E}(Y_{n+1} | X_1, \dots, X_n) &= 1 - 2^{n+1} \mathbb{P}(N > n + 1 | X_1, \dots, X_n) \\ &= 1 - 2^{n+1} \mathbb{P}(N > n, X_{n+1} = 0 | X_1, \dots, X_n) \\ &= 1 - 2^n \cdot 1_{\{n < N\}} = Y_n. \end{aligned}$$

Notice that

$$\mathbb{E}(Y_{N-1}) = \mathbb{E}(1 - 2^{N-1}) = 1 - \sum_{n=1}^{\infty} 2^{n-1} 2^{-n} = -\infty,$$

which means that you need an infinite capital to win consistently. But then gambling becomes pointless.

**Definition 8.3** 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  $n$ . Consider an adapted sequence with  $\mathbb{E}(|Y_n|) < \infty$  for all  $n \geq 0$ . If for all  $n \geq 0$ ,

- martingale property  $\mathbb{E}(Y_{n+1} | \mathcal{F}_n) = Y_n$ ,
- submartingale property  $\mathbb{E}(Y_{n+1} | \mathcal{F}_n) \geq Y_n$ ,
- supermartingale property  $\mathbb{E}(Y_{n+1} | \mathcal{F}_n) \leq Y_n$ ,

then  $(Y_n, \mathcal{F}_n)$  is called a martingale, submartingale, or supermartingale.

**Exercise 8.4** Consider the sequence of means  $m_n = E(Y_n)$ . 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 8.5** Consider a simple random walk  $S_n = S_0 + X_1 + \dots + X_n$  with

$$P(X_i = 1) = p, \quad P(X_i = -1) = 1 - p =: q, \quad S_0 = k.$$

The centred sequence  $S_n - n(p - q)$  is a martingale:

$$E(S_{n+1} - (n+1)(p - q) | X_1, \dots, X_n) = S_n + E(X_{n+1}) - (n+1)(p - q) = S_n - n(p - q).$$

Another martingale is  $Y_n = (q/p)^{S_n}$ :

$$E(Y_{n+1} | X_1, \dots, X_n) = p(q/p)^{S_n+1} + q(q/p)^{S_n-1} = (q/p)^{S_n} = Y_n$$

with  $E(Y_n) = E(Y_0) = (q/p)^k$ . It is called De Moivre martingale.

**Example 8.6** Let  $S_n = X_1 + \dots + X_n$ , where  $X_i$  are iid r.v. with zero means and finite variances  $\sigma^2$ . Then  $S_n^2 - n\sigma^2$  is a martingale

$$E(S_{n+1}^2 - (n+1)\sigma^2 | X_1, \dots, X_n) = S_n^2 + 2S_n E(X_{n+1}) + E(X_{n+1}^2) - (n+1)\sigma^2 = S_n^2 - n\sigma^2.$$

**Example 8.7** Stopped de Moivre martingale. Consider the same simple random walk and let  $T$  be a stopping time of the random walk. Denote  $Z_n = Y_{T \wedge n}$ , where  $(Y_n)$  is the de Moivre martingale. It is easy to see that  $Z_n$  is also a martingale:

$$\begin{aligned} E(Z_{n+1} | X_1, \dots, X_n) &= E(Y_{n+1} 1_{\{T > n\}} + \sum_{j \leq n} Y_j 1_{\{T=j\}} | X_1, \dots, X_n) \\ &= Y_n 1_{\{T > n\}} + \sum_{j \leq n} Y_j 1_{\{T=j\}} = Z_n. \end{aligned}$$

**Example 8.8** Branching processes (recall Section 5.5). Let  $Z_n$  be a branching process with  $Z_0 = 1$  and the mean offspring number  $\mu$ . 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  $h(s) = E(s^{X_n})$  is the generating function of the offspring number. The process  $V_n = \eta^{Z_n}$  is a martingale

$$E(V_{n+1} | Z_1, \dots, Z_n) = E(\eta^{X_1 + \dots + X_{Z_n}} | Z_1, \dots, Z_n) = h(\eta)^{Z_n} = V_n.$$

## 8.2 Convergence in $L^2$

**Lemma 8.9** If  $(Y_n)$  is a martingale with  $E(Y_n^2) < \infty$ , then  $Y_{n+1} - Y_n$  and  $Y_n$  are uncorrelated. It follows that  $(Y_n^2)$  is a submartingale.

PROOF. The first assertion comes from

$$E(Y_n(Y_{n+1} - Y_n) | \mathcal{F}_n) = Y_n(E(Y_{n+1} | \mathcal{F}_n) - Y_n) = 0.$$

The second claim is derived as follows

$$\begin{aligned} E(Y_{n+1}^2 | \mathcal{F}_n) &= E((Y_{n+1} - Y_n)^2 + 2Y_n(Y_{n+1} - Y_n) + Y_n^2 | \mathcal{F}_n) \\ &= E((Y_{n+1} - Y_n)^2 | \mathcal{F}_n) + Y_n^2 \geq Y_n^2. \end{aligned}$$

Notice that

$$E(Y_{n+1}^2) = E(Y_n^2) + E((Y_{n+1} - Y_n)^2)$$

so that  $E(Y_n^2)$  is non-decreasing and there always exists a finite or infinite limit  $\lim_{n \rightarrow \infty} E(Y_n^2)$ .



**Exercise 8.10** Let  $J(x)$  be a convex function. If  $(Y_n)$  is a martingale with  $EJ(Y_n) < \infty$ , then  $J(Y_n)$  is a submartingale. (Hint: use Jensen inequality.)

**Lemma 8.11** *Doob-Kolmogorov inequality.* If  $(Y_n)$  is a martingale with  $E(Y_n^2) < \infty$ , then for any  $\epsilon > 0$

$$P(\max_{1 \leq i \leq n} |Y_i| \geq \epsilon) \leq \frac{E(Y_n^2)}{\epsilon^2}.$$

PROOF. Let  $B_1 = \{|Y_1| \geq \epsilon\}$  and  $B_k = \{|Y_1| < \epsilon, \dots, |Y_{k-1}| < \epsilon, |Y_k| \geq \epsilon\}$ . Then using a submartingale property (in the second inequality) we get

$$E(Y_n^2) \geq \sum_{i=1}^n E(Y_n^2 1_{B_i}) \geq \sum_{i=1}^n E(Y_i^2 1_{B_i}) \geq \epsilon^2 \sum_{i=1}^n P(B_i) = \epsilon^2 P(\max_{1 \leq i \leq n} |Y_i| \geq \epsilon).$$

**Theorem 8.12** If  $(Y_n)$  is a martingale with

$$\sup_n E(Y_n^2) \leq M$$

for some finite  $M$ , then there exists a random variable  $Y$  such that  $Y_n \rightarrow Y$  a.s. and in mean square as  $n \rightarrow \infty$ .

PROOF. Step 1. For

$$A_m(\epsilon) = \bigcup_{i \geq 1} \{|Y_{m+i} - Y_m| \geq \epsilon\}$$

we will show that

$$P(A_m(\epsilon)) \rightarrow 0, \quad m \rightarrow \infty \text{ for any } \epsilon > 0.$$

For a given  $m$ , put  $S_n = Y_{m+n} - Y_m$ . It is also a martingale, since

$$E(S_{n+1} | S_1, \dots, S_n) = E(E(S_{n+1} | \mathcal{F}_{m+n}) | S_1, \dots, S_n) = E(S_n | S_1, \dots, S_n) = S_n.$$

Apply the Doob-Kolmogorov inequality to this martingale to find that

$$P(|Y_{m+i} - Y_m| \geq \epsilon \text{ for some } i \in [1, n]) \leq \epsilon^{-2} E((Y_{m+n} - Y_m)^2) = \epsilon^{-2} (E(Y_{m+n}^2) - E(Y_m^2)).$$

Letting  $n \rightarrow \infty$  we obtain

$$P(A_m(\epsilon)) \leq \epsilon^{-2} (M - E(Y_m^2))$$

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,

$$P\left(\bigcup_{\epsilon > 0} \bigcap_{m \geq 1} A_m(\epsilon)\right) = \lim_{\epsilon \rightarrow 0} P\left(\bigcap_{m \geq 1} A_m(\epsilon)\right) \leq \lim_{\epsilon \rightarrow 0} \lim_{m \rightarrow \infty} P(A_m(\epsilon)) = 0.$$

Therefore, the sequence  $(Y_n)$  is a.s. Cauchy convergent:

$$P\left(\bigcap_{\epsilon > 0} \bigcup_{m \geq 1} A_m^c(\epsilon)\right) = 1,$$

which implies the existence of  $Y$  such that  $Y_n \rightarrow Y$  a.s.

Step 3. Finally, we prove the convergence in mean square using the Fatou lemma

$$\begin{aligned} E((Y_n - Y)^2) &= E(\liminf_{m \rightarrow \infty} (Y_n - Y_m)^2) \leq \liminf_{m \rightarrow \infty} E((Y_n - Y_m)^2) \\ &= \liminf_{m \rightarrow \infty} (E(Y_m^2) - E(Y_n^2)) = M - E(Y_n^2) \rightarrow 0, \quad n \rightarrow \infty. \end{aligned}$$

**Example 8.13** Branching processes (see Section 5.5). Let  $Z_n$  be a branching process with  $Z_0 = 1$  and an offspring distribution having mean  $\mu$  and variance  $\sigma^2$ . Since  $E(Z_{n+1} | Z_n) = \mu Z_n$ , the ratio  $W_n = \mu^{-n} Z_n$  is a martingale with

$$E(W_n^2) = 1 + (\sigma/\mu)^2 (1 + \mu^{-1} + \dots + \mu^{-n+1}).$$

In the supercritical case,  $\mu > 1$ , we have  $E(W_n^2) \rightarrow 1 + \frac{\sigma^2}{\mu(\mu-1)}$ , and there is a r.v.  $W$  such that  $W_n \rightarrow W$  a.s. and in  $L^2$ . The Laplace transform of the limit  $\phi(\theta) = E(e^{-\theta W})$  satisfies the functional equation

$$\phi(\mu\theta) = h(\phi(\theta)).$$

### 8.3 Doob decomposition

**Definition 8.14** 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  $P(S_n \leq S_{n+1}) = 1$  for all  $n \geq 0$ .

**Theorem 8.15** *Doob decomposition.* A submartingale  $(Y_n, \mathcal{F}_n)$  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  $n \geq 0$

$$M_{n+1} - M_n = Y_{n+1} - E(Y_{n+1}|\mathcal{F}_n), \quad S_{n+1} - S_n = E(Y_{n+1}|\mathcal{F}_n) - Y_n.$$

To see uniqueness, suppose another such decomposition  $Y_n = M'_n + S'_n$  exists. 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 8.16** 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 8.17** Such transforms with  $S_n \geq 0$  are usually interpreted as gambling systems, with  $(Y_n)$  being a supermartingale defined as the capital of a gambler wagering a unit stake at each round. Optional skipping is one such strategy: the gambler either wagers a unit stake,  $S_n = 1$ , or skips the round,  $S_n = 0$ . The following result says that there are no sure ways to beat the casino.

**Theorem 8.18** Let  $(Z_n)$  be the transform of  $(Y_n)$  by  $(S_n)$  such that  $E|Z_n| < \infty$  for all  $n$ . Then

(i) If  $(Y_n)$  is a martingale, then  $(Z_n)$  is a martingale.

(ii) If  $(Y_n)$  is a submartingale (supermartingale) and in addition  $S_n \geq 0$  for all  $n$ , then  $(Z_n)$  is a submartingale (supermartingale).

PROOF. All the assertions follow immediately from

$$E(Z_{n+1}|\mathcal{F}_n) - Z_n = E(Z_{n+1} - Z_n|\mathcal{F}_n) = E(S_{n+1}(Y_{n+1} - Y_n)|\mathcal{F}_n) = S_{n+1}(E(Y_{n+1}|\mathcal{F}_n) - Y_n).$$

**Example 8.19** Optional stopping. Consider a gambler wagering 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 8.20** Optional starting. If a gambler does not play until the  $(T+1)$ -th round, where  $T$  is a stopping time, then

$$Z_n = Y_0 + (Y_n - Y_T)1_{\{n \geq T+1\}} = Y_0 + \sum_{i=T+1}^n (Y_i - Y_{i-1}).$$

The corresponding predictable sequence is  $S_n = 1_{\{T \leq n-1\}}$ .

### 8.4 Hoeffding inequality

**Definition 8.21** 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

$$E|D_n| < \infty, \quad E(D_{n+1}|\mathcal{F}_n) = 0, \quad Y_n = Y_0 + D_1 + \dots + D_n.$$

Sums of martingale differences extend the classical setting of summing independent variables.

**Theorem 8.22** Let  $(Y_n, \mathcal{F}_n)$  be a martingale with bounded martingale differences:  $P(|D_n| \leq K_n) = 1$  for a sequence of real numbers  $K_n$ . Then for any  $x > 0$

$$P(|Y_n - Y_0| \geq x) \leq 2 \exp\left(-\frac{x^2}{2(K_1^2 + \dots + K_n^2)}\right).$$

PROOF. Let  $\theta > 0$ . Later in the proof we put  $\theta = x / \sum_{i=1}^n K_i^2$ .

Step 1. The function  $e^{\theta x}$  is convex over  $x \in [-1, 1]$ , 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 zero-mean random variable, such that  $P(|D| \leq 1) = 1$ , then

$$E(e^{\theta D}) \leq \frac{e^{-\theta} + e^{\theta}}{2} + \frac{e^{\theta} - e^{-\theta}}{2}E(D) = \frac{e^{-\theta} + e^{\theta}}{2} = \sum_{k=0}^{\infty} \frac{\theta^{2k}}{(2k)!} < \sum_{k=0}^{\infty} \frac{\theta^{2k}}{2^k(k)!} = e^{\theta^2/2}.$$

Step 2. Applying the previous step to the scaled martingale differences  $D'_n = D_n/K_n$ , we obtain

$$E(e^{\theta(Y_n - Y_0)} | \mathcal{F}_{n-1}) = e^{\theta(Y_{n-1} - Y_0)} E(e^{\theta D_n} | \mathcal{F}_{n-1}) = e^{\theta(Y_{n-1} - Y_0)} E(e^{\theta K_n D'_n} | \mathcal{F}_{n-1}) \leq e^{\theta(Y_{n-1} - Y_0)} e^{\theta^2 K_n^2 / 2}.$$

Take expectations and iterate to find

$$E(e^{\theta(Y_n - Y_0)}) \leq E(e^{\theta(Y_{n-1} - Y_0)}) e^{\theta^2 K_n^2 / 2} \leq \exp\left(\frac{\theta^2}{2} \sum_{i=1}^n K_i^2\right).$$

Step 3. Due to the Markov inequality we have for any  $x > 0$

$$P(Y_n - Y_0 \geq x) = P(e^{\theta(Y_n - Y_0)} \geq e^{\theta x}) \leq e^{-\theta x} E(e^{\theta(Y_n - Y_0)}) \leq \exp\left(-\theta x + \frac{\theta^2}{2} \sum_{i=1}^n K_i^2\right).$$

Set  $\theta = x / \sum_{i=1}^n K_i^2$  to minimize the exponent. Then

$$P(Y_n - Y_0 \geq x) \leq \exp\left(-\frac{x^2}{2(K_1^2 + \dots + K_n^2)}\right).$$

Since  $(-Y_n)$  is also a martingale, we get

$$P(Y_n - Y_0 \leq -x) = P(-Y_n + Y_0 \geq x) \leq \exp\left(-\frac{x^2}{2(K_1^2 + \dots + K_n^2)}\right).$$

**Example 8.23** An application to large deviations. Let  $X_n$  be iid  $\text{Ber}(p)$  random variables. If  $S_n = X_1 + \dots + X_n$ , 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 inequality for any  $x > 0$

$$P\left(\left|\frac{S_n - np}{\sqrt{n}}\right| \geq x\right) = P(|S_n - np| \geq x\sqrt{n}) \leq 2 \exp\left(-\frac{x^2}{2(\max(p, 1-p))^2}\right).$$

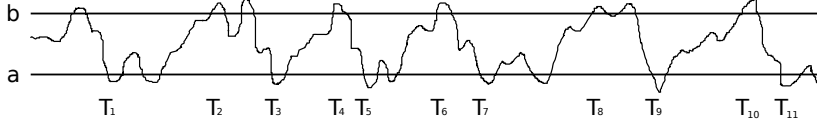
In particular, if  $p = 1/2$ ,

$$P(|S_n - n/2| \geq x\sqrt{n}) \leq 2e^{-2x^2}.$$

Putting here  $x = 3$  we get  $P(|S_n - n/2| \geq 3\sqrt{n}) \leq 3 \cdot 10^{-8}$ .

## 8.5 Convergence in $L^1$

On the figure below five time intervals are shown:  $(T_1, T_2]$ ,  $(T_3, T_4]$ ,  $\dots$ ,  $(T_9, T_{10}]$  for the upcrossings of the interval  $(a, b)$ . If for all rational intervals  $(a, b)$  the number of upcrossings  $U(a, b)$  is finite, then the corresponding trajectory has a (possibly infinite) limit.



All  $T_k$  are stopping times. Definition 7.12 for the stopping times can be extended as follows.

**Definition 8.24** A random variable  $T$  taking values in the set  $\{0, 1, 2, \dots\} \cup \{\infty\}$  is called a stopping time with respect to the filtration  $(\mathcal{F}_n)$ , if

$$\{T = n\} \in \mathcal{F}_n, \quad \text{for all } n = 0, 1, 2, \dots$$

**Exercise 8.25** First, we 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  $T_1 = 0$ . Moreover,

for an even natural number  $k \geq 2$ , put  $T_k = n$  if  $Y_n \geq b$  and  $Y_i < b$  for  $T_{k-1} \leq i \leq n-1$ ,

and

for an odd natural number  $k \geq 3$ , put  $T_k = n$  if  $Y_n \leq a$  and  $Y_i > a$  for  $T_{k-1} \leq i \leq n-1$ .

Show that all crossing times  $T_k$  are stopping times. (Hint: use induction over  $k$ .)

**Lemma 8.26** *Snell upcrossing inequality.* Let  $a < b$  and  $U_n(a, b)$  is the number of upcrossings of the interval  $(a, b)$  for a submartingale  $(Y_0, \dots, Y_n)$ . Then

$$\mathbb{E}[U_n(a, b)] \leq \frac{\mathbb{E}[(Y_n - a) \vee 0]}{b - a}.$$

**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  $(Y_0, \dots, Y_{i-1})$  only. Now, since  $Z_n = (Y_n - a) \vee 0$  forms a submartingale, we get

$$\begin{aligned} \mathbb{E}((Z_i - Z_{i-1}) \cdot I_i) &= \mathbb{E}[\mathbb{E}(I_i \cdot (Z_i - Z_{i-1}) | \mathcal{F}_{i-1})] = \mathbb{E}[I_i \cdot (\mathbb{E}(Z_i | \mathcal{F}_{i-1}) - Z_{i-1})] \\ &\leq \mathbb{E}[\mathbb{E}(Z_i | \mathcal{F}_{i-1}) - Z_{i-1}] = \mathbb{E}(Z_i) - \mathbb{E}(Z_{i-1}). \end{aligned}$$

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

$$(b - a) \mathbb{E} U_n(a, b) \leq \mathbb{E} \sum_{i=1}^n (\mathbb{E}(Z_i | \mathcal{F}_{i-1}) - Z_{i-1}) I_i \leq \mathbb{E} \sum_{i=1}^n (\mathbb{E}(Z_i | \mathcal{F}_{i-1}) - Z_{i-1}) = \mathbb{E}(Z_n) - \mathbb{E}(Z_0) \leq \mathbb{E}(Z_n).$$

**Theorem 8.27** Suppose  $(Y_n, \mathcal{F}_n)$  is a submartingale such that for some finite constant  $M$ ,

$$\sup_n \mathbb{E}(Y_n \vee 0) \leq M.$$

Then

- (i) there exists a r.v.  $Y$  such that  $Y_n \rightarrow Y$  almost surely,
- (ii) the limit  $Y$  has a finite mean,
- (iii) if  $(Y_n)$  is uniformly integrable, see Definition 3.22, then  $Y_n \rightarrow Y$  in  $L^1$ .

PROOF. (i) Using Snell 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,  $\mathbb{P}(U(a, b) < \infty) = 1$  for any given interval  $(a, b)$ . Since there are only countably many rationals, it follows that

$$\mathbb{P}\{U(a, b) < \infty \text{ for all rational } (a, b)\} = 1,$$

which implies that  $Y_n$  converges almost surely.

- (ii) Denote  $Y_n^+ = Y_n \vee 0$ . Since  $|Y_n| = 2Y_n^+ - Y_n$  and  $\mathbb{E}(Y_n | \mathcal{F}_0) \geq Y_0$ , we get

$$\mathbb{E}(|Y_n| | \mathcal{F}_0) \leq 2\mathbb{E}(Y_n^+ | \mathcal{F}_0) - Y_0.$$

By Fatou lemma

$$\mathbb{E}(|Y| | \mathcal{F}_0) = \mathbb{E}(\liminf_{n \rightarrow \infty} |Y_n| | \mathcal{F}_0) \leq \liminf_{n \rightarrow \infty} \mathbb{E}(|Y_n| | \mathcal{F}_0) \leq 2 \liminf_{n \rightarrow \infty} \mathbb{E}(Y_n^+ | \mathcal{F}_0) - Y_0,$$

and it remains to observe that again due to Fatou lemma,

$$\mathbb{E}(\liminf_{n \rightarrow \infty} \mathbb{E}(Y_n^+ | \mathcal{F}_0)) \leq \liminf_{n \rightarrow \infty} \mathbb{E}(Y_n^+ | \mathcal{F}_0) \leq M.$$

- (iii) Finally, recall that according to Theorem 3.25, given  $Y_n \xrightarrow{\mathbb{P}} Y$ , 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 8.28** 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 mean.

**Corollary 8.29** A non-negative supermartingale converges almost surely to a r.v. with a finite mean. A non-positive submartingale converges almost surely to a r.v. with a finite mean.

**Example 8.30** 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  $S_n \rightarrow \infty$  for  $p > q$  and  $S_n \rightarrow -\infty$  for  $p < q$ , in both cases we get  $\mathbb{P}(Y = 0) = 1$ . Note that  $Y_n$  does not converge in mean, since  $\mathbb{E}(Y_n) = \mathbb{E}(Y_0) \neq 0$ .

**Exercise 8.31** Knapsack problem. It is required to pack a knapsack of volume  $c$  to maximum benefit. Suppose you have  $n$  objects, the  $i$ th object having volume  $V_i$  and worth  $W_i$ , where  $V_1, \dots, V_n, W_1, \dots, W_n$  are independent non-negative random variables with finite means. For any vector  $(z_1, \dots, z_n)$  of 0 and 1 such that

$$z_1 V_1 + \dots + z_n V_n \leq c,$$

define the packing benefit to be

$$W(z_1, \dots, z_n) = z_1 W_1 + \dots + z_n W_n.$$

Let  $Z$  be the maximal possible  $W(z_1, \dots, z_n)$  over all eligible vectors  $(z_1, \dots, z_n)$ . 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  $\mathcal{F}_i = \sigma(V_1, W_1, \dots, V_i, W_i)$  for  $i = 1, \dots, n$ . And let  $Y_i = E(Z|\mathcal{F}_i)$ ,  $i = 0, 1, \dots, n$  be a Doob martingale, see next section. 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  $j$ th object. Clearly,

$$E(Z(j)|\mathcal{F}_j) = E(Z(j)|\mathcal{F}_{j-1}).$$

Since obviously  $Z(j) \leq Z \leq Z(j) + 1$ , we have

$$Y_i - Y_{i-1} = E(Z|\mathcal{F}_i) - E(Z|\mathcal{F}_{i-1}) \leq 1 + E(Z(i)|\mathcal{F}_i) - E(Z(i)|\mathcal{F}_{i-1}) = 1,$$

and

$$Y_i - Y_{i-1} = E(Z|\mathcal{F}_i) - E(Z|\mathcal{F}_{i-1}) \geq E(Z(i)|\mathcal{F}_i) - E(Z(i)|\mathcal{F}_{i-1}) - 1 = -1.$$

## 8.6 Doob's martingale

**Theorem 8.32** *Let  $Z$  be a r.v. on  $(\Omega, \mathcal{F}, P)$  such that  $E(|Z|) < \infty$ . For a filtration  $(\mathcal{F}_n)$  define  $Y_n = E(Z|\mathcal{F}_n)$ . Then  $(Y_n, \mathcal{F}_n)$  is a uniformly integrable martingale, which is called Doob's martingale.*

PROOF. Put  $Z_n = E(|Z| |\mathcal{F}_n)$  and observe that  $|Y_n| \leq Z_n$  by the Jensen inequality. The finiteness of the means follows from

$$E(|Y_n|) \leq E(Z_n) = E(|Z|),$$

and the martingale property is straightforward

$$E(Y_{n+1}|\mathcal{F}_n) = E(E(Z|\mathcal{F}_{n+1})|\mathcal{F}_n) = E(Z|\mathcal{F}_n) = Y_n.$$

To show uniform integrability observe that the inequality  $|Y_n| \leq Z_n$  implies that for any  $a$ ,

$$|Y_n|1_{\{|Y_n| \geq a\}} \leq Z_n 1_{\{Z_n \geq a\}}.$$

Since on the other hand, by Definition 2.7 of conditional expectation,

$$E((|Z| - Z_n)1_{\{Z_n \geq a\}}) = 0,$$

we find that

$$E(|Y_n|1_{\{|Y_n| \geq a\}}) \leq E(|Z|1_{\{Z_n \geq a\}}).$$

To derive from here the uniform integrability

$$\sup_n E(|Y_n|1_{\{|Y_n| \geq a\}}) \rightarrow 0, \quad a \rightarrow \infty,$$

it is enough to show that

$$\sup_n E(|Z|1_{\{Z_n \geq a\}}) \rightarrow 0, \quad a \rightarrow \infty.$$

The latter is obtained using the Markov inequality. Indeed, for arbitrary  $C > 0$ ,

$$E(|Z|1_{\{Z_n \geq a\}}) \leq C \cdot P(Z_n \geq a) + E(|Z|1_{\{|Z| \geq C\}}) \leq C \cdot \frac{E(Z_n)}{a} + E(|Z|1_{\{|Z| \geq C\}}),$$

implying that

$$\sup_n E(|Z|1_{\{Z_n \geq a\}}) \leq C \cdot \frac{E|Z|}{a} + E(|Z|1_{\{|Z| \geq C\}}).$$

It remains to let first  $a \rightarrow \infty$  and then  $C \rightarrow \infty$ .

### Remarks

- By Theorem 8.27, Doob's martingale converges  $E(Z|\mathcal{F}_n) \rightarrow Y$  a.s. and in mean as  $n \rightarrow \infty$ .
- It is actually the case (without proof) that the limit  $Y = E(Z|\mathcal{F}_\infty)$ , where  $\mathcal{F}_\infty$  is the smallest  $\sigma$ -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  $Y_n = E(Z|\mathcal{F}_n)$ .

## 8.7 Bounded stopping times. Optional sampling theorem

The stopped de Moivre martingale from Example 8.7 is also a martingale. A general statement of this type follows next.

**Theorem 8.33** *Let  $(Y_n, \mathcal{F}_n)$  be a submartingale and let  $T$  be a stopping time. Then both  $(Y_{T \wedge n}, \mathcal{F}_n)$  and  $(Y_n - Y_{T \wedge n}, \mathcal{F}_n)$  are also submartingales.*

PROOF. The random variable  $Z_n = Y_{T \wedge n}$  is  $\mathcal{F}_n$ -measurable:

$$Z_n = \sum_{i=0}^{n-1} Y_i 1_{\{T \geq i\}} + Y_n 1_{\{T \geq n\}},$$

with

$$\mathbb{E}|Z_n| \leq \sum_{i=0}^n \mathbb{E}|Y_i| < \infty.$$

It remains to see that the equality

$$Z_{n+1} - Z_n = (Y_{n+1} - Y_n) 1_{\{T > n\}}$$

implies on one hand,

$$\mathbb{E}(Z_{n+1} | \mathcal{F}_n) - Z_n = \mathbb{E}(Y_{n+1} - Y_n | \mathcal{F}_n) 1_{\{T > n\}} \geq 0,$$

and on the other hand,

$$\mathbb{E}(Y_{n+1} - Z_{n+1} | \mathcal{F}_n) - (Y_n - Z_n) = \mathbb{E}(Y_{n+1} - Y_n | \mathcal{F}_n) 1_{\{T \leq n\}} \geq 0.$$

**Corollary 8.34** *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 8.35** For a stopping time  $T$  with respect to filtration  $(\mathcal{F}_n)$ , denote by  $\mathcal{F}_T$  the  $\sigma$ -algebra of all events  $A$  such that

$$A \cap \{T = n\} \in \mathcal{F}_n \quad \text{for all } n.$$

**Exercise 8.36** Turning to Example 1.13 of consecutive Bernoulli trials, consider

$$T = \min\{n : S_n = 1\}.$$

Describe the corresponding sigma-algebra  $\mathcal{F}_T$ .

**Exercise 8.37** If  $T \leq S$  are two stopping times, then  $\mathcal{F}_T \subset \mathcal{F}_S$ .

**Definition 8.38** 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  $0 = T_0 \leq T_1 \leq T_2 \leq \dots$  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. By the previous theorem,  $Z_n = Y_{T \wedge n}$  is a submartingale. Therefore  $Z_N = Y_T$ , we have  $\mathbb{E}|Y_T| = \mathbb{E}|Z_N| < \infty$  and

$$\mathbb{E}(Y_T | \mathcal{F}_0) = \mathbb{E}(Z_N | \mathcal{F}_0) \geq Z_0 = Y_0.$$

(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  $S \leq T \leq N$  and put  $W = E(Y_T | \mathcal{F}_S)$ . To show that  $W \geq Y_S$ , observe that for  $A \in \mathcal{F}_S$  we have

$$E(W1_A) = E(Y_T 1_A) = \sum_{k \leq N} E(Y_T 1_{A \cap \{S=k\}}) = \sum_{k \leq N} E\left(1_{A \cap \{S=k\}} E(Y_T | \mathcal{F}_k)\right),$$

and since in view of Theorem 8.33,

$$E(Y_T | \mathcal{F}_k) = E(Y_{T \wedge N} | \mathcal{F}_k) \geq Y_{T \wedge k} \quad \text{for all } k \leq N,$$

we conclude

$$E(W1_A) \geq E\left(\sum_{k \leq N} 1_{A \cap \{S=k\}} Y_{T \wedge k}\right) = E\left(\sum_{k \leq N} 1_{A \cap \{S=k\}} Y_k\right) = E(Y_S 1_A),$$

so that  $E((W - Y_S)1_A) \geq 0$  for all  $A \in \mathcal{F}_S$ . In particular, for any given  $\epsilon > 0$ , we can put  $A_\epsilon = \{W \leq Y_S - \epsilon\}$  and obtain

$$0 \leq E((W - Y_S)1_{A_\epsilon}) \leq -\epsilon P(A_\epsilon).$$

Thus  $P(A_\epsilon) = 0$  and  $W \geq Y_S$  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  $E(Y_T | \mathcal{F}_0) = Y_0$  and  $E(Y_T) = E(Y_0)$ .

## 8.8 Unbounded stopping times

**Theorem 8.41** Let  $(Y_n, \mathcal{F}_n)$  be a martingale and  $T$  be a stopping time. If

- (a)  $P(T < \infty) = 1$ ,
- (b)  $E|Y_T| < \infty$ ,
- (c)  $E(Y_n 1_{\{T > n\}}) \rightarrow 0$  as  $n \rightarrow \infty$ ,

then  $E(Y_T) = E(Y_0)$ .

PROOF. From  $Y_T = Y_{T \wedge n} + (Y_T - Y_n)1_{\{T > n\}}$  using that  $E(Y_{T \wedge n}) = E(Y_0)$  we obtain

$$E(Y_T) = E(Y_0) + E(Y_T 1_{\{T > n\}}) - E(Y_n 1_{\{T > n\}}).$$

It remains to apply (c) and observe that using (a), (b) and the dominated convergence theorem we get

$$E(Y_T 1_{\{T > n\}}) \rightarrow 0.$$

**Theorem 8.42** Let  $(Y_n, \mathcal{F}_n)$  be a martingale and  $T$  be a stopping time. If  $E(T) < \infty$  and there exists a constant  $c$  such that for any  $n$

$$E(|Y_{n+1} - Y_n| | \mathcal{F}_n) 1_{\{T > n\}} \leq c 1_{\{T > n\}},$$

then  $E(Y_T) = E(Y_0)$ .

PROOF. Since  $T \wedge n \rightarrow T$ , we have  $Y_{T \wedge n} \rightarrow Y_T$  a.s. It follows that

$$E(Y_0) = E(Y_{T \wedge n}) \rightarrow E(Y_T)$$

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

$$\sup_n |Y_{T \wedge n}| \leq |Y_0| + W, \quad W := |Y_1 - Y_0| + \dots + |Y_T - Y_{T-1}|,$$

and to verify that  $E(W) < \infty$ . Indeed, since

$$E(|Y_i - Y_{i-1}| 1_{\{T \geq i\}} | \mathcal{F}_{i-1}) \leq c 1_{\{T \geq i\}},$$



we have

$$\mathbb{E}(|Y_i - Y_{i-1}|1_{\{T \geq i\}}) \leq c\mathbb{P}(T \geq i),$$

and therefore

$$\begin{aligned} \mathbb{E}(W) &= \mathbb{E}\left(\sum_{i=1}^T |Y_i - Y_{i-1}|\right) = \sum_{i=1}^{\infty} \mathbb{E}(|Y_i - Y_{i-1}|1_{\{T \geq i\}}) = \sum_{i=1}^{\infty} \mathbb{E}\left(\mathbb{E}(|Y_i - Y_{i-1}|1_{\{T \geq i\}} | \mathcal{F}_{i-1})\right) \\ &\leq c \sum_{i=1}^{\infty} \mathbb{P}(T \geq i) \leq c\mathbb{E}(T) < \infty. \end{aligned}$$

**Example 8.43** Consider a simple random walk  $S_n = k + X_1 + \dots + X_n$  with

$$\mathbb{P}(X_i = 1) = p, \quad \mathbb{P}(X_i = -1) = q.$$

Let  $T$  be the first time when it hits 0 or  $N$ , assuming  $0 \leq k \leq N$ . Next we show that for  $p \neq q$ ,

$$\mathbb{P}(S_T = 0) = \frac{(p/q)^{N-k} - 1}{(p/q)^N - 1}.$$

According to Definition 7.12,  $T$  is a stopping time. The 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 by applying Theorem 8.42 to De Moivre martingale  $Y_n = (q/p)^{S_n}$ , we get

$$\mathbb{E}(Y_T) = \mathbb{E}(Y_0).$$

Put  $p_k := \mathbb{P}(S_T = 0)$  so that  $\mathbb{P}(S_T = N) = 1 - p_k$ . 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  $N \rightarrow \infty$ , we obtain  $\mathbb{P}(S_T = -1) \rightarrow \frac{q}{p}$ . Thus, if  $(-X)$  is the lowest level visited, then  $X$  has a shifted geometric distribution with parameter  $1 - \frac{q}{p}$ .

**Example 8.44** Wald's equality. Let  $(X_n)$  be iid r.v. with finite mean  $\mu$  and  $S_n = X_1 + \dots + X_n$ , then  $Y_n = S_n - n\mu$  is a martingale with respect to  $\mathcal{F}_n = \sigma\{X_1, \dots, X_n\}$ . Now

$$\mathbb{E}(|Y_{n+1} - Y_n| | \mathcal{F}_n) = \mathbb{E}|X_{n+1} - \mu| = \mathbb{E}|X_1 - \mu| < \infty.$$

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 8.45** *Wald's identity.* Let  $(X_n)$  be iid r.v. with a finite  $M(t) = \mathbb{E}(e^{tX})$ . Put  $S_n = X_1 + \dots + X_n$ . If  $T$  is a stopping time with finite mean such that  $|S_n|1_{\{T > n\}} \leq c1_{\{T > n\}}$ , then

$$\mathbb{E}(e^{tS_T} M(t)^{-T}) = 1 \text{ whenever } M(t) \geq 1.$$

PROOF. Define  $Y_0 = 1$ ,  $Y_n = e^{tS_n} M(t)^{-n}$ , and let  $\mathcal{F}_n = \sigma\{X_1, \dots, X_n\}$ . It is clear that  $(Y_n)$  is a martingale and thus the claim follows from Theorem 8.42 after we verify the key condition. By the definition of  $Y_n$ ,

$$\mathbb{E}(|Y_{n+1} - Y_n| | \mathcal{F}_n) = Y_n \mathbb{E}|e^{tX} M(t)^{-1} - 1| \leq Y_n \mathbb{E}(e^{tX} M(t)^{-1} + 1) = 2Y_n.$$

Furthermore, given  $M(t) \geq 1$ ,

$$Y_n 1_{\{T > n\}} = e^{tS_n} M(t)^{-n} 1_{\{T > n\}} \leq e^{tS_n} 1_{\{T > n\}} \leq e^{t|S_n|} 1_{\{T > n\}} \leq e^{c|t|} 1_{\{T > n\}}.$$

**Example 8.46** 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  $(-a, b)$  for some  $a > 0$  and  $b > 0$ . Clearly,  $|S_n| 1_{\{T > n\}} \leq c 1_{\{T > n\}}$  with  $c = a \vee b$ .

By Lemma 8.45 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  $pe^t + qe^{-t} = s^{-1}$ , we obtain a quadratic equation for  $e^t$  having two solutions  $\lambda_i = \lambda_i(s)$ :

$$\lambda_1 = \frac{1 + \sqrt{1 - 4pqs^2}}{2ps}, \quad \lambda_2 = \frac{1 - \sqrt{1 - 4pqs^2}}{2ps}, \quad s \in [0, 1].$$

This observation yields 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

$$\mathbb{E}(s^T 1_{\{S_T = -a\}}) = \frac{\lambda_1^a \lambda_2^a (\lambda_1^b - \lambda_2^b)}{\lambda_1^{a+b} - \lambda_2^{a+b}}, \quad \mathbb{E}(s^T 1_{\{S_T = b\}}) = \frac{\lambda_1^a - \lambda_2^a}{\lambda_1^{a+b} - \lambda_2^{a+b}}.$$

Summing up these two relations we get a formula for the probability generating function of the stopping time  $T$

$$\mathbb{E}(s^T) = \frac{\lambda_1^a (1 - \lambda_2^{a+b}) + \lambda_2^a (\lambda_1^{a+b} - 1)}{\lambda_1^{a+b} - \lambda_2^{a+b}}.$$

## 8.9 Maximal inequality and convergence in $L^r$

For a random variable  $X$ , we write  $X^+ = X \wedge 0$  and  $X^- = (-X) \wedge 0$ , so that

$$X = X^+ - X^-, \quad |X| = X^+ + X^-.$$

**Theorem 8.47** *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, 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. 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 8.33,  $\mathbb{E}(Y_n^+ - Y_{T \wedge n}^+) \geq 0$ . 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  $E(Y_n^+ 1_{\{T>n\}}) = E(Y_{T \wedge n}^+ 1_{\{T>n\}})$ , we have

$$E(Y_n^+ 1_{\{T \leq n\}}) = E(Y_n^+) - E(Y_{T \wedge n}^+ 1_{\{T>n\}}) \geq E(Y_{T \wedge n}^+ 1_{\{T \leq n\}}) = E(Y_T^+ 1_{\{T \leq n\}}) \geq \epsilon P(T \leq n).$$

Using this we get even a stronger inequality

$$P(A) \leq \frac{E(Y_n^+ 1_A)}{\epsilon}, \text{ where } A = \left\{ \max_{0 \leq i \leq n} Y_i \geq \epsilon \right\}. \quad (*)$$

(ii) If  $(Y_n)$  is a supermartingale, then by the first part of Theorem 8.33,  $E(-Y_{T \wedge n}) \geq E(-Y_0)$ . Thus

$$E(Y_0) \geq E(Y_{T \wedge n}) = E(Y_T 1_{\{T \leq n\}}) + E(Y_n 1_{\{T>n\}}) \geq \epsilon P(T \leq n) - E(Y_n^- 1_{\{T>n\}})$$

giving the second assertion.

**Corollary 8.48** *If  $(Y_n)$  is a submartingale, then for any  $\epsilon > 0$ ,*

$$P\left(\max_{0 \leq i \leq n} |Y_i| \geq \epsilon\right) \leq \frac{2E(Y_n^+) - E(Y_0)}{\epsilon}.$$

PROOF. Let  $\epsilon > 0$ . If  $(Y_n)$  is a submartingale, then  $(-Y_n)$  is a supermartingale so that according to Theorem 8.47 (ii),

$$P\left(\min_{0 \leq i \leq n} Y_i \leq -\epsilon\right) = P\left(\max_{0 \leq i \leq n} (-Y_i) \geq \epsilon\right) \leq \frac{E(-Y_0) + E[(-Y_n)^-]}{\epsilon} = \frac{E(Y_n^+) - E(Y_0)}{\epsilon}.$$

Adding this to Theorem 8.47 (i) we arrive at the asserted inequality.

**Exercise 8.49** Show that if  $(Y_n)$  is a martingale, then

$$P\left(\max_{0 \leq i \leq n} |Y_i| \geq \epsilon\right) \leq \frac{E|Y_n|}{\epsilon}.$$

**Corollary 8.50** *Doob-Kolmogorov inequality. If  $(Y_n)$  is a martingale with finite second moments, then  $(Y_n^2)$  is a submartingale, and by Theorem 8.47 (i), for any  $\epsilon > 0$ ,*

$$P\left(\max_{1 \leq i \leq n} |Y_i| \geq \epsilon\right) = P\left(\max_{1 \leq i \leq n} Y_i^2 \geq \epsilon^2\right) \leq \frac{E(Y_n^2)}{\epsilon^2}.$$

**Corollary 8.51** *Kolmogorov inequality. Let  $(X_n)$  are independent r.v. with zero means and finite variances  $(\sigma_n^2)$ , then for any  $\epsilon > 0$ ,*

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

**Theorem 8.52** *Convergence in  $L^r$ . Let  $r > 1$ . Suppose  $(Y_n, \mathcal{F}_n)$  is a martingale such that*

$$\sup_n E(|Y_n|^r) \leq M$$

*for some constant  $M$ . Then there exists  $Y$  such that  $Y_n \rightarrow Y$  a.s. as  $n \rightarrow \infty$ , and moreover,*

$$Y_n \rightarrow Y \text{ in } L^r.$$

PROOF. Combining Corollary 8.28 and Lyapunov inequality we get the a.s. convergence  $Y_n \rightarrow Y$ . We prove  $Y_n \xrightarrow{L^r} Y$  using Theorem 3.26. For this we need to verify that  $(|Y_n^r|)_{n \geq 0}$  is uniformly integrable. Observe first that

$$\bar{Y}_n := \max_{0 \leq i \leq n} |Y_i|$$

have finite  $r$ -th moment

$$E(\bar{Y}_n^r) \leq E(|Y_0|^r + \dots + |Y_n|^r) < \infty.$$

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 (\*) to  $A_x = \{\bar{Y}_n \geq x\}$ , we obtain

$$\mathbb{E}(\bar{Y}_n^r) \leq \int_0^\infty r x^{r-2} \mathbb{E}(|Y_n| 1_{A_x}) dx = r \mathbb{E}\left(|Y_n| \int_{0 \leq x \leq \bar{Y}_n} x^{r-2} dx\right) = \frac{r}{r-1} \mathbb{E}\left(|Y_n| \bar{Y}_n^{r-1}\right).$$

Now, applying the Hölder inequality we find

$$\mathbb{E}(\bar{Y}_n^r) \leq \frac{r}{r-1} \mathbb{E}\left(|Y_n| \bar{Y}_n^{r-1}\right) \leq \frac{r}{r-1} \left(\mathbb{E}(|Y_n|^r)\right)^{1/r} \left(\mathbb{E}(\bar{Y}_n^r)\right)^{(r-1)/r},$$

and we conclude

$$\mathbb{E}\left(\max_{0 \leq i \leq n} |Y_i|^r\right) = \mathbb{E}(\bar{Y}_n^r) \leq \left(\frac{r}{r-1}\right)^r \mathbb{E}(|Y_n|^r) \leq \left(\frac{r}{r-1}\right)^r M.$$

Thus, by monotone convergence,  $\mathbb{E}(\sup_n |Y_n|^r) < \infty$  implying that  $(|Y_n^r|)_{n \geq 0}$  is uniformly integrable, see Exercise 3.23.

## 8.10 Backward martingales. Martingale proof of the strong LLN

**Definition 8.53** Let  $(\mathcal{G}_n)$  be a decreasing sequence of  $\sigma$ -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  $n \geq 0$ ,

$$\begin{aligned} \mathbb{E}(|Y_n|) &< \infty, \\ \mathbb{E}(Y_n | \mathcal{G}_{n+1}) &= Y_{n+1}. \end{aligned}$$

**Theorem 8.54** Let  $(Y_n, \mathcal{G}_n)$  be a backward martingale. Then  $Y_n = \mathbb{E}(Y_0 | \mathcal{G}_n)$  and there is a r.v.  $Y$  such that  $Y_n \rightarrow Y$  almost surely and in mean.

PROOF. Using the tower property for conditional expectations we prove the first statement

$$Y_n = \mathbb{E}(Y_{n-1} | \mathcal{G}_n) = \mathbb{E}(\mathbb{E}(Y_{n-2} | \mathcal{G}_{n-1}) | \mathcal{G}_n) = \mathbb{E}(Y_{n-2} | \mathcal{G}_n) = \dots = \mathbb{E}(Y_0 | \mathcal{G}_n).$$

To prove a.s. convergence, we apply the upcrossing inequality, Lemma 8.26, to the martingale  $(Y_n, \mathcal{G}_n), \dots, (Y_0, \mathcal{G}_0)$ . The number  $U_n(a, b)$  of  $[a, b]$  upcrossings by  $(Y_n, \dots, Y_0)$  satisfies

$$\mathbb{E} U_n(a, b) \leq \frac{\mathbb{E}(Y_0 - a)^+}{b - a},$$

so that letting  $n \rightarrow \infty$  and repeating the proof of Theorem 8.27 we arrive at the stated a.s. convergence.

The convergence in mean follows from the observation that the sequence  $Y_n = \mathbb{E}(Y_0 | \mathcal{G}_n)$  is uniformly integrable, which is obtained by repeating the proof of Theorem 8.32 dealing with the Doob martingale.

**Theorem 8.55** 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

$$\frac{X_1 + \dots + X_n}{n} \rightarrow \mathbb{E}X_1$$

almost surely and in  $L^1$ . On the other hand, if

$$\frac{X_1 + \dots + X_n}{n} \xrightarrow{\text{a.s.}} \mu$$

for some constant  $\mu$ , then  $\mathbb{E}|X_1| < \infty$  and  $\mu = \mathbb{E}X_1$ .

PROOF. Let  $E|X_1| < \infty$ . Set  $S_n = X_1 + \dots + X_n$  and let  $\mathcal{G}_n = \sigma(S_n, S_{n+1}, \dots)$ , then

$$E(S_n | \mathcal{G}_{n+1}) = E(S_n | S_{n+1}, X_{n+2}, X_{n+3}, \dots) = E(S_n | S_{n+1}),$$

and by symmetry,

$$E(S_n | S_{n+1}) = \sum_{i=1}^n E(X_i | S_{n+1}) = nE(X_1 | S_{n+1}).$$

On the other hand,

$$S_{n+1} = E(S_{n+1} | S_{n+1}) = (n+1)E(X_1 | S_{n+1}),$$

implying

$$n^{-1}E(S_n | \mathcal{G}_{n+1}) = E(X_1 | S_{n+1}) = (n+1)^{-1}S_{n+1}.$$

We conclude that  $S_n/n$  is a backward martingale, and according to Theorem 8.54 there exists  $Y$  such that  $S_n/n \rightarrow Y$  a.s. and in mean. Convergence in mean yields

$$E(Y) = E(S_n/n) = E(X_1).$$

It remains to see that by Kolmogorov zero-one law,  $P(Y > c)$  is either 0 or 1 for any given constant  $c$ . Therefore,  $Y = E(X_1)$  a.s.

To prove the reverse part, assume that  $S_n/n \xrightarrow{\text{a.s.}} \mu$  for some finite constant  $\mu$ . Then  $X_n/n \xrightarrow{\text{a.s.}} 0$  by the theory of convergent real series. Indeed, from  $(a_1 + \dots + a_n)/n \rightarrow \mu$  it follows that

$$\frac{a_n}{n} = \frac{a_1 + \dots + a_{n-1}}{n(n-1)} + \frac{a_1 + \dots + a_n}{n} - \frac{a_1 + \dots + a_{n-1}}{n-1} \rightarrow 0$$

Now, in view of  $X_n/n \xrightarrow{\text{a.s.}} 0$ , the second Borell-Cantelli lemma gives

$$\sum_n P(|X_n| \geq n) < \infty,$$

since otherwise  $P(n^{-1}|X_n| \geq 1 \text{ i.o.}) = 1$ . Thus

$$E|X_1| = \int_0^\infty P(|X_1| > x) dx \leq \sum_n P(|X_1| \geq n) = \sum_n P(|X_n| \geq n) < \infty.$$

Applying the first part of this theorem we see that  $\mu = EX_1$ .

## 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  $0 \leq s < t$ ,
- the path  $t \rightarrow W_t$  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 extension theorem.* Assume that for any vector  $(t_1, \dots, t_n)$  with  $t_i \in [0, 1]$  there given a joint distribution function  $F_{(t_1, \dots, t_n)}(x_1, \dots, x_n)$ . Suppose that these distribution functions satisfy two consistency conditions

$$(i) F_{(t_1, \dots, t_n, t_{n+1})}(x_1, \dots, x_n, \infty) = F_{(t_1, \dots, t_n)}(x_1, \dots, x_n),$$

$$(ii) \text{ if } \pi \text{ is a permutation of } (1, \dots, n), \text{ then } F_{(t_{\pi(1)}, \dots, t_{\pi(n)})}(x_{\pi(1)}, \dots, x_{\pi(n)}) = F_{(t_1, \dots, t_n)}(x_1, \dots, x_n).$$

Put  $\Omega = \{\text{functions } \omega : [0, 1] \rightarrow \mathbb{R}\}$  and  $\mathcal{F}$  is the  $\sigma$ -algebra generated by the finite-dimensional sets  $\{\omega : \omega(t_i) \in B_i, i = 1, \dots, n\}$ , where  $B_i$  are Borel subsets of  $\mathbb{R}$ . Then there is a unique probability measure  $P$  on  $(\Omega, \mathcal{F})$  such that a stochastic process defined by  $X(t, \omega) = \omega(t)$  has the finite-dimensional distributions  $F_{(t_1, \dots, t_n)}(x_1, \dots, x_n)$ .

The problem is that the set  $\{\omega : t \rightarrow \omega(t) \text{ is continuous}\}$  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  $0 \leq m \leq 2^n, n \geq 1\}$ .

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  $0 \leq s < t, s, t \in \mathcal{Q}$ .

According to Theorem 9.2 the process  $X(t), t \in \mathcal{Q}$  can be defined on  $(\Omega_q, \mathcal{F}_q, P_q)$  where index  $q$  means the restriction  $t \in \mathcal{Q}$ .

Step 2. For  $m2^{-n} \leq t < (m+1)2^{-n}$  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  $X_n(t) \rightarrow X(t)$  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  $X_n(t) \rightarrow X(t)$  a.s. uniformly over  $t \in \mathcal{Q}$ . To prove the latter we use the Weierstrass M-test by observing that  $X_n(t) = Z_1(t) + \dots + Z_n(t)$ , where  $Z_i(t) = X_i(t) - X_{i-1}(t)$ , and showing that

$$\sum_{i=1}^{\infty} \sup_{t \in \mathcal{Q}} |Z_i(t)| < \infty. \quad (*)$$

Using independence and normality of the increments one can show for  $x_i = c\sqrt{i2^{-i} \log 2}$  that

$$P(\sup_{t \in \mathcal{Q}} |Z_i(t)| > x_i) \leq 2^{i-1} \frac{2^{-ic^2}}{c\sqrt{i \log 2}}.$$

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  $(*)$  follows.

Step 4. Define  $W(t)$  for any  $t \in [0, 1]$  by moving our probability measure to  $(C, \mathcal{C})$ , where  $C =$  continuous  $\omega : [0, 1] \rightarrow \mathbb{R}$  and  $\mathcal{C}$  is the  $\sigma$ -algebra generated by the coordinate maps  $t \rightarrow \omega(t)$ . To do this, we observe that the map  $\psi$  that takes a uniformly continuous point in  $\Omega_q$  to its unique continuous extension in  $C$  is measurable, and we set  $P(A) = P_q(\psi^{-1}(A))$ .

## 9.2 Properties of the Wiener process

We will prove some of the following properties of the standard Wiener process.

(i) The vector  $(W(t_1), \dots, W(t_n))$  has the multivariate normal distribution with zero means and covariances  $\text{Cov}(W(t_i), W(t_j)) = \min(t_i, t_j)$ .

(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  $Z \sim N(0,1)$ , 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. (i) If  $0 \leq s < t$ , then

$$\mathbb{E}(W(s)W(t)) = \mathbb{E}[W(s)^2 + W(s)(W(t) - W(s))] = \mathbb{E}[W(s)^2] = s.$$

(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

$$\begin{aligned} \mathbb{P}(M(t) \geq x, W(t) < x) &= \mathbb{P}(M(t) \geq x, W(t) - W(T(x)) < 0, T(x) \leq t) \\ &= \mathbb{P}(M(t) \geq x, W(t) - W(T(x)) \geq 0, T(x) \leq t) = \mathbb{P}(M(t) \geq x, W(t) \geq x). \end{aligned}$$

Thus

$$\begin{aligned} \mathbb{P}(M(t) \geq x) &= 2\mathbb{P}(M(t) \geq x, W(t) \geq x) = 2\mathbb{P}(W(t) \geq x) \\ &= \mathbb{P}(W(t) \geq x) + \mathbb{P}(W(t) \leq -x) = \mathbb{P}(|W(t)| \geq x). \end{aligned}$$

(vi) We have

$$\mathbb{P}(T(x) \leq t) = \mathbb{P}(M(t) \geq x) = \mathbb{P}(|W(t)| \geq x) = \frac{2}{\sqrt{2\pi t}} \int_x^\infty e^{-\frac{y^2}{2t}} dy = \int_0^t \frac{|x|}{\sqrt{2\pi u^3}} e^{-\frac{x^2}{2u}} du.$$

(vii) Bringing  $e^{\theta W(s)}$  outside

$$\mathbb{E}(e^{\theta W(t)} | \mathcal{F}_s) = e^{\theta W(s)} \mathbb{E}(e^{\theta(W(t)-W(s))} | \mathcal{F}_s) = e^{\theta W(s)} e^{\theta^2(t-s)/2}.$$

(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  $a \rightarrow -\infty$  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

$$dX_t = \mu(t, X_t)dt + \sigma(t, X_t)dW_t. \quad (*)$$

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 8.18).

**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

$$dX_t = -\alpha(X_t - \theta)dt + \sigma dW_t,$$

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  $\theta$ , 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

$$\begin{aligned} \text{forward equation} \quad & \frac{\partial f}{\partial t} = -\frac{\partial}{\partial y}[\mu(t, y)f] + \frac{1}{2} \frac{\partial^2}{\partial y^2}[\sigma^2(t, y)f], \\ \text{backward equation} \quad & \frac{\partial f}{\partial s} = -\mu(s, x) \frac{\partial f}{\partial x} - \frac{1}{2} \sigma^2(s, x) \frac{\partial^2 f}{\partial x^2}. \end{aligned}$$

**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 (\*). Then  $Y_t = f(t, X_t)$  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

$$\mathbb{E}(X_t) = \theta + e^{-\alpha t}(X_0 - \theta), \quad \text{Var}(X_t) = \sigma^2(1 - e^{-2\alpha t})/2\alpha, \quad (**)$$

implying that  $X_t$  loses 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  $\theta$  and variance  $\sigma^2/2\alpha$ .

To verify these formula for the mean and variance we apply the following simple version of Ito 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 the equation from Example 9.5 we obtain

$$d(X_t e^{\alpha t}) = \alpha X_t e^{\alpha t} dt + e^{\alpha t}(\alpha(\theta - X_t) dt + \sigma dW_t) = \theta e^{\alpha t} \alpha dt + \sigma e^{\alpha t} dW_t.$$

Integration gives

$$X_t e^{\alpha t} - X_0 = \theta(e^{\alpha t} - 1) + \sigma \int_0^t e^{\alpha u} dW_u,$$

implying (\*\*), since in view of Example 9.7 and the formula in Example 6.18, we have

$$\mathbb{E}\left(\int_0^t e^{\alpha u} dW_u\right)^2 = \int_0^t e^{2\alpha u} du = \frac{e^{2\alpha t} - 1}{2\alpha}.$$

Observe that the correlation coefficient between  $X(s)$  and  $X(s+t)$  equals

$$\rho(s, s+t) = e^{-\alpha t} \sqrt{\frac{1 - e^{-2\alpha s}}{1 - e^{-2\alpha(s+t)}}} \rightarrow e^{-\alpha t}, \quad s \rightarrow \infty.$$

**Example 9.10** Geometric Brownian motion  $Y_t = e^{\mu t + \sigma W_t}$ . Due to the Ito formula

$$dY_t = \left(\mu + \frac{1}{2}\sigma^2\right)Y_t dt + \sigma Y_t dW_t,$$

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

$$dX_t^2 = 2X_t dX_t + \sigma_1(t, X_t)^2 dt, \quad dY_t^2 = 2Y_t dY_t + \sigma_2(t, Y_t)^2 dt, \\ d(X_t + Y_t)^2 = 2(X_t + Y_t)(dX_t + dY_t) + (\sigma_1(t, X_t) + \sigma_2(t, Y_t))^2 dt.$$

## 9.5 The Black-Scholes formula

**Lemma 9.13** Let  $(W_t, 0 \leq t \leq T)$  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  $0 = t_0 < t_1 < \dots < t_n = T$  and  $x_0, x_1, \dots, x_n \in \mathbb{R}$ . Writing  $\{W(t_i) \in dx_i\}$  for the event  $\{x_i < W(t_i) \leq x_i + dx_i\}$  we have that

$$\begin{aligned} \mathbb{Q}(W(t_1) \in dx_1, \dots, W(t_n) \in dx_n) &= \mathbb{E}(e^{\nu W_T - \nu^2 T/2} 1_{\{W(t_1) \in dx_1, \dots, W(t_n) \in dx_n\}}) \\ &= e^{\nu x_n - \nu^2 T/2} \prod_{i=1}^n \frac{1}{\sqrt{2\pi(t_i - t_{i-1})}} \exp\left(-\frac{(x_i - x_{i-1})^2}{2(t_i - t_{i-1})}\right) dx_i \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi(t_i - t_{i-1})}} \exp\left(-\frac{(x_i - x_{i-1} - \nu(t_i - t_{i-1}))^2}{2(t_i - t_{i-1})}\right) dx_i. \end{aligned}$$

**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  $\sigma$  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  $S_T \leq K$ , 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

$$d_1(t, x) = \frac{\log(x/K) + (r + \sigma^2/2)(T - t)}{\sigma\sqrt{T - t}}, \quad d_2(t, x) = \frac{\log(x/K) + (r - \sigma^2/2)(T - t)}{\sigma\sqrt{T - t}}.$$

**Definition 9.15** Let  $\mathcal{F}_t$  be the  $\sigma$ -algebra generated by  $(S_u, 0 \leq u \leq t)$ . A portfolio is a pair  $(\alpha_t, \beta_t)$  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  $(\alpha_t, \beta_t)$  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

$$e^{-rt} S_t = \exp\{(\nu\sigma - \sigma^2/2)t + \sigma \tilde{W}_t\} = e^{\sigma \tilde{W}_t - \sigma^2 t/2}$$

is a martingale. Take without a proof that there exists a self-financing portfolio  $(\alpha_t, \beta_t)$  replicating the European call option in question. If the market contains no arbitrage opportunities, we have  $V_t = V_t(\alpha, \beta)$  and therefore

$$\begin{aligned} d(e^{-rt}V_t) &= e^{-rt}dV_t - re^{-rt}V_t dt = e^{-rt}\alpha_t(dS_t - rS_t dt) + e^{-rt}\beta_t(dB_t - rB_t dt) \\ &= \alpha_t e^{-rt}S_t((\mu - r)dt + \sigma dW_t) = \alpha_t e^{-rt}S_t \sigma d\tilde{W}_t. \end{aligned}$$

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

$$V_t = e^{rt} \mathbb{E}_{\mathbb{Q}}(e^{-rT} V_T | \mathcal{F}_t) = e^{-r(T-t)} \mathbb{E}_{\mathbb{Q}}((S_T - K)^+ | \mathcal{F}_t) = e^{-r(T-t)} \mathbb{E}_{\mathbb{Q}}((ae^Z - K) \vee 0 | \mathcal{F}_t),$$

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  $(Z | \mathcal{F}_t)_{\mathbb{Q}} \sim N(\gamma, \tau^2)$  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  $Z \sim N(\gamma, \tau^2)$

$$\mathbb{E}(ae^Z - K)^+ = ae^{\gamma + \tau^2/2} \Phi\left(\frac{\log(a/K) + \gamma}{\tau} + \tau\right) - K \Phi\left(\frac{\log(a/K) + \gamma}{\tau}\right).$$