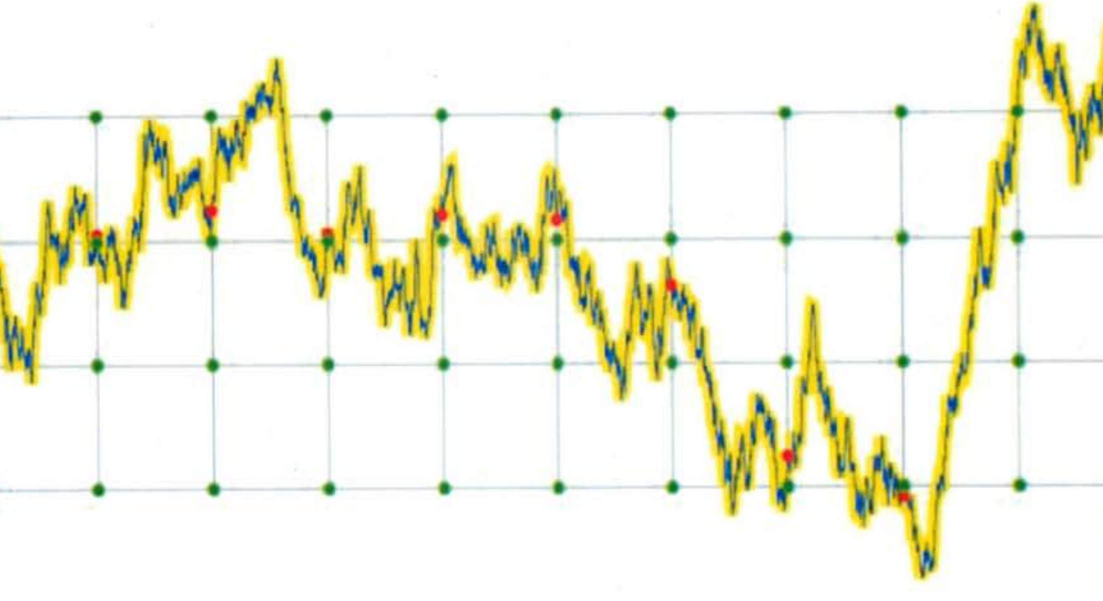
Probability Theory and Stochastic Processes with Applications



Oliver Knill

Overseas Press

Probability and Stochastic Processes with Applications

Probability and Stochastic Processes with Applications

Oliver Knill

OVERSEAS PRESS (INDIA) PVT. LTD.

Regd. Office:
Overseas Press India Private Limited
7/28, Ansari Road, Daryaganj
New Delhi-110 002
Email : info@overseaspub.com
Website : www.overseaspub.com

Sales Office :
Overseas Press India Private Limited
2/15, Ansari Road, Daryaganj
New Delhi-110 002
Email : orders@overseaspub.com
Website : www.overseaspub.com

Edition : 2009

10 digit ISBN : 81 - 89938 - 40 - 1
13 digit ISBN : 978 - 81 - 89938 - 40 - 6

Published by Narinder Kumar Lijhara for Overseas Press India Private Limited,
7/28, Ansari Road, Daryaganj, New Delhi-110002 and Printed in India.

Contents

Preface

These notes grew from an introduction to probability theory taught during the first and second term of 1994 at Caltech. There was a mixed audience of undergraduates and graduate students in the first half of the course which covered Chapters 2 and 3, and mostly graduate students in the second part which covered Chapter 4 and two sections of Chapter 5.

Having been online for many years on my personal web sites, the text got reviewed, corrected and indexed in the summer of 2006. It obtained some enhancements which benefited from some other teaching notes and research, I wrote while teaching probability theory at the University of Arizona in Tucson or when incorporating probability in calculus courses at Caltech and Harvard University.

Most of Chapter 2 is standard material and subject of virtually any course on probability theory. Also Chapters 3 and 4 is well covered by the literature but not in this combination.

The last chapter "selected topics" got considerably extended in the summer of 2006. While in the original course, only localization and percolation problems were included, I added other topics like estimation theory, Vlasov dynamics, multi-dimensional moment problems, random maps, circle-valued random variables, the geometry of numbers, Diophantine equations and harmonic analysis. Some of this material is related to research I got interested in over time.

While the text assumes no prerequisites in probability, a basic exposure to calculus and linear algebra is necessary. Some real analysis as well as some background in topology and functional analysis can be helpful.

I would like to get feedback from readers. I plan to keep this text alive and update it in the future. You can email this to knill@math.harvard.edu and also indicate on the email if you don't want your feedback to be acknowledged in an eventual future edition of these notes.

To get a more detailed and analytic exposure to probability, the students of the original course have consulted the book [105] which contains much more material than covered in class. Since my course had been taught, many other books have appeared. Examples are [21, 34].

For a less analytic approach, see [40, 91, 97] or the still excellent classic [26]. For an introduction to martingales, we recommend [108] and [47] from both of which these notes have benefited a lot and to which the students of the original course had access too.

For Brownian motion, we refer to [73, 66], for stochastic processes to [17], for stochastic differential equation to [2, 55, 76, 66, 46], for random walks to [100], for Markov chains to [27, 87], for entropy and Markov operators [61]. For applications in physics and chemistry, see [106].

For the selected topics, we followed [32] in the percolation section. The books [101, 30] contain introductions to Vlasov dynamics. The book of [1] gives an introduction for the moment problem, [75, 64] for circle-valued random variables, for Poisson processes, see [49, 9]. For the geometry of numbers for Fourier series on fractals [45].

The book [109] contains examples which challenge the theory with counter examples. [33, 92, 70] are sources for problems with solutions.

Probability theory can be developed using nonstandard analysis on finite probability spaces [74]. The book [42] breaks some of the material of the first chapter into attractive stories. Also texts like [89, 78] are not only for mathematical tourists.

We live in a time, in which more and more content is available online. Knowledge diffuses from papers and books to online websites and databases which also ease the digging for knowledge in the fascinating field of probability theory.

Oliver Knill

Chapter 1

Introduction

1.1 What is probability theory?

Probability theory is a fundamental pillar of modern mathematics with relations to other mathematical areas like algebra, topology, analysis, geometry or dynamical systems. As with any fundamental mathematical construction, the theory starts by adding more structure to a set Ω. In a similar way as introducing algebraic operations, a topology, or a time evolution on a set, probability theory adds a **measure theoretical structure** to Ω which generalizes "counting" on finite sets: in order to measure the **probability** of a subset $A \subset \Omega$, one singles out a class of subsets $\mathcal{A}$, on which one can hope to do so. This leads to the notion of a σ-algebra $\mathcal{A}$. It is a set of subsets of Ω in which on can perform finitely or **countably many** operations like taking unions, complements or intersections. The elements in $\mathcal{A}$ are called **events**. If a point ω in the "laboratory" Ω denotes an "experiment", an "event" $A \in \mathcal{A}$ is a subset of Ω, for which one can assign a probability $\mathrm{P}[A] \in [0,1]$. For example, if $\mathrm{P}[A] = 1/3$, the event happens with probability $1/3$. If $\mathrm{P}[A] = 1$, the event takes place almost certainly. The **probability measure** P has to satisfy obvious properties like that the **union** $A \cup B$ of two disjoint events A, B satisfies $\mathrm{P}[A \cup B] = \mathrm{P}[A] + \mathrm{P}[B]$ or that the **complement** A^c of an event A has the probability $\mathrm{P}[A^c] = 1 - \mathrm{P}[A]$. With a probability space $(\Omega, \mathcal{A}, \mathrm{P})$ alone, there is already some interesting mathematics: one has for example the **combinatorial problem** to find the probabilities of events like the event to get a "royal flush" in poker. If Ω is a subset of an Euclidean space like the plane, $\mathrm{P}[A] = \int_A f(x,y)\ dxdy$ for a suitable nonnegative **function** f, we are led to **integration problems** in calculus. Actually, in many applications, the probability space is part of Euclidean space and the σ-algebra is the smallest which contains all **open sets**. It is called the **Borel σ-algebra**. An important example is the Borel σ-algebra on the real line.

Given a probability space $(\Omega, \mathcal{A}, \mathrm{P})$, one can define **random variables** X. A random variable is a function X from Ω to the real line $\mathbb{R}$ which is **measurable** in the sense that the inverse of a measurable Borel set B in $\mathbb{R}$ is in $\mathcal{A}$. The interpretation is that if ω is an **experiment**, then $X(\omega)$ measures an **observable quantity** of the experiment. The technical condition of measurability resembles the notion of a **continuity** for a function f from a topological space $(\Omega, \mathcal{O})$ to the topological space $(\mathbb{R}, \mathcal{U})$. A function is continuous if $f^{-1}(U) \in \mathcal{O}$ for all open sets $U \in \mathcal{U}$. In probability theory, where functions are often denoted with capital letters, like $X, Y, \dots$, a random variable X is measurable if $X^{-1}(B) \in \mathcal{A}$ for all Borel sets $B \in \mathcal{B}$. Any continuous function is measurable for the Borel σ-algebra. As in calculus, where one does not have to worry about continuity most of the time, also in probability theory, one often does not have to sweat about measurability issues. Indeed, one could suspect that notions like σ-algebras or measurability were introduced by mathematicians to scare normal folks away from their realms. This is not the case. Serious issues are avoided with those constructions. Mathematics is eternal: a once established result will be true also in thousands of years. A theory in which one could prove a theorem as well as its negation would be worthless: it would formally allow to prove any other result, whether true or false. So, these notions are not only introduced to keep the theory "clean", they are essential for the "survival" of the theory. We give some examples of "paradoxes" to illustrate the need for building a careful theory. Back to the fundamental notion of random variables: because they are just functions, one can add and multiply them by defining $(X + Y)(\omega) = X(\omega) + Y(\omega)$ or $(XY)(\omega) = X(\omega)Y(\omega)$. Random variables form so an **algebra** $\mathcal{L}$. The **expectation** of a random variable X is denoted by $\mathrm{E}[X]$ if it exists. It is a real number which indicates the "mean" or "average" of the observation X. It is the value, one would **expect** to measure in the experiment. If $X = 1_B$ is the random variable which has the value 1 if ω is in the event B and 0 if ω is not in the event B, then the expectation of X is just the probability of B. The constant random variable $X(\omega) = a$ has the expectation $\mathrm{E}[X] = a$. These two basic examples as well as the **linearity** requirement $\mathrm{E}[aX + bY] = a\mathrm{E}[X] + b\mathrm{E}[Y]$ determine the expectation for all random variables in the algebra $\mathcal{L}$: first one defines expectation for finite sums $\sum_{i=1}^{n} a_i 1_{B_i}$ called **elementary random variables**, which approximate general measurable functions. Extending the expectation to a subset $\mathcal{L}^1$ of the entire algebra is part of **integration theory**. While in calculus, one can live with the **Riemann integral** on the real line, which defines the integral by **Riemann sums** $\int_a^b f(x)\, dx \sim \frac{1}{n} \sum_{i/n \in [a,b]} f(i/n)$, the integral defined in measure theory is the **Lebesgue integral**. The later is more fundamental and probability theory is a major motivator for using it. It allows to make statements like that the probability of the set of real numbers with periodic decimal expansion has probability 0. In general, the probability of A is the expectation of the random variable $X(x) = f(x) = 1_A(x)$. In calculus, the integral $\int_0^1 f(x)\, dx$ would not be defined because a Riemann integral can give 1 or 0 depending on how the Riemann approximation is done. Probability theory allows to **introduce** the Lebesgue integral by defining $\int_a^b f(x)\, dx$ as the limit of $\frac{1}{n} \sum_{i=1}^{n} f(x_i)$ for $n \to \infty$, where x_i are **random uniformly distributed points** in the interval $[a, b]$. This **Monte Carlo definition** of the Lebesgue integral is based on the **law of large numbers** and is as intuitive

to state as the Riemann integral which is the limit of $\frac{1}{n}\sum_{x_j=j/n\in[a,b]} f(x_j)$ for $n \to \infty$.
With the fundamental notion of expectation one can define the **variance**, $\mathrm{Var}[X] = \mathrm{E}[X^2] - \mathrm{E}[X]^2$ and the **standard deviation** $\sigma[X] = \sqrt{\mathrm{Var}[X]}$ of a random variable X for which $X^2 \in \mathcal{L}^1$. One can also look at the **covariance** $\mathrm{Cov}[XY] = \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y]$ of two random variables X, Y for which $X^2, Y^2 \in \mathcal{L}^1$. The **correlation** $\mathrm{Corr}[X,Y] = \mathrm{Cov}[XY]/(\sigma[X]\sigma[Y])$ of two random variables with positive variance is a number which tells how much the random variable X is related to the random variable Y. If $\mathrm{E}[XY]$ is interpreted as an **inner product**, then the standard deviation is the **length** of $X - \mathrm{E}[X]$ and the correlation has the geometric interpretation as $\cos(\alpha)$, where α is the **angle** between the centered random variables $X - \mathrm{E}[X]$ and $Y - \mathrm{E}[Y]$. For example, if $\mathrm{Cov}[X,Y] = 1$, then $Y = \lambda X$ for some $\lambda > 0$, if $\mathrm{Cov}[X,Y] = -1$, they are anti-parallel. If the correlation is zero, the geometric interpretation is that the two random variables are **perpendicular**. Decorrelated random variables still can have relations to each other but if for any measurable real functions f and g, the random variables $f(X)$ and $g(X)$ are uncorrelated, then the random variables X, Y are **independent**.

A random variable X can be described well by its **distribution function** F_X. This is a real-valued function defined as $F_X(s) = \mathrm{P}[X \leq s]$ on $\mathbb{R}$, where $\{X \leq s\}$ is the event of all experiments ω satisfying $X(\omega) \leq s$. The distribution function does not encode the internal structure of the random variable X; it does not reveal the structure of the probability space for example. But the function F_X allows the construction of a probability space with exactly this distribution function. There are two important types of distributions, **continuous distributions** with a **probability density function** $f_X = F_X'$ and **discrete distributions** for which F is piecewise constant. An example of a continuous distribution is the **standard normal distribution**, where $f_X(x) = e^{-x^2/2}/\sqrt{2\pi}$. One can characterize it as the distribution with maximal **entropy** $I(f) = -\int \log(f(x))f(x)\,dx$ among all distributions which have zero mean and variance 1. An example of a discrete distribution is the **Poisson distribution** $\mathrm{P}[X = k] = e^{-\lambda}\frac{\lambda^k}{k!}$ on $\mathbb{N} = \{0, 1, 2, \dots\}$. One can describe random variables by their **moment generating functions** $M_X(t) = \mathrm{E}[e^{Xt}]$ or by their **characteristic function** $\phi_X(t) = \mathrm{E}[e^{iXt}]$. The later is the Fourier transform of the **law** $\mu_X = F_X'$ which is a measure on the real line $\mathbb{R}$.

The law μ_X of the random variable is a probability measure on the real line satisfying $\mu_X((a,b]) = F_X(b) - F_X(a)$. By the Lebesgue decomposition theorem, one can decompose any measure μ into a **discrete part** μ_{pp}, an absolutely continuous part μ_{ac} and a **singular continuous part** μ_{sc}. Random variables X for which μ_X is a discrete measure are called **discrete random variables**, random variables with a continuous law are called **continuous random variables**. Traditionally, these two type of random variables are the most important ones. But singular continuous random variables appear too: in spectral theory, dynamical systems or fractal geometry. Of course, the law of a random variable X does not need to be pure. It can mix the three types. A random variable can be mixed discrete and continuous for example.

Inequalities play an important role in probability theory. The **Chebychev inequality** $\mathrm{P}[|X - \mathrm{E}[X]| \geq c] \leq \frac{\mathrm{Var}[X]}{c^2}$ is used very often. It is a special case of the **Chebychev-Markov inequality** $h(c) \cdot \mathrm{P}[X \geq c] \leq \mathrm{E}[h(X)]$ for monotone nonnegative functions h. Other inequalities are the **Jensen inequality** $\mathrm{E}[h(X)] \geq h(\mathrm{E}[X])$ for convex functions h, the **Minkowski inequality** $||X + Y||_p \leq ||X||_p + ||Y||_p$ or the **Hölder inequality** $||XY||_1 \leq ||X||_p||Y||_q, 1/p + 1/q = 1$ for random variables, X, Y, for which $||X||_p = \mathrm{E}[|X|^p], ||Y||_q = \mathrm{E}[|Y|^q]$ are finite. Any inequality which appears in analysis can be useful in the toolbox of probability theory.

Independence is an central notion in probability theory. Two events A, B are called **independent**, if $\mathrm{P}[A \cap B] = \mathrm{P}[A] \cdot \mathrm{P}[B]$. An arbitrary set of events A_i is called independent, if for any finite subset of them, the probability of their intersection is the product of their probabilities. Two σ-algebras $\mathcal{A}, \mathcal{B}$ are called independent, if for any pair $A \in \mathcal{A}, B \in \mathcal{B}$, the events A, B are independent. Two random variables X, Y are independent, if they generate independent σ-algebras. It is enough to check that the events $A = \{X \in (a, b)\}$ and $B = \{Y \in (c, d)\}$ are independent for all intervals (a, b) and (c, d). One should think of independent random variables as two aspects of the laboratory Ω which do not influence each other. Each event $A = \{a < X(\omega) < b\ \}$ is independent of the event $B = \{c < Y(\omega) < d\ \}$. While the distribution function F_{X+Y} of the sum of two independent random variables is a **convolution** $\int_{\mathbb{R}} F_X(t-s)\ dF_Y(s)$, the **moment generating functions** and **characteristic functions** satisfy the formulas $M_{X+Y}(t) = M_X(t)M_Y(t)$ and $\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t)$. These identities make M_X, ϕ_X valuable tools to compute the distribution of an arbitrary finite sum of independent random variables.

Independence can also be explained using **conditional probability** with respect to an event B of positive probability: the conditional probability $\mathrm{P}[A|B] = \mathrm{P}[A \cap B]/\mathrm{P}[B]$ of A is the probability that A happens when we know that B takes place. If B is independent of A, then $\mathrm{P}[A|B] = \mathrm{P}[A]$ but in general, the conditional probability is larger. The notion of conditional probability leads to the important notion of **conditional expectation** $\mathrm{E}[X|\mathcal{B}]$ of a random variable X with respect to some sub-σ-algebra $\mathcal{B}$ of the σ algebra $\mathcal{A}$; it is a new random variable which is $\mathcal{B}$-measurable. For $\mathcal{B} = \mathcal{A}$, it is the random variable itself, for the trivial algebra $\mathcal{B} = \{\emptyset, \Omega\ \}$, we obtain the usual expectation $\mathrm{E}[X] = \mathrm{E}[X|\{\emptyset, \Omega\ \}]$. If $\mathcal{B}$ is generated by a finite partition $B_1, \dots, B_n$ of Ω of pairwise disjoint sets covering Ω, then $\mathrm{E}[X|\mathcal{B}]$ is piecewise constant on the sets B_i and the value on B_i is the average value of X on B_i. If $\mathcal{B}$ is the σ-algebra of an independent random variable Y, then $\mathrm{E}[X|Y] = \mathrm{E}[X|\mathcal{B}] = \mathrm{E}[X]$. In general, the conditional expectation with respect to $\mathcal{B}$ is a new random variable obtained by averaging on the elements of $\mathcal{B}$. One has $\mathrm{E}[X|Y] = h(Y)$ for some function h, extreme cases being $\mathrm{E}[X|1] = \mathrm{E}[X], \mathrm{E}[X|X] = X$. An illustrative example is the situation where $X(x,y)$ is a continuous function on the unit square with $\mathrm{P} = dxdy$ as a probability measure and where $Y(x,y) = x$. In that case, $\mathrm{E}[X|Y]$ is a function of x alone, given by $\mathrm{E}[X|Y](x) = \int_0^1 f(x,y)\, dy$. This is called **a conditional integral.**

A set $\{X_t\}_{t \in T}$ of random variables defines a **stochastic process.** The variable $t \in T$ is a parameter called "time". Stochastic processes are to probability theory what **differential equations** are to **calculus.** An example is a family X_n of random variables which evolve with **discrete time** $n \in \mathbb{N}$. Deterministic dynamical system theory branches into **discrete time systems**, the iteration of maps and **continuous time systems**, the theory of ordinary and partial **differential equations.** Similarly, in probability theory, one distinguishes between **discrete time stochastic processes** and **continuous time stochastic processes.** A discrete time stochastic process is a sequence of random variables X_n with certain properties. An important example is when X_n are independent, identically distributed random variables. A continuous time stochastic process is given by a family of random variables X_t, where t is **real time.** An example is a solution of a **stochastic differential equation.** With more general time like $\mathbb{Z}^d$ or $\mathbb{R}^d$ random variables are called **random fields** which play a role in statistical physics. Examples of such processes are **percolation** processes.

While one can realize every discrete time stochastic process X_n by a measure-preserving transformation $T : \Omega \to \Omega$ and $X_n(\omega) = X(T^n(\omega))$, probability theory often focuses a special subclass of systems called **martingales**, where one has a filtration $\mathcal{A}_n \subset \mathcal{A}_{n+1}$ of σ-algebras such that X_n is $\mathcal{A}_n$-measurable and $\mathrm{E}[X_n|\mathcal{A}_{n-1}] = X_{n-1}$, where $\mathrm{E}[X_n|\mathcal{A}_{n-1}]$ is the **conditional expectation** with respect to the sub-algebra $\mathcal{A}_{n-1}$. Martingales are a powerful generalization of the **random walk**, the process of summing up IID random variables with zero mean. Similar as ergodic theory, martingale theory is a natural extension of probability theory and has many applications.

The language of probability fits well into the **classical theory of dynamical systems.** For example, the **ergodic theorem of Birkhoff** for measure-preserving transformations has as a special case the **law of large numbers** which describes the average of partial sums of random variables $\frac{1}{n}\sum_{k=1}^{m} X_k$. There are different versions of the law of large numbers. "Weak laws" make statements about **convergence in probability**, "strong laws" make statements about **almost everywhere convergence.** There are versions of the law of large numbers for which the random variables do not need to have a common distribution and which go beyond Birkhoff's theorem. An other important theorem is the **central limit theorem** which shows that $S_n = X_1 + X_2 + \cdots + X_n$ normalized to have zero mean and variance 1 converges **in law** to the normal distribution or the **law of the iterated logarithm** which says that for centered independent and identically distributed X_k, the scaled sum S_n/Λ_n has accumulation points in the interval $[-\sigma, \sigma]$ if $\Lambda_n = \sqrt{2n \log\log n}$ and σ is the standard deviation of X_k. While stating the weak and strong law of large numbers and the central limit theorem, different convergence notions for random variables appear: **almost sure convergence** is the strongest, it implies **convergence in probability** and the later implies convergence **convergence in law**. There is also $\mathcal{L}^1$**-convergence** which is stronger than convergence in probability.

As in the deterministic case, where the **theory of differential equations** is more technical than the **theory of maps**, building up the formalism for **continuous time stochastic processes** X_t is more elaborate. Similarly as for differential equations, one has first to prove the existence of the objects. The most important continuous time stochastic process definitely is **Brownian motion** B_t. **Standard Brownian motion** is a stochastic process which satisfies $B_0 = 0$, $\mathrm{E}[B_t] = 0$, $\mathrm{Cov}[B_s, B_t] = s$ for $s \leq t$ and for any sequence of times, $0 = t_0 < t_1 < \cdots < t_i < t_{i+1}$, the increments $B_{t_{i+1}} - B_{t_i}$ are all independent random vectors with normal distribution. Brownian motion B_t is a solution of the **stochastic differential equation** $\frac{d}{dt}B_t = \zeta(t)$, where $\zeta(t)$ is called **white noise**. Because white noise is only defined as a **generalized function** and is not a stochastic process by itself, this stochastic differential equation has to be understood in its integrated form $B_t = \int_0^t dB_s = \int_0^t \zeta(s)\, ds$.

More generally, a solution to a stochastic differential equation $\frac{d}{dt}X_t = f(X_t)\zeta(t) + g(X_t)$ is defined as the solution to the integral equation $X_t = X_0 + \int_0^t f(X_s)\, dB_t + \int_0^t g(X_s)\, ds$. Stochastic differential equations can be defined in different ways. The expression $\int_0^t f(X_s)\, dB_t$ can either be defined as an **Ito integral**, which leads to martingale solutions, or the **Stratonovich integral**, which has similar integration rules than classical differentiation equations. Examples of stochastic differential equations are $\frac{d}{dt}X_t = X_t\zeta(t)$ which has the solution $X_t = e^{B_t - t/2}$. Or $\frac{d}{dt}X_t = B_t^4\zeta(t)$ which has as the solution the process $X_t = B_t^5 - 10B_t^3 + 15B_t$. The key tool to solve stochastic differential equations is **Ito's formula** $f(B_t) - f(B_0) = \int_0^t f'(B_s)dB_s + \frac{1}{2}\int_0^t f''(B_s)\, ds$, which is the stochastic analog of the fundamental theorem of calculus. Solutions to stochastic differential equations are examples of **Markov processes** which show diffusion. Especially, the solutions can be used to solve classical partial differential equations like the **Dirichlet problem** $\Delta u = 0$ in a bounded domain D with $u = f$ on the boundary δD. One can get the solution by computing the expectation of f at the end points of Brownian motion starting at x and ending at the boundary $u = \mathrm{E}_x[f(B_T)]$. On a discrete graph, if Brownian motion is replaced by random walk, the same formula holds too. Stochastic calculus is also useful to interpret quantum mechanics as a **diffusion processes** [73, 71] or as a tool to compute solutions to quantum mechanical problems using **Feynman-Kac formulas**.

Some features of stochastic process can be described using the language of **Markov operators** P, which are positive and expectation-preserving transformations on $\mathcal{L}^1$. Examples of such operators are **Perron-Frobenius operators** $X \to X(T)$ for a measure preserving transformation T defining a discrete time evolution or stochastic matrices describing a random walk on a finite graph. Markov operators can be defined by **transition probability functions** which are measure-valued random variables. The interpretation is that from a given point ω, there are different possibilities to go to. A **transition probability measure** $\mathcal{P}(\omega, \cdot)$ gives the distribution of the target. The relation with Markov operators is assured by the **Chapman-Kolmogorov equation** $\mathrm{P}^{n+m} = \mathrm{P}^n \circ \mathrm{P}^m$. Markov processes can be obtained from **random transformations, random walks** or by **stochastic differential equations**. In the case of a finite or countable target space S, one obtains **Markov chains** which can be described by **probability matrices** P, which are the simplest Markov operators. For Markov operators, there is an **arrow of time**: the **relative entropy** with respect to a background measure is non-increasing. Markov processes often are attracted by fixed points of the Markov operator. Such fixed points are called **stationary states**. They describe **equilibria** and often they are measures with maximal entropy. An example is the Markov operator P, which assigns to a probability density f_Y the probability density of $f_{\overline{Y+X}}$ where $\overline{Y+X}$ is the random variable $Y+X$ normalized so that it has mean 0 and variance 1. For the initial function $f=1$, the function $P^n(f_X)$ is the distribution of S^*_n the normalized sum of n IID random variables X_i. This Markov operator has a unique equilibrium point, the **standard normal distribution**. It has maximal entropy among all distributions on the real line with variance 1 and mean 0. The central limit theorem tells that the Markov operator P has the normal distribution as a unique attracting fixed point if one takes the weaker topology of convergence in distribution on $\mathcal{L}^1$. This works in other situations too. For **circle-valued random variables** for example, the uniform distribution maximizes entropy. It is not surprising therefore, that there is a central limit theorem for circle-valued random variables with the uniform distribution as the limiting distribution.

In the same way as mathematics reaches out into other scientific areas, probability theory has connections with many other branches of mathematics. The last chapter of these notes give some examples. The section on **percolation** shows how probability theory can help to understand critical phenomena. In solid state physics, one considers **operator-valued random variables**. The spectrum of random operators are random objects too. One is interested what happens with probability one. **Localization** is the phenomenon in solid state physics that sufficiently random operators often have pure point spectrum. The section on **estimation theory** gives a glimpse of what mathematical statistics is about. In statistics one often does not know the probability space itself so that one has to make a **statistical model** and look at a parameterization of probability spaces. The goal is to give **maximum likelihood estimates** for the parameters from data and to understand how small the quadratic estimation error can be made. A section on **Vlasov dynamics** shows how probability theory appears in problems of **geometric evolution**. Vlasov dynamics is a generalization of the n-body problem to the evolution of of probability measures. One can look at the evolution of smooth measures or measures located on surfaces. This deterministic stochastic system produces an evolution of densities which can form singularities without doing harm to the formalism. It also defines the evolution of surfaces. The section on moment problems is part of **multivariate statistics**. As for random variables, random vectors can be described by their **moments**. Since moments define the law of the random variable, the question arises how one can see from the moments, whether we have a continuous random variable. The section of **random maps** is an other part of dynamical systems theory. Randomized versions of diffeomorphisms can be considered idealization of their undisturbed versions. They often can be understood better than their deterministic versions. For example, many random diffeomorphisms have only finitely many ergodic components. In the section in **circular random variables**, we see that the **Mises** distribution has extremal entropy among all circle-valued random variables with given circular mean and variance. There is also a central limit theorem on the circle: the sum of IID circular random variables converges in law to the uniform distribution. We then look at a problem in the **geometry of numbers**: how many lattice points are there in a neighborhood of the graph of one-dimensional Brownian motion? The analysis of this problem needs a law of large numbers for independent random variables X_k with uniform distribution on $[0, 1]$: for $0 \leq \delta < 1$, and $A_n = [0, 1/n^\delta]$ one has $\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^n \frac{1_{A_n}(X_k)}{n^\delta} = 1$. Probability theory also matters in complexity theory as a section on **arithmetic random variables** shows. It turns out that random variables like $X_n(k) = k$, $Y_n(k) = k^2 + 3 \bmod n$ defined on finite probability spaces become independent in the limit $n \to \infty$. Such considerations matter in **complexity** theory: arithmetic functions defined on large but finite sets behave very much like random functions. This is reflected by the fact that the inverse of arithmetic functions is in general difficult to compute and belong to the complexity class of NP. Indeed, if one could invert arithmetic functions easily, one could solve problems like factoring integers fast. A short section on **Diophantine equations** indicates how the distribution of random variables can shed light on the solution of **Diophantine equations**. Finally, we look at a topic in **harmonic analysis** which was initiated by Norbert Wiener. It deals with the relation of the characteristic function ϕ_X and the continuity properties of the random variable X.

1.2 Some paradoxes in probability theory

Colloquial language is not always precise enough to tackle problems in probability theory. Paradoxes appear, when definitions allow different interpretations. Ambiguous language can lead to wrong conclusions or contradicting solutions. To illustrate this, we mention a few problems. The following four examples should serve as a motivation to introduce probability theory on a rigorous mathematical footing.

1) **Bertrand's paradox (Bertrand 1889)**
We throw at random lines onto the unit disc. What is the probability that

the line intersects the disc with a length $\geq \sqrt{3}$, the length of the inscribed equilateral triangle?

First answer: take an arbitrary point P on the boundary of the disc. The set of all lines through that point are parameterized by an angle ϕ. In order that the chord is longer than $\sqrt{3}$, the line has to lie within a sector of 60° within a range of 180°. The probability is $1/3$.

Second answer: take all lines perpendicular to a fixed diameter. The chord is longer than $\sqrt{3}$ if the point of intersection lies on the middle half of the diameter. The probability is $1/2$.

Third answer: if the midpoints of the chords lie in a disc of radius $1/2$, the chord is longer than $\sqrt{3}$. Because the disc has a radius which is half the radius of the unit disc, the probability is $1/4$.

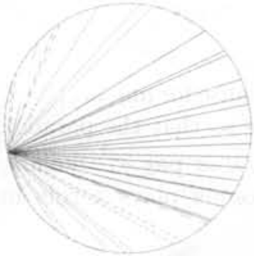


Figure. *Random angle.*

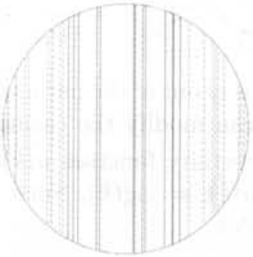


Figure. *Random translation.*

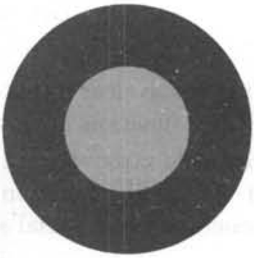


Figure. *Random area.*

Like most paradoxes in mathematics, a part of the question in Bertrand's problem is not well defined. Here it is the term "random line". The solution of the paradox lies in the fact that the three answers depend on the **chosen probability distribution**. There are several "natural" distributions. The actual answer depends on how the experiment is performed.

2) **Petersburg paradox (D.Bernoulli, 1738)**
In the Petersburg casino, you pay an entrance fee c and you get the prize 2^T, where T is the number of times, the casino flips a coin until "head" appears. For example, if the sequence of coin experiments would give "tail, tail, tail, head", you would win $2^3 - c = 8 - c$, the win minus the entrance fee. Fair would be an entrance fee which is equal to the expectation of the win, which is

$$\sum_{k=1}^{\infty} 2^k \mathrm{P}[T = k] = \sum_{k=1}^{\infty} 1 = \infty .$$

The paradox is that nobody would agree to pay even an entrance fee $c = 10$.

The problem with this casino is that it is not quite clear, what is "fair". For example, the situation $T = 20$ is so improbable that it never occurs in the life-time of a person. Therefore, for any practical reason, one has not to worry about large values of T. This, as well as the finiteness of money resources is the reason, why casinos do not have to worry about the following bullet proof **martingale strategy** in roulette: bet c dollars on red. If you win, stop, if you lose, bet $2c$ dollars on red. If you win, stop. If you lose, bet $4c$ dollars on red. Keep doubling the bet. Eventually after n steps, red will occur and you will win $2^n c - (c + 2c + \cdots + 2^{n-1}c) = c$ dollars. This example motivates the concept of martingales. Theorem (3.2.7) or proposition (3.2.9) will shed some light on this. Back to the Petersburg paradox. How does one resolve it? What would be a reasonable entrance fee in "real life"? Bernoulli proposed to replace the expectation $\mathrm{E}[G]$ of the profit $G = 2^T$ with the expectation $(\mathrm{E}[\sqrt{G}])^2$, where $u(x) = \sqrt{x}$ is called a **utility function**. This would lead to a fair entrance

$$(\mathrm{E}[\sqrt{G}])^2 = (\sum_{k=1}^{\infty} 2^{k/2}2^{-k})^2 = \frac{1}{(\sqrt{2}-1)^2} \sim 5.828\ldots .$$

It is not so clear if that is a way out of the paradox because for any proposed utility function $u(k)$, one can modify the casino rule so that the paradox reappears: pay $(2^k)^2$ if the utility function $u(k) = \sqrt{k}$ or pay e^{2^k} dollars, if the utility function is $u(k) = \log(k)$. Such reasoning plays a role in economics and social sciences.

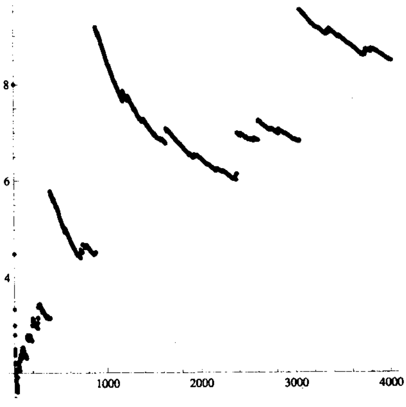


Figure. *The picture to the right shows the average profit development during a typical tournament of 4000 Petersburg games. After these 4000 games, the player would have lost about 10 thousand dollars, when paying a 10 dollar entrance fee each game. The player would have to play a very, very long time to catch up. Mathematically, the player will do so and have a profit in the long run, but it is unlikely that it will happen in his or her life time.*

3) **The three door problem (1991)** Suppose you're on a game show and you are given a choice of three doors. Behind one door is a car and behind the others are goats. You pick a door-say No. 1 - and the host, who knows what's behind the doors, opens another door-say, No. 3-which has a goat. (In all games, he opens a door to reveal a goat). He then says to you, "Do

you want to pick door No. 2?" (In all games he always offers an option to switch). Is it to your advantage to switch your choice?

The problem is also called "Monty Hall problem" and was discussed by Marilyn vos Savant in a "Parade" column in 1991 and provoked a big controversy. (See [98] for pointers and similar examples.) The problem is that intuitive argumentation can easily lead to the conclusion that it does not matter whether to change the door or not. Switching the door doubles the chances to win:

No switching: you choose a door and win with probability $1/3$. The opening of the host does not affect any more your choice.
Switching: when choosing the door with the car, you loose since you switch. If you choose a door with a goat. The host opens the other door with the goat and you win. There are two such cases, where you win. The probability to win is $2/3$.

4) **The Banach-Tarski paradox (1924)**
It is possible to cut the standard unit ball $\Omega = \{x \in \mathbb{R}^3 \mid |x| \leq 1 \}$ into 5 disjoint pieces $\Omega = Y_1 \cup Y_2 \cup Y_3 \cup Y_4 \cup Y_5$ and rotate and translate the pieces with transformations T_i so that $T_1(Y_1) \cup T_2(Y_2) = \Omega$ and $T_3(Y_3) \cup T_4(Y_4) \cup T_5(Y_5) = \Omega'$ is a second unit ball $\Omega' = \{x \in \mathbb{R}^3 \mid |x - (3,0,0)| \leq 1\}$ and all the transformed sets again don't intersect.
While this example of Banach-Tarski is spectacular, the existence of bounded subsets A of the circle for which one can not assign a translational invariant probability $\mathrm{P}[A]$ can already be achieved in one dimension. The Italian mathematician **Giuseppe Vitali** gave in 1905 the following example: define an equivalence relation on the circle $\mathbb{T} = [0, 2\pi)$ by saying that two angles are **equivalent** $x \sim y$ if $(x-y)/\pi$ is a rational angle. Let A be a subset in the circle which contains exactly one number from each equivalence class. The **axiom of choice** assures the existence of A. If $x_1, x_2, \dots$ is a enumeration of the set of rational angles in the circle, then the sets $A_i = A + x_i$ are pairwise disjoint and satisfy $\bigcup_{i=1}^\infty A_i = \mathbb{T}$. If we could assign a translational invariant probability $\mathrm{P}[A_i]$ to A, then the basic rules of probability would give

$$1 = \mathrm{P}[\mathbb{T}] = \mathrm{P}[\bigcup_{i=1}^\infty A_i] = \sum_{i=1}^\infty \mathrm{P}[A_i] = \sum_{i=1}^\infty p \, .$$

But there is no real number $p = \mathrm{P}[A] = \mathrm{P}[A_i]$ which makes this possible. Both the Banach-Tarski as well as Vitalis result shows that one can not hope to define a probability space on the algebra $\mathcal{A}$ of **all** subsets of the unit ball or the unit circle such that the probability measure is translational and rotational invariant. The natural concepts of "length" or "volume", which are rotational and translational invariant only makes sense for a smaller algebra. This will lead to the notion of σ-algebra. In the context of topological spaces like Euclidean spaces, it leads to **Borel σ-algebras**, algebras of sets generated by the compact sets of the topological space. This language will be developed in the next chapter.

1.3 Some applications of probability theory

Probability theory is a central topic in mathematics. There are close relations and intersections with other fields like computer science, ergodic theory and dynamical systems, cryptology, game theory, analysis, partial differential equation, mathematical physics, economical sciences, statistical mechanics and even number theory. As a motivation, we give some problems and topics which can be treated with probabilistic methods.

1) **Random walks**: (statistical mechanics, gambling, stock markets, quantum field theory).

Assume you walk through a lattice. At each vertex, you choose a direction at random. What is the probability that you return back to your starting point? Polya's theorem (3.8.1) says that in two dimensions, a random walker almost certainly returns to the origin arbitrarily often, while in three dimensions, the walker with probability 1 only returns a finite number of times and then escapes for ever.

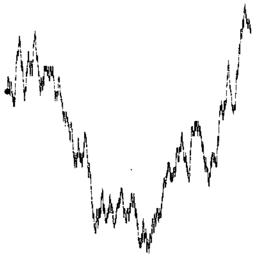


Figure. *A random walk in one dimensions displayed as a graph (t, B_t).*

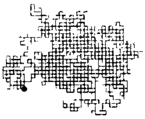


Figure. *A piece of a random walk in two dimensions.*

Figure. *A piece of a random walk in three dimensions.*

2) **Percolation problems** (model of a porous medium, statistical mechanics, critical phenomena).

Each bond of a rectangular lattice in the plane is connected with probability p and disconnected with probability $1-p$. Two lattice points x, y in the lattice are in the same **cluster**, if there is a path from x to y. One says that **"percolation occurs"** if there is a positive probability that an infinite cluster appears. One problem is to find the **critical probability** p_c, the infimum of all p, for which percolation occurs. The problem can be extended to situations, where the switch probabilities are not independent to each other. Some random variables like the size of the largest cluster are of interest near the critical probability p_c.

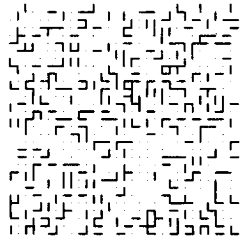


Figure. *Bond percolation with p=0.2.*

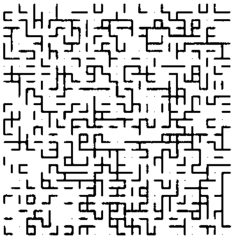


Figure. *Bond percolation with p=0.4.*

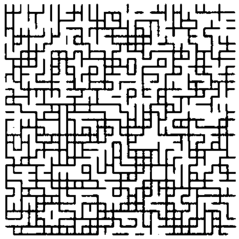


Figure. *Bond percolation with p=0.6.*

A variant of bond percolation is **site percolation** where the nodes of the lattice are switched on with probability p.

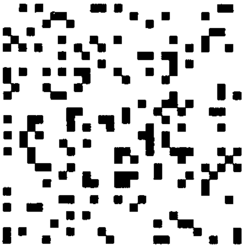


Figure. *Site percolation with p=0.2.*

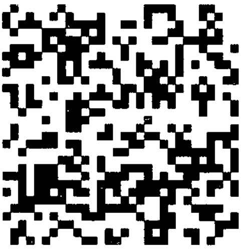


Figure. *Site percolation with p=0.4.*

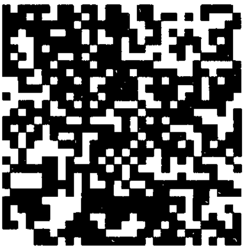


Figure. *Site percolation with p=0.6.*

Generalized percolation problems are obtained, when the independence of the individual nodes is relaxed. A class of such **dependent percolation** problems can be obtained by choosing two irrational numbers α, β like $\alpha = \sqrt{2} - 1$ and $\beta = \sqrt{3} - 1$ and switching the node (n, m) on if $(n\alpha + m\beta) \bmod 1 \in [0, p)$. The probability of switching a node on is again p, but the random variables

$$X_{n,m} = 1_{(n\alpha+m\beta) \bmod 1 \in [0,p)}$$

are no more independent.

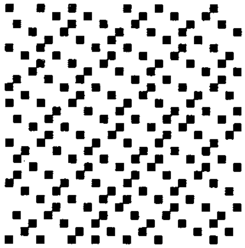


Figure. *Dependent site percolation with p=0.2.*

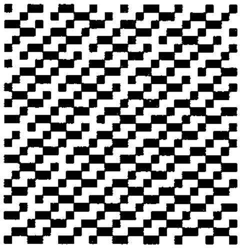


Figure. *Dependent site percolation with p=0.4.*

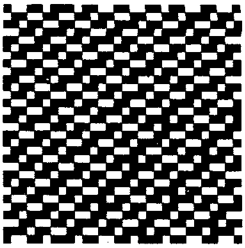


Figure. *Dependent site percolation with p=0.6.*

Even more general percolation problems are obtained, if also the distribution of the random variables $X_{n,m}$ can depend on the position (n, m).

3) **Random Schrödinger operators.** (quantum mechanics, functional analysis, disordered systems, solid state physics)

Consider the linear map $Lu(n) = \sum_{|m-n|=1} u(n) + V(n)u(n)$ on the space of sequences $u = (\ldots, u_{-2}, u_{-1}, u_0, u_1, u_2, \ldots)$. We assume that $V(n)$ takes random values in $\{0, 1\}$. The function V is called the potential. The problem is to determine the spectrum or spectral type of the infinite matrix L on the **Hilbert space** l^2 of all sequences u with finite $||u||_2^2 = \sum_{n=-\infty}^{\infty} u_n^2$. The operator L is the Hamiltonian of an electron in a one-dimensional disordered crystal. The spectral properties of L have a relation with the **conductivity** properties of the crystal. Of special interest is the situation, where the values $V(n)$ are all independent random variables. It turns out that if $V(n)$ are IID random variables with a continuous distribution, there are many eigenvalues for the infinite dimensional matrix L - at least with probability 1. This phenomenon is called **localization**.

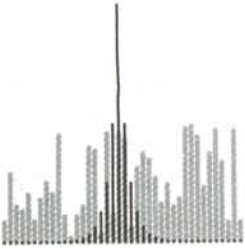


Figure. *A wave $\psi(t) = e^{iLt}\psi(0)$ evolving in a random potential at $t = 0$. Shown are both the potential V_n and the wave $\psi(0)$.*

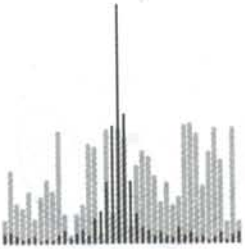


Figure. *A wave $\psi(t) = e^{iLt}\psi(0)$ evolving in a random potential at $t = 1$. Shown are both the potential V_n and the wave $\psi(1)$.*

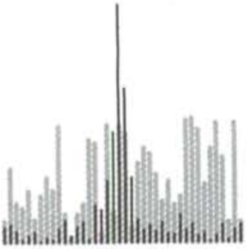


Figure. *A wave $\psi(t) = e^{iLt}\psi(0)$ evolving in a random potential at $t = 2$. Shown are both the potential V_n and the wave $\psi(2)$.*

More general operators are obtained by allowing $V(n)$ to be random variables with the same distribution but where one does not persist on independence any more. A well studied example is the **almost Mathieu operator**, where $V(n) = \lambda\cos(\theta + n\alpha)$ and for which $\alpha/(2\pi)$ is irrational.

4) **Classical dynamical systems** (celestial mechanics, fluid dynamics, mechanics, population models)

The study of deterministic dynamical systems like the **logistic map** $x \mapsto 4x(1-x)$ on the interval $[0, 1]$ or the **three body problem** in celestial mechanics has shown that such systems or subsets of it can behave like random systems. Many effects can be described by **ergodic theory**, which can be seen as a brother of probability theory. Many results in probability theory generalize to the more general setup of ergodic theory. An example is **Birkhoff's ergodic theorem** which generalizes the law of large numbers.

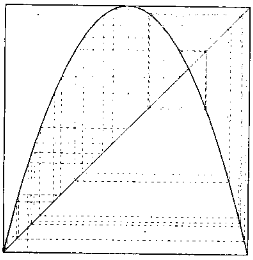


Figure. *Iterating the logistic map*

$$T(x) = 4x(1-x)$$

on $[0,1]$ *produces independent random variables. The invariant measure* P *is continuous.*

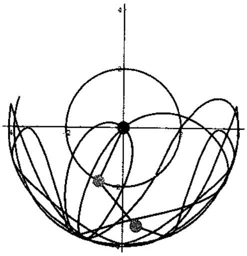


Figure. *The simple mechanical system of a double pendulum exhibits complicated dynamics. The differential equation defines a measure preserving flow* T_t *on a probability space.*

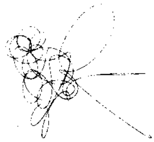


Figure. *A short time evolution of the Newtonian three body problem. There are energies and subsets of the energy surface which are invariant and on which there is an invariant probability measure.*

Given a dynamical system given by a map T or a flow T_t on a subset Ω of some Euclidean space, one obtains for every invariant probability measure P a probability space $(\Omega, \mathcal{A}, \mathrm{P})$. An observed quantity like a coordinate of an individual particle is a random variable X and defines a stochastic process $X_n(\omega) = X(T^n\omega)$. For many dynamical systems including also some 3 body problems, there are invariant measures and observables X for which X_n are IID random variables. Probability theory is therefore intrinsically relevant also in classical dynamical systems.

5) **Cryptology**. (computer science, coding theory, data encryption)

Coding theory deals with the mathematics of encrypting codes or deals with the design of error correcting codes. Both aspects of coding theory have important applications. A good code can repair loss of information due to bad channels and hide the information in an encrypted way. While many aspects of coding theory are based in discrete mathematics, number theory, algebra and algebraic geometry, there are probabilistic and combinatorial aspects to the problem. We illustrate this with the example of a public key encryption algorithm whose security is based on the fact that it is hard to factor a large integer $N = pq$ into its prime factors p, q but easy to verify that p, q are factors, if one knows them. The number N can be public but only the person, who knows the factors p, q can read the message. Assume, we want to crack the code and find the factors p and q.

The simplest method is to try to find the factors by trial and error but this is impractical already if N has 50 digits. We would have to search through 10^{25} numbers to find the factor p. This corresponds to probe 100 million times every second over a time span of 15 billion years. There are better methods known and we want to illustrate one of them now: assume we want to find the factors of $N = 1111111111111111111111111111111111111111111111$. The method goes as follows: start with an integer a and iterate the quadratic map $T(x) = x^2 + c \bmod N$ on $\{0, 1., , , .N-1\}$. If we assume the numbers $x_0 = a, x_1 = T(a), x_2 = T(T(a)) \dots$ to be random, how many such numbers do we have to generate, until two of them are the same modulo one of the prime factors p? The answer is surprisingly small and based on the **birthday paradox**: the probability that in a group of 23 students, two of them have the same birthday is larger than 1/2: the probability of the event that we have no birthday match is $1(364/365)(363/365)\cdots(343/365) = 0.492703\dots$, so that the probability of a birthday match is $1 - 0.492703 = 0.507292$. This is larger than 1/2. If we apply this thinking to the sequence of numbers x_i generated by the **pseudo random number generator** T, then we expect to have a chance of 1/2 for finding a match modulo p in $\sqrt{p}$ iterations. Because $p \leq \sqrt{n}$, we have to try $N^{1/4}$ numbers, to get a factor: if x_n and x_m are the same modulo p, then $\gcd(x_n - x_m, N)$ produces the factor p of N. In the above example of the 46 digit number N, there is a prime factor $p = 35121409$. The Pollard algorithm finds this factor with probability 1/2 in $\sqrt{p} = 5926$ steps. This is an estimate only which gives the order of magnitude. With the above N, if we start with $a = 17$ and take $a = 3$, then we have a match $x_{27720} = x_{13860}$. It can be found very fast.

This probabilistic argument would give a rigorous probabilistic estimate if we would pick truly random numbers. The algorithm of course generates such numbers in a **deterministic** way and they are not truly random. The generator is called a **pseudo random number generator**. It produces numbers which are random in the sense that many statistical tests can not distinguish them from true random numbers. Actually, many random number generators built into computer operating systems and programming languages are pseudo random number generators.

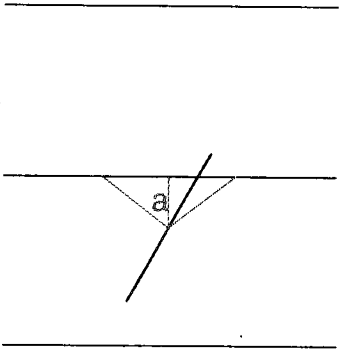
Probabilistic thinking is often involved in designing, investigating and attacking data encryption codes or random number generators.

6) **Numerical methods**. (integration, Monte Carlo experiments, algorithms) In applied situations, it is often very difficult to find integrals directly. This happens for example in statistical mechanics or quantum electrodynamics, where one wants to find integrals in spaces with a large number of dimensions. One can nevertheless compute numerical values using **Monte Carlo Methods** with a manageable amount of effort. Limit theorems assure that these numerical values are reasonable. Let us illustrate this with a very simple but famous example, the **Buffon needle problem**.

A **stick** of length 2 is thrown onto the plane filled with parallel lines, all of which are distance $d = 2$ apart. If the center of the stick falls within distance y of a line, then the interval of angles leading to an intersection with a grid line has length $2\arccos(y)$ among a possible range of angles

$[0, \pi]$. The probability of hitting a line is therefore $\int_0^1 2\arccos(y)/\pi = 2/\pi$. This leads to a Monte Carlo method to compute π. Just throw randomly n sticks onto the plane and count the number k of times, it hits a line. The number $2n/k$ is an approximation of π. This is of course not an effective way to compute π but it illustrates the principle.

Figure. *The Buffon needle problem is a Monte Carlo method to compute π. By counting the number of hits in a sequence of experiments, one can get random approximations of π. The* **law of large numbers** *assures that the approximations will converge to the expected limit. All Monte Carlo computations are theoretically based on limit theorems.*



Chapter 2

Limit theorems

2.1 Probability spaces, random variables, independence

In this section we define the basic notions of a "probability space" and "random variables" on an arbitrary set Ω.

Definition. A set $\mathcal{A}$ of subsets of Ω is called a **σ-algebra** if the following three properties are satisfied:

$$\boxed{\begin{array}{l} \text{(i) } \Omega \in \mathcal{A}, \\ \text{(ii) } A \in \mathcal{A} \Rightarrow A^c = \Omega \setminus A \in \mathcal{A}, \\ \text{(iii) } A_n \in \mathcal{A} \Rightarrow \bigcup_{n \in \mathbb{N}} A_n \in \mathcal{A} \end{array}}$$

A pair $(\Omega, \mathcal{A})$ for which $\mathcal{A}$ is a σ-algebra in Ω is called a **measurable space**.

Properties. If $\mathcal{A}$ is a σ-algebra, and A_n is a sequence in $\mathcal{A}$, then the following properties follow immediately by checking the axioms:
1) $\bigcap_{n \in \mathbb{N}} A_n \in \mathcal{A}$.
2) $\limsup_n A_n := \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} A_n \in \mathcal{A}$.
3) $\liminf_n A_n := \bigcup_{n=1}^{\infty} \bigcap_{m=n}^{\infty} A_n \in \mathcal{A}$.
4) $\mathcal{A}, \mathcal{B}$ are algebras, then $\mathcal{A} \cap \mathcal{B}$ is an algebra.
5) If $\{\mathcal{A}_\lambda\}_{i \in I}$ is a family of σ- sub-algebras of $\mathcal{A}$. then $\bigcap_{i \in I} \mathcal{A}_i$ is a σ-algebra.

Example. For an arbitrary set Ω, $\mathcal{A} = \{\emptyset, \Omega\})$ is a σ-algebra. It is called the **trivial** σ-algebra.

Example. If Ω is an arbitrary set, then $\mathcal{A} = \{A \subset \Omega\})$ is a σ-algebra. The set of all subsets of Ω is the largest σ-algebra one can define on a set.

Example. A finite set of subsets $A_1, A_2, \dots, A_n$ of Ω which are pairwise disjoint and whose union is Ω, it is called a **partition** of Ω. It generates the σ-algebra: $\mathcal{A} = \{A = \bigcup_{j \in J} A_j\ \}$ where J runs over all subsets of $\{1, .., n\}$. This σ-algebra has 2^n elements. Every finite σ-algebra is of this form. The smallest nonempty elements $\{A_1, \dots, A_n\}$ of this algebra are called **atoms**.

Definition. For any set $\mathcal{C}$ of subsets of Ω, we can define $\sigma(\mathcal{C})$, the smallest σ-algebra $\mathcal{A}$ which contains $\mathcal{C}$. The σ-algebra $\mathcal{A}$ is the intersection of all σ-algebras which contain $\mathcal{C}$. It is again a σ-algebra.

Example. For $\Omega = \{1, 2, 3\}$, the set $\mathcal{C} = \{\{1, 2\}, \{2, 3\ \}\}$ generates the σ-algebra $\mathcal{A}$ which consists of all 8 subsets of Ω.

Definition. If $(E, \mathcal{O})$ is a topological space, where $\mathcal{O}$ is the set of **open sets** in E. then $\sigma(\mathcal{O})$ is called the **Borel σ-algebra** of the topological space. If $\mathcal{A} \subset \mathcal{B}$, then $\mathcal{A}$ is called a **subalgebra** of $\mathcal{B}$. A set B in $\mathcal{B}$ is also called a **Borel set**.

Remark. One sometimes defines the Borel σ-algebra as the σ-algebra generated by the set of **compact sets** $\mathcal{C}$ of a topological space. Compact sets in a topological space are sets for which every open cover has a finite subcover. In Euclidean spaces $\mathbb{R}^n$, where compact sets coincide with the sets which are both bounded and closed, the Borel σ-algebra generated by the compact sets is the same as the one generated by open sets. The two definitions agree for a large class of topological spaces like "locally compact separable metric spaces".

Remark. Often, the Borel σ-algebra is enlarged to the σ-algebra of all **Lebesgue measurable** sets, which includes all sets B which are a subset of a Borel set A of measure 0. The smallest σ-algebra $\overline{\mathcal{B}}$ which contains all these sets is called the **completion** of $\mathcal{B}$. The completion of the Borel σ-algebra is the σ-algebra of all Lebesgue measurable sets. It is in general strictly larger than the Borel σ-algebra. But it can also have pathological features like that the composition of a Lebesgue measurable function with a continuous functions does not need to be Lebesgue measurable any more. (See [109], Example 2.4).

Example. The σ-algebra generated by the **open balls** $\mathcal{C} = \{A = B_r(x)\ \}$ of a metric space (X, d) need not to agree with the family of Borel subsets, which are generated by $\mathcal{O}$, the set of **open sets** in (X, d).
Proof. Take the metric space $(\mathbb{R}, d)$ where $d(x, y) = 1_{\{x=y\}}$ is the **discrete metric**. Because any subset of $\mathbb{R}$ is open, the Borel σ-algebra is the set of all subsets of $\mathbb{R}$. The open balls in $\mathbb{R}$ are either single points or the whole space. The σ-algebra generated by the open balls is the set of countable subset of $\mathbb{R}$ together with their complements.

Example. If $\Omega = [0,1] \times [0,1]$ is the unit square and $\mathcal{C}$ is the set of all sets of the form $[0,1] \times [a,b]$ with $0 < a < b < 1$, then $\sigma(\mathcal{C})$ is the σ-algebra of all sets of the form $[0,1] \times A$, where A is in the Borel σ-algebra of $[0,1]$.

Definition. Given a measurable space $(\Omega, \mathcal{A})$. A function $\mathrm{P} : \mathcal{A} \to \mathbb{R}$ is called a **probability measure** and $(\Omega, \mathcal{A}, \mathrm{P})$ is called a **probability space** if the following three properties called **Kolmogorov axioms** are satisfied:

(i) $\mathrm{P}[A] \geq 0$ for all $A \in \mathcal{A}$,
(ii) $\mathrm{P}[\Omega] = 1$,
(iii) $A_n \in \mathcal{A}$ disjoint $\Rightarrow \mathrm{P}[\bigcup_n A_n] = \sum_n \mathrm{P}[A_n]$

The last property is called **σ-additivity**.

Properties. Here are some basic properties of the probability measure which immediately follow from the definition:
1) $\mathrm{P}[\emptyset] = 0$.
2) $A \subset B \Rightarrow \mathrm{P}[A] \leq \mathrm{P}[B]$.
3) $\mathrm{P}[\bigcup_n A_n] \leq \sum_n \mathrm{P}[A_n]$.
4) $\mathrm{P}[A^c] = 1 - \mathrm{P}[A]$.
5) $0 \leq \mathrm{P}[A] \leq 1$.
6) $A_1 \subset A_2, \subset \cdots$ with $A_n \in \mathcal{A}$ then $\mathrm{P}[\bigcup_{n=1}^{\infty} A_n] = \lim_{n\to\infty} \mathrm{P}[A_n]$.

Remark. There are different ways to build the axioms for a probability space. One could for example replace (i) and (ii) with properties 4),5) in the above list. Statement 6) is equivalent to σ-additivity if P is only assumed to be additive.

Remark. The name "Kolmogorov axioms" honors a monograph of Kolmogorov from 1933 [53] in which an axiomatization appeared. Other mathematicians have formulated similar axiomatizations at the same time, like Hans Reichenbach in 1932. According to Doob, axioms (i)-(iii) were first proposed by G. Bohlmann in 1908 [22].

Definition. A map X from a measure space $(\Omega, \mathcal{A})$ to an other measure space $(\Delta, \mathcal{B})$ is called **measurable**, if $X^{-1}(B) \in \mathcal{A}$ for all $B \in \mathcal{B}$. The set $X^{-1}(B)$ consists of all points $x \in \Omega$ for which $X(x) \in B$. This **pull back set** $X^{-1}(B)$ is defined even if X is non-invertible. For example, for $X(x) = x^2$ on $(\mathbb{R}, \mathcal{B})$ one has $X^{-1}([1,4]) = [1,2] \cup [-2,-1]$.

Definition. A function $X : \Omega \to \mathbb{R}$ is called a **random variable**, if it is a measurable map from $(\Omega, \mathcal{A})$ to $(\mathbb{R}, \mathcal{B})$, where $\mathcal{B}$ is the Borel σ-algebra of

$\mathbb{R}$. Denote by $\mathcal{L}$ the set of all real random variables. The set $\mathcal{L}$ is an **algebra** under addition and multiplication: one can add and multiply random variables and gets new random variables. More generally, one can consider random variables taking values in a second measurable space $(E, \mathcal{B})$. If $\mathrm{E} = \mathbb{R}^d$, then the random variable X is called a **random vector**. For a random vector $X = (X_1, \dots, X_d)$, each component X_i is a random variable.

Example. Let $\Omega = \mathbb{R}^2$ with Borel σ-algebra $\mathcal{A}$ and let

$$\mathrm{P}[A] = \frac{1}{2\pi} \int \int_A e^{-(x^2-y^2)/2} \, dxdy \, .$$

Any continuous function X of two variables is a random variable on Ω. For example, $X(x, y) = xy(x + y)$ is a random variable. But also $X(x, y) = 1/(x + y)$ is a random variable, even so it is not continuous. The vector-valued function $X(x, y) = (x, y, x^3)$ is an example of a random vector.

Definition. Every random variable X defines a σ-algebra

$$X^{-1}(\mathcal{B}) = \{X^{-1}(B) \mid B \in \mathcal{B}\} \, .$$

We denote this algebra by $\sigma(X)$ and call it the **σ-algebra generated by** X.

Example. A constant map $X(x) = c$ defines the trivial algebra $\mathcal{A} = \{\emptyset, \Omega\}$.

Example. The map $X(x, y) = x$ from the square $\Omega = [0, 1] \times [0, 1]$ to the real line $\mathbb{R}$ defines the algebra $\mathcal{B} = \{A \times [0, 1]\}$, where A is in the Borel σ-algebra of the interval $[0, 1]$.

Example. The map X from $\mathbb{Z}_6 = \{0, 1, 2, 3, 4, 5\}$ to $\{0, 1\} \subset \mathbb{R}$ defined by $X(x) = x \bmod 2$ has the value $X(x) = 0$ if x is even and $X(x) = 1$ if x is odd. The σ-algebra generated by X is $\mathcal{A} = \{\emptyset, \{1, 3, 5\}, \{0, 2, 4\}, \Omega\}$.

Definition. Given a set $B \in \mathcal{A}$ with $\mathrm{P}[B] > 0$, we define

$$\mathrm{P}[A|B] = \frac{\mathrm{P}[A \cap B]}{\mathrm{P}[B]} \, ,$$

the **conditional probability** of A with respect to B. It is the probability of the event A, under the condition that the event B happens.

Example. We throw two fair dice. Let A be the event that the first dice is 6 and let B be the event that the sum of two dices is 11. Because $\mathrm{P}[B] = 2/36 = 1/18$ and $\mathrm{P}[A \cap B] = 1/36$ (we need to throw a 6 and then a 5), we have $\mathrm{P}[A|B] = (1/16)/(1/18) = 1/2$. The interpretation is that since we know that the event B happens, we have only two possibilities: $(5, 6)$ or $(6, 5)$. On this space of possibilities, only the second is compatible with the event B.

Exercice. a) Verify that the **Sicherman dices** with faces $(1,3,4,5,6,8)$ and $(1,2,2,3,3,4)$ have the property that the probability of getting the value k is the same as with a pair of standard dice. For example, the probability to get 5 with the Sicherman dices is $3/36$ because the three cases $(1,4),(3,2),(3,2)$ lead to a sum 5. Also for the standard dice, we have three cases $(1,4),(2,3),(3,2)$.
b) Three dices A,B,C are called **non-transitive**, if the probability that $A > B$ is larger than $1/2$, the probability that $B > C$ is larger than $1/2$ and the probability that $C > A$ is larger than $1/2$. Verify the nontransitivity property for $A = (1,4,4,4,4,4)$, $B = (3,3,3,3,3,6)$ and $C = (2,2,2,5,5,5)$.

Properties. The following properties of conditional probability are called **Keynes postulates**. While they follow immediately from the definition of conditional probability, they are historically interesting because they appeared already in 1921 as part of an axiomatization of probability theory:

1) $\mathrm{P}[A|B] \geq 0$.
2) $\mathrm{P}[A|A] = 1$.
3) $\mathrm{P}[A|B] + \mathrm{P}[A^c|B] = 1$.
4) $\mathrm{P}[A \cap B|C] = \mathrm{P}[A|C] \cdot \mathrm{P}[B|A \cap C]$.

Definition. A finite set $\{A_1, \dots, A_n\} \subset \mathcal{A}$ is called a **finite partition** of Ω if $\bigcup_{j=1}^n A_j = \Omega$ and $A_j \cap A_i = \emptyset$ for $i \neq j$. A finite partition covers the entire space with finitely many, pairwise disjoint sets.

If all possible experiments are partitioned into different events A_j and the probabilities that B occurs under the condition A_j, then one can compute the probability that A_i occurs knowing that B happens:

Theorem 2.1.1 (Bayes rule). Given a finite partition $\{A_1, .., A_n\}$ in $\mathcal{A}$ and $B \in \mathcal{A}$ with $\mathrm{P}[B] > 0$, one has

$$\mathrm{P}[A_i|B] = \frac{\mathrm{P}[B|A_i]\mathrm{P}[A_i]}{\sum_{j=1}^n \mathrm{P}[B|A_j]\mathrm{P}[A_j]} .$$

Proof. Because the denominator is $\mathrm{P}[B] = \sum_{j=1}^n \mathrm{P}[B|A_j]\mathrm{P}[A_j]$, the Bayes rule just says $\mathrm{P}[A_i|B]\mathrm{P}[B] = \mathrm{P}[B|A_i]\mathrm{P}[A_i]$. But these are by definition both $\mathrm{P}[A_i \cap B]$. □

Example. A fair dice is rolled first. It gives a random number k from $\{1,2,3,4,5,6\}$. Next, a fair coin is tossed k times. Assume, we know that all coins show heads, what is the probability that the score of the dice was equal to 5?

Solution. Let B be the event that all coins are heads and let A_j be the event that the dice showed the number j. The problem is to find $\mathrm{P}[A_5|B]$. We know $\mathrm{P}[B|A_j] = 2^{-j}$. Because the events $A_j, j = 1,\dots,6$ form a partition of Ω, we have $\mathrm{P}[B] = \sum_{j=1}^6 \mathrm{P}[B \cap A_j] = \sum_{j=1}^6 \mathrm{P}[B|A_j]\mathrm{P}[A_j] = \sum_{j=1}^6 2^{-j}/6 = (1+1/2+1/3+1/4+1/5+1/6)/6 = 49/120$. By Bayes rule,

$$\mathrm{P}[A_5|B] = \frac{\mathrm{P}[B|A_5]\mathrm{P}[A_5]}{(\sum_{j=1}^6 \mathrm{P}[B|A_j]\mathrm{P}[A_j])} = \frac{(1/32)(1/6)}{49/120} = \frac{5}{392} ,$$

which is about 1 percent.

Example. The **Girl-Boy problem**: "Dave has two child. One child is a boy. What is the probability that the other child is a girl"?

Most people would intuitively say 1/2 because the second event looks independent of the first. However, it is not and the initial intuition is misleading. Here is the solution: first introduce the probability space of all possible events $\Omega = \{BG, GB, BB, GG\}$ with $\mathrm{P}[\{BG\}] = \mathrm{P}[\{GB\}] = \mathrm{P}[\{BB\}] = \mathrm{P}[\{GG\}] = 1/4$. Let $B = \{BG, GB, BB\}$ be the event that there is at least one boy and $A = \{GB, BG, GG\}$ be the event that there is at least one girl. We have

$$\mathrm{P}[A|B] = \frac{\mathrm{P}[A \cap B]}{\mathrm{P}[B]} = \frac{(1/2)}{(3/4)} = \frac{2}{3} .$$

Definition. Two events A, B in s probability space $(\Omega, \mathcal{A}, \mathrm{P})$ are called **independent**, if

$$\mathrm{P}[A \cap B] = \mathrm{P}[A] \cdot \mathrm{P}[B] .$$

Example. The probability space $\Omega = \{1,2,3,4,5,6\}$ and $p_i = \mathrm{P}[\{i\}] = 1/6$ describes a fair dice which is thrown once. The set $A = \{1,3,5\}$ is the event that "the dice produces an odd number". It has the probability 1/2. The event $B = \{1,2\}$ is the event that the dice shows a number smaller than 3. It has probability 1/3. The two events are independent because $\mathrm{P}[A \cap B] = \mathrm{P}[\{1\}] = 1/6 = \mathrm{P}[A] \cdot \mathrm{P}[B]$.

Definition. Write $J \subset_f I$ if J is a **finite subset** of I. A family $\{\mathcal{A}_i\}_{i \in I}$ of σ-sub-algebras of $\mathcal{A}$ is called **independent**, if for every $J \subset_f I$ and every choice $A_j \in \mathcal{A}_j$ $\mathrm{P}[\bigcap_{j \in J} A_j] = \prod_{j \in J} \mathrm{P}[A_j]$. A family $\{X_j\}_{j \in J}$ of random variables is called **independent**, if $\{\sigma(X_j)\}_{j \in J}$ are independent σ-algebras. A family of sets $\{A_j\}_{j \in I}$ is called **independent**, if the σ-algebras $\mathcal{A}_j = \{\emptyset, A_j, A_j^c, \Omega\}$ are independent.

Example. On $\Omega = \{1,2,3,4\}$ the two σ-algebras $\mathcal{A} = \{\emptyset, \{1,3\}, \{2,4\}, \Omega\}$ and $\mathcal{B} = \{\emptyset, \{1,2\}, \{3,4\}, \Omega\}$ are independent.

Properties. (1) If a σ-algebra $\mathcal{F} \subset \mathcal{A}$ is independent to itself, then $\mathrm{P}[A \cap A] = \mathrm{P}[A] = \mathrm{P}[A]^2$ so that for every $A \in \mathcal{F}$, $\mathrm{P}[A] \in \{0,1\}$. Such a σ-algebra is called **P-trivial.**
(2) Two sets $A, B \in \mathcal{A}$ are independent if and only if $\mathrm{P}[A \cap B] = \mathrm{P}[A] \cdot \mathrm{P}[B]$.
(3) If A, B are independent, then A, B^c are independent too.
(4) If $\mathrm{P}[B] > 0$, and A, B are independent, then $\mathrm{P}[A|B] = \mathrm{P}[A]$ because $\mathrm{P}[A|B] = (\mathrm{P}[A] \cdot \mathrm{P}[B])/\mathrm{P}[B] = \mathrm{P}[A]$.
(5) For independent sets A, B, the σ-algebras $\mathcal{A} = \{\emptyset, A, A^c, \Omega\}$ and $\mathcal{B} = \{\emptyset, B, B^c, \Omega\}$ are independent.

Definition. A family $\mathcal{I}$ of subsets of Ω is called a **π-system**, if $\mathcal{I}$ is closed under intersections: if A, B are in $\mathcal{I}$, then $A \cap B$ is in $\mathcal{I}$. A σ-additive map from a π-system $\mathcal{I}$ to $[0, \infty)$ is called a **measure**.

Example. 1) The family $\mathcal{I} = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1,2\}, \{2,3\}, \Omega\}$ is a π-system on $\Omega = \{1,2,3\}$.
2) The set $\mathcal{I} = \{[a,b) \mid 0 \le a < b \le 1\} \cup \{\emptyset\}$ of all half closed intervals is a π-system on $\Omega = [0,1]$ because the intersection of two such intervals $[a,b)$ and $[c,d)$ is either empty or again such an interval $[c,b)$.

Definition. We use the notation $A_n \nearrow A$ if $A_n \subset A_{n+1}$ and $\bigcup_n A_n = A$. Let Ω be a set. $(\Omega, \mathcal{D})$ is called a **Dynkin system** if $\mathcal{D}$ is a set of subsets of Ω satisfying

(i) $\Omega \in \mathcal{A}$,
(ii) $A, B \in \mathcal{D}$, $A \subset B \Rightarrow B \setminus A \in \mathcal{D}$.
(iii) $A_n \in \mathcal{D}, A_n \nearrow A \Rightarrow A \in \mathcal{D}$

Lemma 2.1.2. $(\Omega, \mathcal{A})$ is a σ-algebra if and only if it is a π-system and a Dynkin system.

Proof. If $\mathcal{A}$ is a σ-algebra, then it certainly is both a π-system and a Dynkin system. Assume now, $\mathcal{A}$ is both a π-system and a Dynkin system. Given $A, B \in \mathcal{A}$. The Dynkin property implies that $A^c = \Omega \setminus A, B^c = \Omega \setminus B$ are in $\mathcal{A}$ and by the π-system property also $A \cup B = \Omega \setminus (A^c \cap B^c) \in \mathcal{A}$. Given a sequence $A_n \in \mathcal{A}$. Define $B_n = \bigcup_{k=1}^n A_k \in \mathcal{A}$ and $A = \bigcup_n A_n$. Then

$A_n \nearrow A$ and by the Dynkin property $A \in \mathcal{A}$. Also $\bigcap_n A_n = \Omega \setminus \bigcup_n A_n^c \in \mathcal{A}$ so that $\mathcal{A}$ is a σ-algebra. □

Definition. If $\mathcal{I}$ is any set of subsets of Ω, we denote by $d(\mathcal{I})$ the smallest Dynkin system, which contains $\mathcal{I}$ and call it the **Dynkin system generated by** $\mathcal{I}$.

Lemma 2.1.3. If $\mathcal{I}$ is a π- system, then $d(\mathcal{I}) = \sigma(\mathcal{I})$.

Proof. By the previous lemma, we need only to show that $d(\mathcal{I})$ is a $\pi-$ system.
(i) Define $\mathcal{D}_1 = \{B \in d(\mathcal{I}) \mid B \cap C \in d(\mathcal{I}), \forall C \in \mathcal{I}\ \}$. Because $\mathcal{I}$ is a π-system, we have $\mathcal{I} \subset \mathcal{D}_1$.
Claim. $\mathcal{D}_1$ is a Dynkin system.
Proof. Clearly $\Omega \in \mathcal{D}_1$. Given $A, B \in \mathcal{D} - 1$ with $A \subset B$. For $C \in \mathcal{I}$ we compute $(B \setminus A) \cap C = (B \cap C) \setminus (A \cap C)$ which is in $d(\mathcal{I})$. Therefore $A \setminus B \in \mathcal{D}_1$. Given $A_n \nearrow A$ with $A_n \in \mathcal{D}_1$ and $C \in \mathcal{I}$. Then $A_n \cap C \nearrow A \cap C$ so that $A \cap C \in d(\mathcal{I})$ and $A \in \mathcal{D}_1$.
(ii) Define $\mathcal{D}_2 = \{A \in d(\mathcal{I}) \mid B \cap A \in d(\mathcal{I}), \forall B \in d(\mathcal{I})\ \}$. From (i) we know that $\mathcal{I} \subset \mathcal{D}_2$. Like in (i), we show that $\mathcal{D}_2$ is a Dynkin-system. Therefore $\mathcal{D}_2 = d(\mathcal{I})$, which means that $d(\mathcal{I})$ is a π-system. □

Lemma 2.1.4. (Extension lemma) Given a π-system $\mathcal{I}$. If two measures μ, ν on $\sigma(\mathcal{I})$ satisfy $\mu(\Omega), \nu(\Omega) < \infty$ and $\mu(A) = \nu(A)$ for $A \in \mathcal{I}$, then $\mu = \nu$..

Proof. Proof of lemma (2.1.5). The set $\mathcal{D} = \{A \in \sigma(\mathcal{I}) \mid \mu(A) = \nu(A)\ \}$ is Dynkin system: first of all $\Omega \in \mathcal{D}$. Given $A, B \in \mathcal{D}, A \subset B$. Then $\mu(B \setminus A) = \mu(B) - \mu(A) = \nu(B) - \nu(A) = \nu(B \setminus A)$ so that $B \setminus A \in \mathcal{D}$. Given $A_n \in \mathcal{D}$ with $A_n \nearrow A$, then the σ additivity gives $\mu(A) = \limsup_n \mu(A_n) = \limsup_n \nu(A_n) = \nu(A)$, so that $A \in \mathcal{D}$. Since $\mathcal{D}$ is a Dynkin system containing the π-system $\mathcal{I}$, we know that $\sigma(\mathcal{I}) = d(\mathcal{I}) \subset \mathcal{D}$ which means that $\mu = \nu$ on $\sigma(\mathcal{I})$. □

Definition. Given a probability space $(\Omega, \mathcal{A}, \mathrm{P})$. Two π-systems $\mathcal{I}, \mathcal{J} \subset \mathcal{A}$ are called **P-independent**, if for all $A \in \mathcal{I}$ and $B \in \mathcal{J}$, $\mathrm{P}[A \cap B] = \mathrm{P}[A] \cdot \mathrm{P}[B]$.

Lemma 2.1.5. Given a probability space $(\Omega, \mathcal{A}, \mathrm{P})$. Let $\mathcal{G}, \mathcal{H}$ be two σ-subalgebras of $\mathcal{A}$ and $\mathcal{I}$ and $\mathcal{J}$ be two π-systems satisfying $\sigma(\mathcal{I}) = \mathcal{G}$, $\sigma(\mathcal{J}) = \mathcal{H}$. Then $\mathcal{G}$ and $\mathcal{H}$ are independent if and only if $\mathcal{I}$ and $\mathcal{J}$ are independent.

Proof. (i) Fix $I \in \mathcal{I}$ and define on $(\Omega, \mathcal{H})$ the measures $\mu(H) = \mathrm{P}[I \cap H], \nu(H) = \mathrm{P}[I]\mathrm{P}[H]$ of total probability $\mathrm{P}[I]$. By the independence of $\mathcal{I}$ and $\mathcal{J}$, they coincide on $\mathcal{J}$ and by the extension lemma (2.1.4), they agree on $\mathcal{H}$ and we have $\mathrm{P}[I \cap H] = \mathrm{P}[I]\mathrm{P}[H]$ for all $I \in \mathcal{I}$ and $H \in \mathcal{H}$.
(ii) Define for fixed $H \in \mathcal{H}$ the measures $\mu(G) = \mathrm{P}[G \cap H]$ and $\nu(G) = \mathrm{P}[G]\mathrm{P}[H]$ of total probability $\mathrm{P}[H]$ on $(\Omega, \mathcal{G})$. They agree on $\mathcal{I}$ and so on $\mathcal{G}$. We have shown that $\mathrm{P}[G \cap H] = \mathrm{P}[G]\mathrm{P}[H]$ for all $G \in \mathcal{G}$ and all $H \in \mathcal{H}$. □

Definition. $\mathcal{A}$ is an **algebra** if $\mathcal{A}$ is a set of subsets of Ω satisfying

(i) $\Omega \in \mathcal{A}$,
(ii) $A \in \mathcal{A} \Rightarrow A^c \in \mathcal{A}$,
(iii) $A, B \in \mathcal{A} \Rightarrow A \cup B \in \mathcal{A}$

A σ-additive map from $\mathcal{A}$ to $[0, \infty)$ is called a **measure**.

Theorem 2.1.6 (Carathéodory continuation theorem). Any measure on an algebra $\mathcal{R}$ has a unique continuation to a measure on $\sigma(\mathcal{R})$.

Before we launch into the proof of this theorem, we need two lemmas:

Definition. Let $\mathcal{A}$ be an algebra and $\lambda : \mathcal{A} \mapsto [0, \infty]$ with $\lambda(\emptyset) = 0$. A set $A \in \mathcal{A}$ is called a **λ-set**, if $\lambda(A \cap G) + \lambda(A^c \cap G) = \lambda(G)$ for all $G \in \mathcal{A}$.

Lemma 2.1.7. The set $\mathcal{A}_\lambda$ of λ-sets of an algebra $\mathcal{A}$ is again an algebra and satisfies $\sum_{k=1}^n \lambda(A_k \cap G) = \lambda((\bigcup_{k=1}^n A_k) \cap G)$ for all finite disjoint families $\{A_k\}_{k=1}^n$ and all $G \in \mathcal{A}$.

Proof. From the definition is clear that $\Omega \in \mathcal{A}_\lambda$ and that if $B \in \mathcal{A}_\lambda$, then $B^c \in \mathcal{A}_\lambda$. Given $B, C \in \mathcal{A}_\lambda$. Then $A = B \cap C \in \mathcal{A}_\lambda$. Proof. Since $C \in \mathcal{A}_\lambda$, we get

$$\lambda(C \cap A^c \cap G) + \lambda(C^c \cap A^c \cap G) = \lambda(A^c \cap G) .$$

This can be rewritten with $C \cap A^c = C \cap B^c$ and $C^c \cap A^c = C^c$ as

$$\lambda(A^c \cap G) = \lambda(C \cap B^c \cap G) + \lambda(C^c \cap G) . \tag{2.1}$$

Because B is a λ-set, we get using $B \cap C = A$.

$$\lambda(A \cap G) + \lambda(B^c \cap C \cap G) = \lambda(C \cap G) . \tag{2.2}$$

Since C is a λ-set, we have

$$\lambda(C \cap G) + \lambda(C^c \cap G) = \lambda(G) . \tag{2.3}$$

Adding up these three equations shows that $B \cap C$ is a λ-set. We have so verified that $\mathcal{A}_\lambda$ is an algebra. If B and C are disjoint in $\mathcal{A}_\lambda$ we deduce from the fact that B is a λ-set

$$\lambda(B \cap (B \cup C) \cap G) + \lambda(B^c \cap (B \cup C) \cap G) = \lambda((B \cup C) \cap G) .$$

This can be rewritten as $\lambda(B \cap G) + \lambda(C \cap G) = \lambda((B \cup C) \cap G)$. The analog statement for finitely many sets is obtained by induction. □

Definition. Let $\mathcal{A}$ be a σ-algebra. A map $\lambda : \mathcal{A} \to [0, \infty]$ is called an **outer measure**, if

$$\boxed{\begin{array}{l} \lambda(\emptyset) = 0, \\ A, B \in \mathcal{A} \text{ with } A \subset B \Rightarrow \lambda(A) \leq \lambda(B). \\ A_n \in \mathcal{A} \ \Rightarrow \lambda(\bigcup_n A_n) \leq \sum_n \mathrm{P}[A_n] \ (\sigma \text{ subadditivity}) \end{array}}$$

Lemma 2.1.8. (Carathéodory's lemma) If λ is an outer measure on a measurable space $(\Omega, \mathcal{A})$, then $\mathcal{A}_\lambda \subset \mathcal{A}$ defines a σ-algebra on which λ is countably additive.

Proof. Given a disjoint sequence $A_n \in \mathcal{A}_\lambda$. We have to show that $A = \bigcup_n A_n \in \mathcal{A}_\lambda$ and $\lambda(A) = \sum_n \lambda(A_n)$. By the above lemma (2.1.7), $B_n = \bigcup_{k=1}^n A_k$ is in $\mathcal{A}_\lambda$. By the monotonicity, additivity and the σ -subadditivity, we have

$$\begin{aligned} \lambda(G) &= \lambda(B_n \cap G) + \lambda(B_n^c \cap G) \geq \lambda(B_n \cap G) + \lambda(A^c \cap G) \\ &= \sum_{k=1}^n \lambda(A_k \cap G) + \lambda(A^c \cap G) \geq \lambda(A \cap G) + \lambda(A^c \cap G) . \end{aligned}$$

Subadditivity for λ gives $\lambda(G) \leq \lambda(A \cap G) + \lambda(A^c \cap G)$. All the inequalities in this proof are therefore equalities. We conclude that $A \in \mathcal{L}$ and that λ is σ additive on $\mathcal{A}_\lambda$. □

We now prove the Caratheodory's continuation theorem (2.1.6):

Proof. Given an algebra $\mathcal{R}$ with a measure μ. Define $\mathcal{A} = \sigma(\mathcal{R})$ and the σ-algebra $\mathcal{P}$ consisting of all subsets of Ω. Define on $\mathcal{P}$ the function

$$\lambda(A) = \inf\{\sum_{n \in \mathbb{N}} \mu(A_n) \mid \{A_n\}_{n \in \mathbb{N}} \text{ sequence in } \mathcal{R} \text{ satisfying } A \subset \bigcup_n A_n \} .$$

(i) λ is an outer measure on $\mathcal{P}$.
$\lambda(\emptyset) = 0$ and $\lambda(A) \supset \lambda(B)$ for $B \geq A$ are obvious. To see the σ subadditivity, take a sequence $A_n \in \mathcal{P}$ with $\lambda(A_n) < \infty$ and fix $\epsilon > 0$. For all $n \in \mathbb{N}$, one can (by the definition of λ) find a sequence $\{B_{n,k}\}_{k \in \mathbb{N}}$ in $\mathcal{R}$ such that $A_n \subset \bigcup_{k \in \mathbb{N}} B_{n,k}$ and $\sum_{k \in \mathbb{N}} \mu(B_{n,k}) \leq \lambda(A_n) + \epsilon 2^{-n}$. Define $A = \bigcup_{n \in \mathbb{N}} A_n \subset \bigcup_{n,k \in \mathbb{N}} B_{n,k}$, so that $\lambda(A) \leq \sum_{n,k} \mu(B_{n,k}) \leq \sum_n \lambda(A_n) + \epsilon$. Since ϵ was arbitrary, the σ-subadditivity is proven.

(ii) $\lambda = \mu$ on $\mathcal{R}$.
Given $A \in \mathcal{R}$. Clearly $\lambda(A) \leq \mu(A)$. Suppose that $A \subset \bigcup_n A_n$, with $A_n \in \mathcal{R}$. Define a sequence $\{B_n\}_{n \in \mathbb{N}}$ of disjoint sets in $\mathcal{R}$ inductively $B_1 = A_1$, $B_n = A_n \cap (\bigcup_{k<n} A_k)^c$ such that $B_n \subset A_n$ and $\bigcup_n B_n = \bigcup_n A_n \supset A$. From the σ-additivity of μ on $\mathcal{R}$ follows

$$\mu(A) \leq \bigcup_n \mu(B_n) \leq \bigcup_n \mu(A_n) ,$$

so that $\mu(A) \geq \lambda(A)$.

(iii) Let $\mathcal{P}_\lambda$ be the set of λ-sets in $\mathcal{P}$. Then $\mathcal{R} \subset \mathcal{P}_\lambda$.
Given $A \in \mathcal{R}$ and $G \in \mathcal{P}$. There exists a sequence $\{B_n\}_{n \in \mathbb{N}}$ in $\mathcal{R}$ such that $G \subset \bigcup_n B_n$ and $\sum_n \mu(B_n) \leq \lambda(G) + \epsilon$. By the definition of λ

$$\sum_n \mu(B_n) = \sum_n \mu(A \cap B_n) + \sum_n \mu(A^c \cap B_n) \geq \lambda(A \cap G) + \lambda(A^c \cap G)$$

because $A \cap G \subset \bigcup_n A \cap B_n$ and $A^c \cap G \subset \bigcup_n A^c \cap B_n$. Since ϵ is arbitrary, we get $\lambda(A) \geq \lambda(A \cap G) + \lambda(A^c \cap G)$. On the other hand, since λ is sub-additive, we have also $\lambda(A) \leq \lambda(A \cap G) + \lambda(A^c \cap G)$ and A is a λ-set.

(iv) By (i) λ is an outer measure on $(\Omega, \mathcal{P}_\lambda)$. Since by step (iii), $\mathcal{R} \subset \mathcal{P}_\lambda$, we know by Caratheodory's lemma that $\mathcal{A} \subset \mathcal{P}_\lambda$, so that we can define μ on $\mathcal{A}$ as the restriction of λ to $\mathcal{A}$. By step *(ii)*, this is an extension of the measure μ on $\mathcal{R}$. □

Here is an overview over the possible set of subsets of Ω we have considered. We also include the notion of **ring** and σ-ring, which is often used in measure theory and which differ from the notions of algebra or σ-algebra in that Ω does not have to be in it. In probability theory, those notions are not needed at first. For an introduction into measure theory, see [3, 37, 18].

Set of Ω subsets	contains	closed under
topology	$\emptyset, \Omega$	arbitrary unions, finite intersections
π-system		finite intersections
Dynkin system	Ω	increasing countable union, difference
ring	$\emptyset$	complement and finite unions
σ-ring	$\emptyset$	countably many unions and complement
algebra	Ω	complement and finite unions
σ-algebra	$\emptyset, \Omega$	countably many unions and complement
Borel σ-algebra	$\emptyset, \Omega$	σ-algebra generated by the topology

Remark. The name "ring" has its origin to the fact that with the "addition" $A + B = A\Delta B = (A \cup B) \setminus (A \cap B)$ and "multiplication" $A \cdot B = A \cap B$, a ring of sets becomes an **algebraic ring** like the set of integers, in which rules like $A \cdot (B + C) = A \cdot B + A \cdot C$ hold. The empty set $\emptyset$ is the zero element because $A\Delta\emptyset = A$ for every set A. If the set Ω is also in the ring, one has a ring with 1 because the identity $A \cap \Omega = A$ shows that Ω is the 1-element in the ring.

Lets add some definitions, which will occur later:

Definition. A nonzero measure μ on a measurable space $(\Omega, \mathcal{A})$ is called **positive**, if $\mu(A) \geq 0$ for all $A \in \mathcal{A}$. If μ^+, μ^- are two positive measures so that $\mu(A) = \mu^+ - \mu^-$ then this is called the **Hahn decomposition** of μ. A measure is called **finite** if it has a Hahn decomposition and the positive measure $|\mu|$ defined by $|\mu|(A) = \mu^+(A) + \mu^-(A)$ satisfies $|\mu|(\Omega) < \infty$.

Definition. Let ν be a measure on the measurable space $(\Omega, \mathcal{A})$. We write $\nu << \mu$ if for every A in the σ-algebra $\mathcal{A}$, the condition $\mu(A) = 0$ implies $\nu(A) = 0$. One says that ν is **absolutely continuous** with respect to μ.

2.2 Kolmogorov's 0 − 1 law, Borel-Cantelli lemma

Definition. Given a family $\{\mathcal{A}_i\}_{i\in I}$ of σ-subalgebras of $\mathcal{A}$. For any nonempty set $J \subset I$, let $\mathcal{A}_J := \bigvee_{j\in J} \mathcal{A}_j$ be the σ-algebra generated by $\bigcup_{j\in J} \mathcal{A}_j$. Define also $\mathcal{A}_\emptyset = \{\emptyset, \Omega\}$. The **tail σ-algebra** $\mathcal{T}$ of $\{\mathcal{A}\}_{i\in I}$ is defined as $\mathcal{T} = \bigcap_{J\subset_f I} \mathcal{A}_{J^c}$, where $\mathcal{J}^c = \mathcal{I} \setminus \mathcal{I}$.

Theorem 2.2.1 (Kolmogorov's 0 − 1 law). If $\{\mathcal{A}_i\}_{i\in I}$ are independent σ-algebras, then the tail σ-algebra $\mathcal{T}$ is P-trivial: $\mathrm{P}[A] = 0$ or $\mathrm{P}[A] = 1$ for every $A \in \mathcal{T}$.

Proof. (i) The algebras $\mathcal{A}_F$ and $\mathcal{A}_G$ are independent, whenever $F, G \subset I$ are disjoint.

Proof. Define for $H \subset I$ the π-system

$$\mathcal{I}_H = \{A \in \mathcal{A} \mid A = \bigcap_{i \in K} A_i, K \subset_f H, A_i \in \mathcal{A}_i\} .$$

The π-systems $\mathcal{I}_F$ and $\mathcal{I}_G$ are independent and generate the σ-algebras $\mathcal{A}_F$ and $\mathcal{A}_G$. Use lemma (2.1.5).
(ii) Especially: $\mathcal{A}_J$ is independent of $\mathcal{A}_{J^c}$ for every $J \subset I$.
(iii) $\mathcal{T}$ is independent of $\mathcal{A}_I$.
Proof. $\mathcal{T} = \bigcap_{J \subset_f I} \mathcal{A}_{J^c}$ is independent of any $\mathcal{A}_K$ for $K \subset_f I$. It is therefore independent to the π-system $\mathcal{I}_I$ which generates $\mathcal{A}_I$. Use again lemma (2.1.5).
(iv) $\mathcal{T}$ is a sub-σ-algebra of $\mathcal{A}_I$. Therefore $\mathcal{T}$ is independent of itself which implies that it is P-trivial. □

Example. Let X_n be a sequence of independent random variables and let

$$A = \{\omega \in \Omega \mid \sum_{n=1}^{\infty} X_n \text{ converges } \} .$$

Then $\mathrm{P}[A] = 0$ or $\mathrm{P}[A] = 1$. Proof. Because $\sum_{n=1}^{\infty} X_n$ converges if and only if $\sum_{n=N}^{\infty} X_n$ converges, $A \in \sigma(A_N, A_{N+1} \dots)$ and so $A \in \mathcal{T}$, the tail σ- algebra defined by the independent σ-algebras $\mathcal{A}_n = \sigma(X_n)$. If for example, X_n takes values $\pm 1/n$, each with probability $1/2$, then $\mathrm{P}[A] = 0$. If X_n takes values $\pm 1/n^2$ each with probability $1/2$, then $\mathrm{P}[A] = 1$. As you might guess, the decision whether $\mathrm{P}[A] = 0$ or $\mathrm{P}[A] = 0$ is related to the convergence or divergence of a series. We will come back to that later.

Example. Let $\{A_n\}_{n \in \mathbb{N}}$ be a sequence of of subsets of Ω. The set

$$A_\infty := \limsup_{n \to \infty} A_n = \bigcap_{m=1}^{\infty} \bigcup_{n \geq m} A_n$$

consists of the set $\{\omega \in \Omega\}$ such that $\omega \in A_n$ for infinitely many $n \in \mathbb{N}$. The set A_∞ is contained in the tail σ-algebra of $\mathcal{A}_n = \{\emptyset, A, A^c, \Omega\}$. It follows from Kolmogorov's $0-1$ law that $\mathrm{P}[A_\infty] \in \{0, 1\}$ if $A_n \in \mathcal{A}$ and $\{A_n\}$ are P-independent.

Remark. In the theory of dynamical systems, a measurable map $T : \Omega \to \Omega$ of a probability space $(\Omega, \mathcal{A}, \mathrm{P})$ onto itself is called a K**-system**, if there exists a σ-subalgebra $\mathcal{F} \subset \mathcal{A}$ which satisfies $\mathcal{F} \subset \sigma(T(\mathcal{F}))$ for which the sequence $\mathcal{F}_n = \sigma(T^n(\mathcal{F}))$ satisfies $\mathcal{F}_{\mathbb{N}} = \mathcal{A}$ and which has a trivial tail σ-algebra $\mathcal{T} = \{\emptyset, \Omega\}$. An example of such a system is a shift map $T(x)_n = x_{n+1}$ on $\Omega = \Delta^{\mathbb{N}}$, where Δ is a compact topological space. The K-system property follows from Kolmogorov's $0-1$ law: take $\mathcal{F} = \bigvee_{k=1}^{\infty} T^k(\mathcal{F}_0)$, with $\mathcal{F}_0 = \{x \in \Omega = \Delta^{\mathbb{Z}} \mid x_0 = r \in \Delta \}$.

Theorem 2.2.2 (First Borel-Cantelli lemma). Given a sequence of events $A_n \in \mathcal{A}$. Then

$$\sum_{n \in \mathbb{N}} \mathrm{P}[A_n] < \infty \Rightarrow \mathrm{P}[A_\infty] = 0 .$$

Proof. $\mathrm{P}[A_\infty] = \lim_{n \to \infty} \mathrm{P}[\bigcup_{k \geq n} A_k] \leq \lim_{n \to \infty} \sum_{k \geq n} \mathrm{P}[A_k] = 0.$ □

Theorem 2.2.3 (Second Borel-Cantelli lemma). For a sequence $A_n \in \mathcal{A}$ of independent events,

$$\sum_{n \in \mathbb{N}} \mathrm{P}[A_n] = \infty \Rightarrow \mathrm{P}[A_\infty] = 1 .$$

Proof. For every integer $n \in \mathbb{N}$,

$$\begin{aligned} \mathrm{P}[\bigcap_{k \geq n} A_k^c] &= \prod_{k \geq n} \mathrm{P}[A_k^c] \\ &= \prod_{k \geq n} (1 - \mathrm{P}[A_k]) \leq \prod_{k \geq n} \exp(-\mathrm{P}[A_k]) \\ &= \exp(-\sum_{k \geq n} \mathrm{P}[A_k]) . \end{aligned}$$

The right hand side converges to 0 for $n \to \infty$. From

$$\mathrm{P}[A_\infty^c] = \mathrm{P}[\bigcup_{n \in \mathbb{N}} \bigcap_{k \geq n} A_k^c] \leq \sum_{n \in \mathbb{N}} \mathrm{P}[\bigcap A_k^c] = 0$$

follows $\mathrm{P}[A_\infty^c] = 0$. □

Example. The following example illustrates that independence is necessary in the second Borel-Cantelli lemma: take the probability space $([0,1], \mathcal{B}, \mathrm{P})$, where $\mathrm{P} = dx$ is the Lebesgue measure on the Borel σ-algebra $\mathcal{B}$ of $[0,1]$. For $A_n = [0, 1/n]$ we get $A_\infty = \emptyset$ and so $\mathrm{P}[A_\infty] = 0$. But because $\mathrm{P}[A_n] = 1/n$ we have $\sum_{n=1}^\infty \mathrm{P}[A_n] = \sum_{n=1}^\infty \frac{1}{n} = \infty$ because the **harmonic series** $\sum_{n=1}^\infty 1/n$ diverges:

$$\sum_{n=1}^{R} \frac{1}{n} \geq \int_1^R \frac{1}{x} \, dx = \log(R) .$$

Example. ("Monkey typing Shakespeare") Writing a novel amounts to enter a sequence of N symbols into a computer. For example, to write "Hamlet", Shakespeare had to enter $N = 180'000$ characters. A monkey is placed in front of a terminal and types symbols at random, one per unit time, producing a random sequence X_n of identically distributed sequence of random variables in the set of all possible symbols. If each letter occurs with probability at least ϵ, then the probability that Hamlet appears when typing the first N letters is ϵ^N. Call A_1 this event and call A_k the event that this happens when typing the $(k-1)N+1$ until the kN'th letter. These sets A_k are all independent and have all equal probability. By the second Borel-Cantelli lemma, the events occur infinitely often. This means that Shakespeare's work is not only written once, but infinitely many times. Before we model this precisely, lets look at the odds for random typing. There are 30^N possibilities to write a word of length N with 26 letters together with a minimal set of punctuation: a space, a comma, a dash and a period sign. The chance to write **"To be, or not to be - that is the question."** with 43 random hits onto the keyboard is $1/10^{63.5}$. Note that the life time of a monkey is bounded above by 131400000 $\sim 10^8$ seconds so that it is even unlikely that this single sentence will ever be typed. To compare the probability probability, it is helpful to put the result into a list of known large numbers [10, 38].

10^4	One "myriad". The largest numbers, the Greeks were considering.
10^5	The largest number considered by the Romans.
10^{10}	The age of the universe in years.
10^{17}	The age of the universe in seconds.
10^{22}	Distance to our neighbor galaxy Andromeda in meters.
10^{23}	Number of atoms in two gram Carbon which is 1 Avogadro.
10^{27}	Estimated size of universe in meters.
10^{30}	Mass of the sun in kilograms.
10^{41}	Mass of our home galaxy "milky way" in kilograms.
10^{51}	Archimedes's estimate of number of sand grains in universe.
10^{80}	The number of protons in the universe.

10^{100}	One "googol". (Name coined by 9 year old nephew of E. Kasner).
10^{153}	Number mentioned in a myth about Buddha.
10^{155}	Size of ninth Fermat number (factored in 1990).
10^{10^6}	Size of large prime number (Mersenne number, Nov 1996).
10^{10^7}	Years, ape needs to write "hound of Baskerville" (random typing).
$10^{10^{33}}$	Inverse is chance that a can of beer tips by quantum fluctuation.
$10^{10^{42}}$	Inverse is probability that a mouse survives on the sun for a week.
$10^{10^{50}}$	Estimated number of possible games of chess.
$10^{10^{51}}$	Inverse is chance to find yourself on Mars by quantum fluctuations
$10^{10^{100}}$	One "Gogoolplex"

Lemma 2.2.4. Given a random variable X on a finite probability space Δ, there exists a sequence $X_1, X_2, \dots$ of independent random variables for which all random variables X_i have the same distribution as X.

Proof. The product space $\Omega = \Delta^{\mathbb{N}}$ is compact by Tychonov's theorem. Let $\mathcal{A}$ be the Borel-σ-algebra on Ω and let Qdenote the probability measure on Δ. The probability measure $\mathrm{P} = Q^{\mathbb{Z}}$ is defined on $(\Omega, \mathcal{A})$ has the property that for any **cylinder set**

$$Z(w) = \{\omega \in \Omega \mid \omega_k = r_k, \omega_{k+1} = r_{k+1}, \dots, \omega_n = r_n \}$$

defined by a "word" $w = [r_k, , \dots r_n]$,

$$\mathrm{P}[Z(w)] = \prod_{i=k}^{n} \mathrm{P}[\omega_i = r_i] = \prod_{i=k}^{n} Q(\{r_i\}) .$$

Finite unions of cylinder sets form an algebra $\mathcal{R}$ which generates $\sigma(\mathcal{R}) = \mathcal{A}$. The measure P is σ-additive on this algebra. By Carathéodory's continuation theorem (2.1.6), there exists a measure P on $(\Omega, \mathcal{A})$. For this probability space $(\Omega, \mathcal{A}, \mathrm{P})$, the random variables $X_i(\omega) = \omega_i)$ are independent and have the same distribution as X. □

Example. In the example of the monkey writing a novel, the process of authoring is given by a sequence of independent random variables $X_n(\omega) = \omega_n$. The event that Hamlet is written during the time $[Nk + 1, N(k + 1)]$ is given by a cylinder set A_k. They have all the same probability. By the second Borel-Cantelli lemma, $\mathrm{P}[A_\infty] = 1$. The set A_∞, the event that the Monkey types this novel arbitrarily often, has probability 1.

Remark. Lemma (2.2.4) can be generalized: given any sequence of probability spaces $(\mathbb{R}, \mathcal{B}, \mathrm{P}_i)$ one can form the product space $(\Omega, \mathcal{A}, \mathrm{P})$. The random variables $X_i(\omega) = \omega_i$ are independent and have the law P_i. For an other construction of independent random variables is given in [105].

Exercice. In this exercise, we experiment with some measures on $\Omega = \mathbb{N}$ [108].
a) The distance $d(n, m) = |n - m|$ defines a topology $\mathcal{O}$ on $\Omega = \mathbb{N}$. What is the Borel σ-algebra $\mathcal{A}$ generated by this topology?
b) Show that for every $\lambda > 0$

$$\mathrm{P}[A] = \sum_{n \in A} e^{-\lambda} \frac{\lambda^n}{n!}$$

is a probability measure on the measurable space $(\Omega, \mathcal{A})$ considered in a).
c) Show that for every $s > 1$

$$\mathrm{P}[A] = \sum_{n \in A} \zeta(s)^{-1} n^{-s}$$

is a probability measure on the measurable space $(\Omega, \mathcal{A})$. The function

$$s \mapsto \zeta(s) = \sum_{n \in \Omega} \frac{1}{n^s}$$

is called the **Riemann zeta function.**
d) Show that the sets $A_p = \{n \in \Omega | \, p \text{ divides } n\}$ with prime p are independent. What happens if p is not prime.
e) Give a probabilistic proof of Euler's formula

$$\frac{1}{\zeta(s)} = \prod_{p \text{ prime}} (1 - \frac{1}{p^s}) .$$

f) Let A be the set of natural numbers which are not divisible by a square different from 1. Prove

$$\mathrm{P}[A] = \frac{1}{\zeta(2s)} .$$

2.3 Integration, Expectation, Variance

In this entire section, $(\Omega, \mathcal{A}, \mathrm{P})$ will denote a fixed probability space.

Definition. A **statement** S about points $\omega \in \Omega$ is a map from Ω to $\{\text{true}, \text{false}\}$. A statement is said to hold **almost everywhere**, if the set $\mathrm{P}[\{\omega \mid S(\omega) = \text{false }\}] = 0$. For example, the statement "let $X_n \to X$ almost everywhere", is a short hand notation for the statement that the set $\{x \in \Omega \mid X_n(x) \to X(x) \}$ is measurable and has measure 1.

Definition. The **algebra of all random variables** is denoted by $\mathcal{L}$. It is a vector space over the field $\mathbb{R}$ of the real numbers in which one can multiply. A **elementary function** or **step function** is an element of $\mathcal{L}$ which is of the form

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

with $\alpha_i \in \mathbb{R}$ and where $A_i \in \mathcal{A}$ are disjoint sets. Denote by $\mathcal{S}$ the algebra of step functions. For $X \in \mathcal{S}$ we can define the **integral**

$$\mathrm{E}[X] := \int_{\Omega} X \, dP = \sum_{i=1}^{n} \alpha_i \mathrm{P}[A_i] .$$

Definition. Define $\mathcal{L}^1 \subset \mathcal{L}$ as the set of random variables X, for which

$$\sup_{Y \in \mathcal{S}, Y \leq |X|} \int Y \, dP$$

is finite. For $X \in \mathcal{L}^1$, we can define the **integral** or **expectation**

$$\mathrm{E}[X] := \int X \, dP = \sup_{Y \in \mathcal{S}, Y \leq X^+} \int Y \, dP - \sup_{Y \in \mathcal{S}, Y \leq X^-} \int Y \, dP ,$$

where $X^+ = X \vee 0 = \max(X, 0)$ and $X^- = -X \vee 0 = \max(-X, 0)$. The vector space $\mathcal{L}^1$ is called the space of **integrable random variables**. Similarly, for $p \geq 1$ write $\mathcal{L}^p$ for the set of random variables X for which $\mathrm{E}[|X|^p] < \infty$.

Definition. It is custom to write L^1 for the space $\mathcal{L}^1$, where random variables X, Y for which $\mathrm{E}[|X - Y|] = 0$ are identified. Unlike $\mathcal{L}^p$, the spaces L^p are Banach spaces. We will come back to this later.

Definition. For $X \in \mathcal{L}^2$, we can define the **variance**

$$\mathrm{Var}[X] := \mathrm{E}[(X - \mathrm{E}[X])^2] = \mathrm{E}[X^2] - \mathrm{E}[X]^2 .$$

The nonnegative number

$$\sigma[X] = \mathrm{Var}[X]^{1/2}$$

is called the **standard deviation** of X.

The names **expectation** and **standard deviation** pretty much describe already the meaning of these numbers. The expectation is the "average", "mean" or "expected" value of the variable and the standard deviation measures how much we can expect the variable to deviate from the mean.

Example. The m'th power random variable $X(x) = x^m$ on $([0,1], \mathcal{B}, \mathrm{P})$ has the expectation

$$\mathrm{E}[X] = \int_0^1 x^m \, dx = \frac{1}{m+1} ,$$

the variance

$$\mathrm{Var}[X] = \mathrm{E}[X^2] - \mathrm{E}[X]^2 = \frac{1}{2m+1} - \frac{1}{(m+1)^2} = \frac{m^2}{(1+m)^2(1+2m)}$$

and the standard deviation $\sigma[X] = \frac{m}{(1+m)\sqrt{(1+2m)}}$. Both the expectation as well as the standard deviation converge to 0 if $m \to \infty$.

Definition. If X is a random variable, then $\mathrm{E}[X^m]$ is called the m'th **moment** of X. The m'th **central moment** of X is defined as $\mathrm{E}[(X - \mathrm{E}[X])^m]$.

Definition. The **moment generating function** of X is defined as $M_X(t) = \mathrm{E}[e^{tX}]$. The moment generating function often allows a fast simultaneous computation of all the moments. The function

$$\kappa_X(t) = \log(M_X(t))$$

is called the **cumulant generating function**.

Example. For $X(x) = x$ on $[0,1]$ we have both

$$M_X(t) = \int_0^1 e^{tx}\, dx = \frac{(e^t-1)}{t} = \sum_{m=1}^{\infty} \frac{t^{m-1}}{m!} = \sum_{m=0}^{\infty} \frac{t^m}{(m+1)!}$$

and

$$M_X(t) = \mathrm{E}[e^{tX}] = \mathrm{E}[\sum_{m=0}^{\infty} \frac{t^m X^m}{m!}] = \sum_{m=0}^{\infty} t^m \frac{\mathrm{E}[X^m]}{m!} .$$

Comparing coefficients shows $\mathrm{E}[X^m] = 1/(m+1)$.

Example. Let $\Omega = \mathbb{R}$. For given $m \in \mathbb{R}, \sigma > 0$, define the probability measure $\mathrm{P}[[a,b]] = \int_a^b f(x)\, dx$ with

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-m)^2}{2\sigma^2}} .$$

This is a probability measure because after a change of variables $y = (x-m)/(\sqrt{2}\sigma)$, the integral $\int_{-\infty}^{\infty} f(x)\, dx$ becomes $\frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-y^2}\, dy = 1$. The random variable $X(x) = x$ on $(\Omega, \mathcal{A}, \mathrm{P})$ is a random variable with **Gaussian distribution** mean m and standard deviation σ. One simply calls it a **Gaussian random variable** or random variable with **normal distribution**. Lets justify the constants m and σ: the expectation of X is $\mathrm{E}[X] = \int X\, dP = \int_{-\infty}^{\infty} x f(x)\, dx = m$. The variance is $\mathrm{E}[(X-m)^2] = \int_{-\infty}^{\infty} x^2 f(x)\, dx = \sigma^2$ so that the constant σ is indeed the standard deviation. The moment generating function of X is $M_X(t) = e^{mt+\sigma^2 t^2/2}$. The cumulant generating function is therefore $\kappa_X(t) = mt + \sigma^2 t^2/2$.

Example. If X is a Gaussian random variable with mean $m = 0$ and standard deviation σ, then the random variable $Y = e^X$ has the mean $\mathrm{E}[e^Y] = e^{\sigma^2/2}$. Proof:

$$\frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{y - \frac{y^2}{2\sigma^2}}\, dy = e^{\sigma^2/2} \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{\frac{(y-\sigma^2)^2}{2\sigma^2}}\, dy = e^{\sigma^2/2} .$$

The random variable Y has the **log normal distribution**.

Example. A random variable $X \in \mathcal{L}^2$ with standard deviation $\sigma = 0$ is a constant random variable. It satisfies $X(\omega) = m$ for all $\omega \in \Omega$.

Definition. If $X \in \mathcal{L}^2$ is a random variable with mean m and standard deviation σ, then the random variable $Y = (X-m)/\sigma$ has the mean $m = 0$ and standard deviation $\sigma = 1$. Such a random variable is called **normalized**. One often only adjusts the mean and calls $X - \mathrm{E}[X]$ the **centered random variable**.

Exercice. The **Rademacher functions** $r_n(x)$ are real-valued functions on $[0,1]$ defined by

$$r_n(x) = \begin{cases} 1 & \frac{2k-1}{n} \leq x < \frac{2k}{n} \\ -1 & \frac{2k}{n} \leq x < \frac{2k+1}{n} \end{cases} .$$

They are random variables on the Lebesgue space $([0,1], \mathcal{A}, \mathrm{P} = dx)$.
a) Show that $1-2x = \sum_{n=1}^{\infty} \frac{r_n(x)}{2^n}$. This means that for fixed x, the sequence $r_n(x)$ is the binary expansion of $1-2x$.
b) Verify that $r_n(x) = \mathrm{sign}(\sin(2\pi 2^{n-1} x))$ for almost all x.
c) Show that the random variables $r_n(x)$ on $[0,1]$ are IID random variables with uniform distribution on $\{-1, 1\}$.
d) Each $r_n(x)$ has the mean $\mathrm{E}[r_n] = 0$ and the variance $\mathrm{Var}[r_n] = 1$.

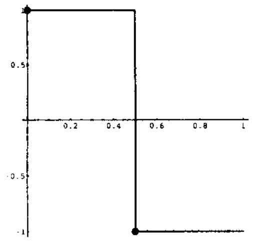


Figure. *The Rademacher Function* $r_1(x)$

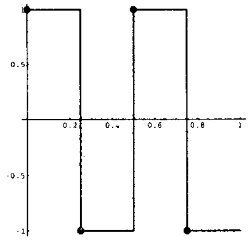


Figure. *The Rademacher Function* $r_2(x)$

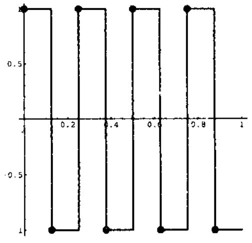


Figure. *The Rademacher Function* $r_3(x)$

2.4 Results from real analysis

In this section we recall some results of **real analysis** with their proofs. In the measure theory or real analysis literature, it is custom to write $\int f(x)\, d\mu(x)$ instead of $\mathrm{E}[X]$ or $f, g, h, \ldots$ instead of $X, Y, Z, \ldots$, but this is just a change of vocabulary. What is special about probability theory is that the measures μ are **probability measures** and so finite.

Theorem 2.4.1 (Monotone convergence theorem, Beppo Lévi 1906). Let X_n be a sequence of random variables in $\mathcal{L}^1$ with $0 \leq X_1 \leq X_2, \dots$ and assume $X = \lim_{n \to \infty} X_n$ converges point wise. If $\sup_n \mathrm{E}[X_n] < \infty$, then $X \in \mathcal{L}^1$ and

$$\mathrm{E}[X] = \lim_{n \to \infty} \mathrm{E}[X_n] .$$

Proof. Because we can replace X_n by $X_n - X_1$, we can assume $X_n \geq 0$. Find for each n a monotone sequence of step functions $X_{n,m} \in \mathcal{S}$ with $X_n = \sup_m X_{n,m}$. Consider the sequence of step functions

$$Y_n := \sup_{1 \leq k \leq n} X_{k,n} \leq \sup_{1 \leq k \leq n} X_{k,n+1} \leq \sup_{1 \leq k \leq n+1} X_{k,n+1} = Y_{n+1} .$$

Since $Y_n \leq \sup_{m=1}^{n} X_m = X_n$ also $\mathrm{E}[Y_n] \leq \mathrm{E}[X_n]$. One checks that $\sup_n Y_n = X$ implies $\sup_n \mathrm{E}[Y_n] = \sup_{Y \in \mathcal{S}, Y \leq X} \mathrm{E}[Y]$ and concludes

$$\mathrm{E}[X] = \sup_{Y \in \mathcal{S}, Y \leq X} \mathrm{E}[Y] = \sup_n \mathrm{E}[Y_n] \leq \sup_n \mathrm{E}[X_n] \leq \mathrm{E}[\sup_n X_n] = \mathrm{E}[X] .$$

We have used the monotonicity $\mathrm{E}[X_n] \leq \mathrm{E}[X_{n+1}]$ in $\sup_n \mathrm{E}[X_n] = \mathrm{E}[X]$. □

Theorem 2.4.2 (Fatou lemma, 1906). Let X_n be a sequence of random variables in $\mathcal{L}^1$ with $|X_n| \leq X$ for some $X \in \mathcal{L}^1$. Then

$$\mathrm{E}[\liminf_{n \to \infty} X_n] \leq \liminf_{n \to \infty} \mathrm{E}[X_n] \leq \limsup_{n \to \infty} \mathrm{E}[X_n] \leq \mathrm{E}[\limsup_{n \to \infty} X_n] .$$

Proof. For $p \geq n$, we have

$$\inf_{m \geq n} X_m \leq X_p \leq \sup_{m \geq n} X_m .$$

Therefore

$$\mathrm{E}[\inf_{m \geq n} X_m] \leq \mathrm{E}[X_p] \leq \mathrm{E}[\sup_{m \geq n} X_m] .$$

Because $p \geq n$ was arbitrary, we have also

$$\mathrm{E}[\inf_{m \geq n} X_m] \leq \inf_{p \geq n} \mathrm{E}[X_p] \leq \sup_{p \geq n} \mathrm{E}[X_p] \leq \mathrm{E}[\sup_{m \geq n} X_m] .$$

Since $Y_n = \inf_{m \geq n} X_m$ is increasing with $\sup_n \mathrm{E}[Y_n] < \infty$ and $Z_n = \sup_{m \geq n} X_m$ is decreasing with $\inf_n \mathrm{E}[Z_n] > -\infty$ we get from Beppo-Levi theorem (2.4.1) that $Y = \sup_n Y_n = \limsup_n X_n$ and $Z = \inf_n Z_n = \liminf_n X_n$ are in $\mathcal{L}^1$ and

$$\begin{aligned} E[\liminf_n X_n] &= \sup_n \mathrm{E}[\inf_{m\geq n} X_m] \leq \sup_n \inf_{m\geq n} \mathrm{E}[X_m] = \liminf_n \mathrm{E}[X_n] \\ &\leq \limsup_n \mathrm{E}[X_n] = \inf_n \sup_{m\geq n} \mathrm{E}[X_m] \\ &\leq \inf_n \mathrm{E}[\sup_{m\geq n} X_m] = \mathrm{E}[\limsup_n X_n] \,. \end{aligned}$$

□

Theorem 2.4.3 (Lebesgue's dominated convergence theorem, 1902). Let X_n be a sequence in $\mathcal{L}^1$ with $|X_n| \leq Y$ for some $Y \in \mathcal{L}^1$. If $X_n \to X$ almost everywhere, then $\mathrm{E}[X_n] \to \mathrm{E}[X]$.

Proof. Since $X = \liminf_n X_n = \limsup_n X_n$ we know that $X \in \mathcal{L}^1$ and from Fatou lemma (2.4.2)

$$\begin{aligned} \mathrm{E}[X] &= \mathrm{E}[\liminf_n X_n] \leq \liminf_n \mathrm{E}[X_n] \\ &\leq \limsup_n \mathrm{E}[X_n] \leq \mathrm{E}[\limsup_n X_n] = \mathrm{E}[X] \,. \end{aligned}$$

□

A special case of Lebesgue's dominated convergence theorem is when $Y = K$ is constant. The theorem is then called the **bounded dominated convergence theorem**. It says that $\mathrm{E}[X_n] \to \mathrm{E}[X]$ if $X_n \leq K$ and $X_n \to X$ almost everywhere.

Definition. Define also for $p \in [1, \infty)$ the vector spaces $\mathcal{L}^p = \{X \in \mathcal{L} \mid |X|^p \in \mathcal{L}^1 \}$ and $\mathcal{L}^\infty = \{X \in \mathcal{L} \mid \exists K \in \mathbb{R}\ X \leq K,$ almost everywhere $\}$.

Example. For $\Omega = [0, 1]$ with the Lebesgue measure $\mathrm{P} = dx$ and Borel σ-algebra $\mathcal{A}$, look at the random variable $X(x) = x^\alpha$, where α is a real number. Because X is bounded for $\alpha > 0$, we have then $X \in \mathcal{L}^\infty$. For $\alpha < 0$, the integral $\mathrm{E}[|X|^p] = \int_0^1 x^{\alpha p}\, dx$ is finite if and only if $\alpha p < 1$ so that X is in $\mathcal{L}^p$ whenever $p > 1/\alpha$.

2.5 Some inequalities

Definition. A function $h : \mathbb{R} \to \mathbb{R}$ is called **convex**, if there exists for all $x_0 \in \mathbb{R}$ a linear map $l(x) = ax + b$ such that $l(x_0) = h(x_0)$ and for all $x \in \mathbb{R}$ the inequality $l(x) \leq h(x)$ holds.

Example. $h(x) = x^2$ is convex, $h(x) = e^x$ is convex, $h(x) = x$ is convex. $h(x) = -x^2$ is not convex, $h(x) = x^3$ is not convex on $\mathbb{R}$ but convex on $\mathbb{R}^+ = [0, \infty)$.

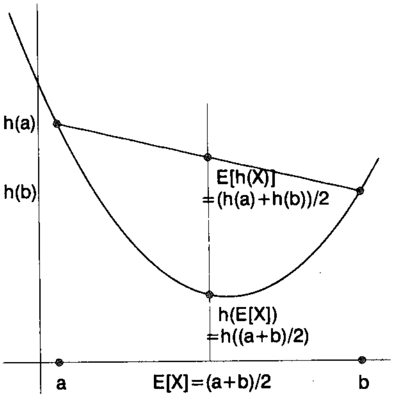


Figure. *The Jensen inequality in the case* $\Omega = \{u, v\}$, $\mathrm{P}[\{u\}] = \mathrm{P}[\{v\}] = 1/2$ *and with* $X(u) = a, X(v) = b$. *The function* h *in this picture is a quadratic function of the form* $h(x) = (x-s)^2 + t$.

Theorem 2.5.1 (Jensen inequality). Given $X \in \mathcal{L}^1$. For any convex function $h : \mathbb{R} \to \mathbb{R}$, we have

$$\mathrm{E}[h(X)] \geq h(\mathrm{E}[X]) ,$$

where the left hand side can also be infinite.

Proof. Let l be the linear map defined at $x_0 = \mathrm{E}[X]$. By the linearity and monotonicity of the expectation, we get

$$h(\mathrm{E}[X]) = l(\mathrm{E}[X]) = \mathrm{E}[l(X)] \leq \mathrm{E}[h(X)] .$$

□

Example. Given $p \leq q$. Define $h(x) = |x|^{q/p}$. Jensen's inequality gives $\mathrm{E}[|X|^q] = \mathrm{E}[h(|X|^p)] \leq h(\mathrm{E}[|X|^p] = \mathrm{E}[|X|^p]^{q/p}$. This implies that $||X||_q := \mathrm{E}[|X|^q]^{1/q} \leq \mathrm{E}[|X|^p]^{1/p} = ||X||_p$ for $p \leq q$ and so

$$\mathcal{L}^\infty \subset \mathcal{L}^p \subset \mathcal{L}^q \subset \mathcal{L}^1$$

for $p \geq q$. The smallest space is $\mathcal{L}^\infty$ which is the space of all bounded random variables.

Exercice. Assume X is a nonnegative random variable for which X and $1/X$ are both in $\mathcal{L}^1$. Show that $\mathrm{E}[X + 1/X] \geq 2$.

We have defined $\mathcal{L}^p$ as the set of random variables which satisfy $\mathrm{E}[|X|^p] < \infty$ for $p \in [1, \infty)$ and $|X| \leq K$ almost everywhere for $p = \infty$. The vector space $\mathcal{L}^p$ has the semi-norm $||X||_p = \mathrm{E}[|X|^p]^{1/p}$ rsp. $||X||_\infty = \inf\{K \in \mathbb{R} \mid |X| \leq K$ almost everywhere $\}$.

Definition. One can construct from $\mathcal{L}^p$ a real **Banach space** $L^p = \mathcal{L}^p/\mathcal{N}$ which is the quotient of $\mathcal{L}^p$ with $\mathcal{N} = \{X \in \mathcal{L}_p \mid ||X||_p = 0\}$. Without this identification, one only has a pre-Banach space in which the property that only the zero element has norm zero is not necessarily true. Especially, for $p = 2$, the space L^2 is a real **Hilbert space** with **inner product** $< X, Y >= \mathrm{E}[XY]$.

Example. The function $f(x) = 1_{\mathbb{Q}}(x)$ which assigns values 1 to rational numbers x on $[0, 1]$ and the value 0 to irrational numbers is different from the constant function $g(x) = 0$ in $\mathcal{L}^p$. But in L^p, we have $f = g$.
The finiteness of the inner product follows from the following inequality:

Theorem 2.5.2 (Hölder inequality, Hölder 1889). Given $p, q \in [1, \infty]$ with $p^{-1} + q^{-1} = 1$ and $X \in \mathcal{L}^p$ and $Y \in \mathcal{L}^q$. Then $XY \in \mathcal{L}^1$ and

$$||XY||_1 \leq ||X||_p ||Y||_q \, .$$

Proof. Without loss of generality, we can restrict the situation to $X, Y \geq 0$ and $||X||_p > 0$. Define the probability measure

$$Q = \frac{X^p P}{\mathrm{E}[X^p]}$$

and define $u = 1_{\{X>0\}} Y/X^{p-1}$. Jensen's inequality gives $Q(u)^q \leq Q(u^q)$ so that

$$\mathrm{E}[|XY|] \leq ||X||_p ||1_{\{Z>0\}} Y||_q \leq ||X||_p ||Y||_q \, .$$

□

A special case of Hölder's inequality is the **Cauchy-Schwarz** inequality

$$||XY||_1 \leq ||X||_2 \cdot ||Y||_2 \, .$$

The semi-norm property of $\mathcal{L}^p$ follows from the following fact:

Theorem 2.5.3 (Minkowski inequality (1896)). Given $p \in [1, \infty]$ and $X, Y \in \mathcal{L}^p$. Then

$$||X + Y||_p \leq ||X||_p + ||Y||_p \, .$$

Proof. We use Hölder's inequality from below to get

$$\mathrm{E}[|X+Y|^p] = \mathrm{E}[|X||X+Y|^{p-1}] + \mathrm{E}[|Y||X+Y|^{p-1}] \leq ||X||_p C + ||Y||_p C ,$$

where $C = |||X+Y|^{p-1}||_q = \mathrm{E}[|X+Y|^p]^{1/q}$ which leads to the claim. □

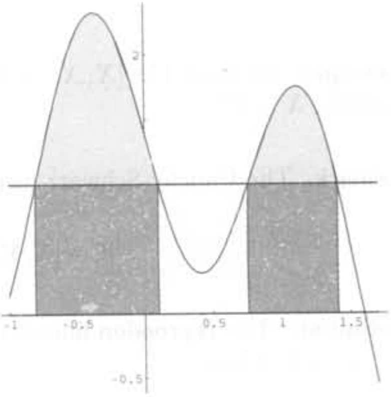
Definition. We use the short-hand notation $\mathrm{P}[X \geq c]$ for $\mathrm{P}[\{\omega \in \Omega \mid X(\omega) \geq c\}]$.

Theorem 2.5.4 (Chebychev-Markov inequality)**.** Let h be a monotone function on $\mathbb{R}$ with $h \geq 0$. For every $c > 0$, and $h(X) \in \mathcal{L}^1$ we have

$$h(c) \cdot \mathrm{P}[X \geq c] \leq \mathrm{E}[h(X)] .$$

Proof. Integrate the inequality $h(c)1_{X \geq c} \leq h(X)$ and use the monotonicity and linearity of the expectation. □

Figure. *The proof of the Chebychev-Markov inequality in the case $h(x) = x$. The left hand side $h(c) \cdot \mathrm{P}[X \geq c]$ is the area of the rectangles $\{X \geq c\} \times [0, h(x)]$ and $\mathrm{E}[h(X)] = \mathrm{E}[X]$ is the area under the graph of X.*



Example. $h(x) = |x|$ leads to $\mathrm{P}[|X| \geq c] \leq ||X||_1/c$ which implies for example the statement

$$\mathrm{E}[|X|] = 0 \Rightarrow \mathrm{P}[X = 0] = 1 .$$

Exercice. Prove the **Chernoff bound**

$$\mathrm{P}[X \geq c] \leq \inf_{t \geq 0} e^{-tc} M_X(t)$$

where $M_X(t) = \mathrm{E}[e^{Xt}]$ is the moment generating function of X.

An important special case of the Chebychev-Markov inequality is the **Chebychev inequality**:

Theorem 2.5.5 (Chebychev inequality). If $X \in \mathcal{L}^2$, then

$$\mathrm{P}[|X - \mathrm{E}[X]| \geq c] \leq \frac{\mathrm{Var}[X]}{c^2} .$$

Proof. Take $h(x) = x^2$ and apply the Chebychev-Markov inequality to the random variable $Y = X - \mathrm{E}[X] \in \mathcal{L}^2$ satisfying $h(Y) \in \mathcal{L}^1$. □

Definition. For $X, Y \in \mathcal{L}^2$ define the **covariance**

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

Two random variables in $\mathcal{L}^2$ are called **uncorrelated** if $\mathrm{Cov}[X,Y] = 0$.

Example. We have $\mathrm{Cov}[X,X] = \mathrm{Var}[X] = \mathrm{E}[(X - \mathrm{E}[X])^2]$ for a random variable $X \in \mathcal{L}^2$.

Remark. The Cauchy-Schwarz-inequality can be restated in the form

$$|\mathrm{Cov}[X,Y]| \leq \sigma[X]\sigma[Y]$$

Definition. The **regression line** of two random variables X, Y is defined as $y = ax + b$, where

$$a = \frac{\mathrm{Cov}[X,Y]}{\mathrm{Var}[X]}, \quad b = \mathrm{E}[Y] - a\mathrm{E}[X] .$$

If $\Omega = \{1, \dots, n\}$ is a finite set, then the random variables X, Y define the vectors

$$X = (X(1), \dots, X(n)), \; Y = (Y(1), \dots, Y(n))$$

or n data points $(X(i), Y(i))$ in the plane. As will follow from the proposition below, the regression line has the property that it minimizes the sum of the squares of the distances from these points to the line.

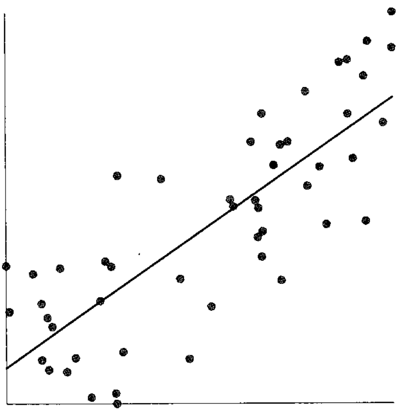


Figure. *Regression line computed from a finite set of data points $(X(i), Y(i))$.*

Example. If X, Y are independent, then $a = 0$. It follows that $b = \mathrm{E}[Y]$.

Example. If $X = Y$, then $a = 1$ and $b = 0$. The best guess for Y is X.

Proposition 2.5.6. If $y = ax + b$ is the regression line of of X, Y, then the random variable $\tilde{Y} = aX + b$ minimizes $\mathrm{Var}[Y - \tilde{Y}]$ under the constraint $\mathrm{E}[Y] = \mathrm{E}[\tilde{Y}]$ and is the best guess for Y, when knowing only $\mathrm{E}[Y]$ and $\mathrm{Cov}[X, Y]$. We check $\mathrm{Cov}[X, Y] = \mathrm{Cov}[X, \tilde{Y}]$.

Proof. To minimize $\mathrm{Var}[aX+b-Y]$ under the constraint $\mathrm{E}[aX+b-Y] = 0$ is equivalent to find (a, b) which minimizes $f(a, b) = \mathrm{E}[(aX + b - Y)^2]$ under the constraint $g(a, b) = \mathrm{E}[aX + b - Y] = 0$. This **least square** solution can be obtained with the Lagrange multiplier method or by solving $b = \mathrm{E}[Y] - a\mathrm{E}[X]$ and minimizing $h(a) = \mathrm{E}[(aX - Y - \mathrm{E}[aX - Y])^2] = a^2(\mathrm{E}[X^2] - \mathrm{E}[X]^2) - 2a(E[XY] - \mathrm{E}[X]\mathrm{E}[Y]) = a^2\mathrm{Var}[X] - 2a\mathrm{Cov}[X, Y]$. Setting $h'(a) = 0$ gives $a = \mathrm{Cov}[X, Y]/\mathrm{Var}[X]$. □

Definition. If the standard deviations $\sigma[X], \sigma[Y]$ are both different from zero, then one can define the **correlation coefficient**

$$\mathrm{Corr}[X, Y] = \frac{\mathrm{Cov}[X, Y]}{\sigma[X]\sigma[Y]}$$

which is a number in $[-1, 1]$. Two random variables in $\mathcal{L}^2$ are called **uncorrelated** if $\mathrm{Corr}[X, Y] = 0$. The other extreme is $|\mathrm{Corr}[X, Y]| = 1$, then $Y = aX + b$ by the Cauchy-Schwarz inequality.

Theorem 2.5.7 (Pythagoras). If two random variables $X, Y \in \mathcal{L}^2$ are independent, then $\mathrm{Cov}[X, Y] = 0$. If X and Y are uncorrelated, then $\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]$.

Proof. We can find monotone sequences of step functions

$$X_n = \sum_{i=1}^{n} \alpha_i 1_{A_i} \to X \; , Y_n = \sum_{j=1}^{n} \beta_j \cdot 1_{B_i} \to Y \; .$$

We can choose these functions in such a way that $A_i \in \mathcal{A} = \sigma(X)$ and $B_j \in \mathcal{B} = \sigma(Y)$. By the Lebesgue dominated convergence theorem (2.4.3), $\mathrm{E}[X_n] \to \mathrm{E}[X]$ and $\mathrm{E}[Y_n] \to \mathrm{E}[Y]$ almost everywhere. Compute $X_n \cdot Y_n = \sum_{i,j=1}^{n} \alpha_i \beta_j 1_{A_i \cap B_j}$. By the Lebesgue dominated convergence theorem (2.4.3) again, $\mathrm{E}[X_n Y_n] \to \mathrm{E}[XY]$. By the independence of X, Y we have $\mathrm{E}[X_n Y_n] = \mathrm{E}[X_n] \cdot \mathrm{E}[Y_n]$ and so $\mathrm{E}[XY] = \mathrm{E}[X]\mathrm{E}[Y]$ which implies $\mathrm{Cov}[X,Y] = \mathrm{E}[XY] - \mathrm{E}[X] \cdot \mathrm{E}[Y] = 0$.
The second statement follows from

$$\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\,\mathrm{Cov}[X,Y] \; .$$

□

Remark. If Ω is a finite set, then the covariance $\mathrm{Cov}[X,Y]$ is the **dot product** between the centered random variables $X - \mathrm{E}[X]$ and $Y - \mathrm{E}[Y]$, and $\sigma[X]$ is the **length** of the vector $X - \mathrm{E}[X]$ and the correlation coefficient $\mathrm{Corr}[X,Y]$ is the cosine of the **angle** α between $X - \mathrm{E}[X]$ and $Y - \mathrm{E}[Y]$ because the dot product satisfies $\vec{v} \cdot \vec{w} = |\vec{v}||\vec{w}|\cos(\alpha)$. So, uncorrelated random variables X, Y have the property that $X - \mathrm{E}[X]$ is **perpendicular** to $Y - \mathrm{E}[Y]$. This geometric interpretation explains, why lemma (2.5.7) is called **Pythagoras theorem.**

For more inequalities in analysis, see the classic [29, 58]. We end this section with a list of properties of variance and covariance:

$\mathrm{Var}[X] \geq 0$.
$\mathrm{Var}[X] = \mathrm{E}[X^2] - \mathrm{E}[X]^2$.
$\mathrm{Var}[\lambda X] = \lambda^2 \mathrm{Var}[X]$.
$\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\mathrm{Cov}[X,Y]$. $\mathrm{Corr}[X,Y] \in [0,1]$.
$\mathrm{Cov}[X,Y] = \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y]$.
$\mathrm{Cov}[X,Y] \leq \sigma[X]\sigma[Y]$.
$\mathrm{Corr}[X,Y] = 1$ if $X - \mathrm{E}[X] = Y - \mathrm{E}[Y]$

2.6 The weak law of large numbers

Consider a sequence $X_1, X_2, \dots$ of random variables on a probability space $(\Omega, \mathcal{A}, \mathrm{P})$. We are interested in the asymptotic behavior of the sums $S_n = X_1 + X_2 + \cdots + X_n$ for $n \to \infty$ and especially in the convergence of the averages S_n/n. The limiting behavior is described by "laws of large numbers". Depending on the definition of convergence, one speaks of "weak" and "strong" laws of large numbers.

We first prove the weak law of large numbers. There exist different versions of this theorem since more assumptions on X_n can allow stronger statements.

Definition. A sequence of random variables Y_n converges **in probability** to a random variable Y, if for all $\epsilon > 0$,

$$\lim_{n \to \infty} \mathrm{P}[|Y_n - Y| \geq \epsilon] = 0 .$$

One calls convergence in probability also **stochastic convergence.**

Remark. If for some $p \in [1, \infty)$, $||X_n - X||_p \to 0$, then $X_n \to X$ in probability since by the Chebychev-Markov inequality (2.5.4), $\mathrm{P}[|X_n - X| \geq \epsilon] \leq ||X - X_n||^p / \epsilon^p$.

Exercice. Show that if two random variables $X, Y \in \mathcal{L}^2$ have non-zero variance and satisfy $|\mathrm{Corr}(X, Y)| = 1$, then $Y = aX + b$ for some real numbers a, b.

Theorem 2.6.1 (Weak law of large numbers for uncorrelated random variables). Assume $X_i \in \mathcal{L}^2$ have common expectation $\mathrm{E}[X_i] = m$ and satisfy $\sup_n \frac{1}{n} \sum_{i=1}^n \mathrm{Var}[X_i] < \infty$. If X_n are pairwise uncorrelated, then $\frac{S_n}{n} \to m$ in probability.

Proof. Since $\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] + 2 \cdot \mathrm{Cov}[X, Y]$ and X_n are pairwise uncorrelated, we get $\mathrm{Var}[X_n + X_m] = \mathrm{Var}[X_n] + \mathrm{Var}[X_m]$ and by induction $\mathrm{Var}[S_n] = \sum_{i=1}^n \mathrm{Var}[X_n]$. Using linearity, we obtain $\mathrm{E}[S_n/n] = m$ and

$$\mathrm{Var}[\frac{S_n}{n}] = \mathrm{E}[\frac{S_n^2}{n^2}] - \frac{\mathrm{E}[S_n]^2}{n^2} = \frac{\mathrm{Var}[S_n]}{n^2} = \frac{1}{n^2} \sum_{i=1}^n \mathrm{Var}[X_n] .$$

The right hand side converges to zero for $n \to \infty$. With Chebychev's inequality (2.5.5), we obtain

$$\mathrm{P}[|\frac{S_n}{n} - m| \geq \epsilon] \leq \frac{\mathrm{Var}[\frac{S_n}{n}]}{\epsilon^2} .$$

□

As an application in analysis, this leads to a constructive proof of a **theorem of Weierstrass** which states that polynomials are dense in the space $C[0,1]$ of all continuous functions on the interval $[0,1]$. Unlike the abstract Weierstrass theorem, the construction with specific polynomials is constructive and gives explicit formulas.

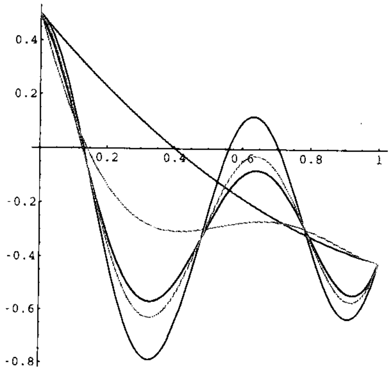


Figure. *Approximation of a function $f(x)$ by Bernstein polynomials $B_2, B_5, B_{10}, B_{20}, B_{30}$.*

Theorem 2.6.2 (Weierstrass theorem). For every $f \in C[0,1]$, the **Bernstein polynomials**

$$B_n(x) = \sum_{k=1}^{n} f(\frac{k}{n}) \left(\begin{array}{c} n \\ k \end{array} \right) x^k (1-x)^{n-k}$$

converge uniformly to f. If $f(x) \geq 0$, then also $B_n(x) \geq 0$.

Proof. For $x \in [0,1]$, let X_n be a sequence of independent $\{0,1\}$- valued random variables with mean value x. In other words, we take the probability space $(\{0,1\}^{\mathbb{N}}, \mathcal{A}, \mathrm{P})$ defined by $\mathrm{P}[\omega_n = 1] = x$. Since $\mathrm{P}[S_n = k] = \left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k}$, we can write $B_n(x) = \mathrm{E}[f(\frac{S_n}{n})]$. We estimate

$$\begin{aligned} |B_n(x) - f(x)| &= |\mathrm{E}[f(\frac{S_n}{n})] - f(x)]| \leq \mathrm{E}[|f(\frac{S_n}{n}) - f(x)|] \\ &\leq 2||f|| \cdot \mathrm{P}[|\frac{S_n}{n} - x| \geq \delta] \\ &+ \sup_{|x-y| \leq \delta} |f(x) - f(y)| \cdot \mathrm{P}[|\frac{S_n}{n} - x| < \delta] \\ &\leq 2||f|| \cdot \mathrm{P}[|\frac{S_n}{n} - x| \geq \delta] \\ &+ \sup_{|x-y| \leq \delta} |f(x) - f(y)| \, . \end{aligned}$$

The second term in the last line is called the **continuity module** of f. It converges to zero for $\delta \to 0$. By the Chebychev inequality (2.5.5) and the proof of the weak law of large numbers, the first term can be estimated from above by

$$2||f||\frac{\mathrm{Var}[X_i]}{n\delta^2} ,$$

a bound which goes to zero for $n \to \infty$ because the variance satisfies $\mathrm{Var}[X_i] = x(1-x) \leq 1/4$. □

In the first version of the weak law of large numbers theorem (2.6.1), we only assumed the random variables to be uncorrelated. Under the stronger condition of independence and a stronger conditions on the moments ($X^4 \in \mathcal{L}^1$), the convergence can be accelerated:

Theorem 2.6.3 (Weak law of large numbers for independent L^4 random variables). Assume $X_i \in \mathcal{L}^4$ have common expectation $\mathrm{E}[X_i] = m$ and satisfy $M = \sup_n ||X||_4 < \infty$. If X_i are independent, then $S_n/n \to m$ in probability. Even $\sum_{n=1}^{\infty} \mathrm{P}[|\frac{S_n}{n} - m| \geq \epsilon]$ converges for all $\epsilon > 0$.

Proof. We can assume without loss of generality that $m = 0$. Because the X_i are independent, we get

$$\mathrm{E}[S_n^4] = \sum_{i_1,i_2,i_3,i_4=1}^{n} \mathrm{E}[X_{i_1}X_{i_2}X_{i_3}X_{i_4}] .$$

Again by independence, a summand $\mathrm{E}[X_{i_1}X_{i_2}X_{i_3}X_{i_4}]$ is zero if an index $i = i_k$ occurs alone, is $\mathrm{E}[X_i^4]$ if all indices are the same and $\mathrm{E}[X_i^2]\mathrm{E}[X_j^2]$, if there are two pairwise equal indices. Since by Jensen's inequality $\mathrm{E}[X_i^2]^2 \leq \mathrm{E}[X_i^4] \leq M$ we get

$$\mathrm{E}[S_n^4] \leq nM + n(n+1)M .$$

Use now the Chebychev-Markov inequality (2.5.4) with $h(x) = x^4$ to get

$$\begin{aligned} \mathrm{P}[|\frac{S_n}{n}| \geq \epsilon] &\leq \frac{\mathrm{E}[(S_n/n)^4]}{\epsilon^4} \\ &\leq M\frac{n+n^2}{\epsilon^4 n^4} \leq 2M\frac{1}{\epsilon^4 n^2} . \end{aligned}$$

□

We can weaken the moment assumption in order to deal with $\mathcal{L}^1$ random variables. Of course, the assumptions have to be made stronger at some other place.

Definition. A family $\{X_i\}_{i \in I}$ of random variables is called **uniformly integrable**, if $\sup_{i \in I} \mathrm{E}[1_{|X_i| \geq R}] \to 0$ for $R \to \infty$. A convenient notation which we will quite often use in the future is $\mathrm{E}[1_A X] = \mathrm{E}[X; A]$ for $X \in \mathcal{L}^1$ and $A \in \mathcal{A}$.

Theorem 2.6.4 (Weak law for uniformly integrable, independent L^1 random variables). Assume $X_i \in \mathcal{L}^1$ are uniformly integrable. If X_i are independent, then $\frac{1}{n} \sum_{i=1}^{n} (X_m - \mathrm{E}[X_m]) \to 0$ in $\mathcal{L}^1$ and therefore in probability.

Proof. Without loss of generality, we can assume that $\mathrm{E}[X_n] = 0$ for all $n \in \mathbb{N}$, because otherwise X_n can be replaced by $Y_n = X_n - \mathrm{E}[X_n]$. Define $f_R(t) = t 1_{[-R,R]}$, the random variables

$$X_n^{(R)} = f_R(X_n) - \mathrm{E}[f_R(X_n)],\ Y_n^{(R)} = X_n - X_n^{(R)}$$

as well as the random variables

$$S_n^{(R)} = \frac{1}{n} \sum_{i=1}^{n} X_n^{(R)},\ T_n^{(R)} = \frac{1}{n} \sum_{i=1}^{n} Y_n^{(R)} .$$

We estimate, using the Minkowski and Cauchy-Schwarz inequalities

$$\begin{aligned} ||S_n||_1 &\leq ||S_n^{(R)}||_1 + ||T_n^{(R)}||_1 \\ &\leq ||S_n^{(R)}||_2 + 2 \sup_{1 \leq l \leq n} \mathrm{E}[|X_l|; |X_l| \geq R] \\ &\leq \frac{R}{\sqrt{n}} + 2 \sup_{l \in \mathbb{N}} \mathrm{E}[|X_l|; |X_l| \geq R] . \end{aligned}$$

In the last step we have used the independence of the random variables and $\mathrm{E}[X_n^{(R)}] = 0$ to get

$$||S_n^{(R)}||_2^2 = \mathrm{E}[(S_n^{(R)})^2] = \frac{\mathrm{E}[(X_n^{(R)})^2]}{n} \leq \frac{R^2}{n} .$$

The claim follows from the uniform integrability assumption
$\sup_{l \in \mathbb{N}} \mathrm{E}[|X_l|; |X_l| \geq R] \to 0$ for $R \to \infty$ □

A special case of the weak law of large numbers is the situation, where all the random variables are IID:

Theorem 2.6.5 (Weak law of large numbers for IID L^1 random variables). Assume $X_i \in \mathcal{L}^1$ are IID random variables with mean m. Then $S_n/n \to m$ in $\mathcal{L}^1$ and so in probability.

Proof. We show that a set of IID $\mathcal{L}^1$ random variables is uniformly integrable: given $X \in \mathcal{L}^1$, we have $K \cdot \mathrm{P}[|X| > K] \leq ||X||_1$ so that $\mathrm{P}[|X| > K] \to 0$ for $K \to \infty$.

Because the random variables X_i are identically distributed, $\mathrm{P}[|X_i|; |X_i| \geq R]$ is independent of i. Consequently any set of IID random variables is also uniformly integrable. We can now use theorem (2.6.4). □

Example. The random variable $X(x) = x^2$ on $[0, 1]$ has the expectation $m = \mathrm{E}[X] = \int_0^1 x^2 \, dx = 1/2$. For every n, we can form the sum $S_n/n = (x_1^2 + x_2^2 + \cdots + x_n^2)/n$. The weak law of large numbers tells us that $\mathrm{P}[|S_n - 1/2| \geq \epsilon] \to 0$ for $n \to \infty$. Geometrically, this means that for every $\epsilon > 0$, the volume of the set of points in the n-dimensional cube for which the distance $r(x_1, .., x_n) = \sqrt{x_1^2 + \cdots + x_n^2}$ to the origin satisfies $\sqrt{n/2} - \epsilon \leq r \leq \sqrt{n/2} + \epsilon$ converges to 1 for $n \to \infty$. In colloquial language, one could rephrase this that asymptotically, as the number of dimensions to go infinity, most of the weight of a n-dimensional cube is concentrated near a shell of radius $1/\sqrt{2} \sim 0.7$ times the length $\sqrt{n}$ of the longest diagonal in the cube.

Exercice. Show that if $X, Y \in \mathcal{L}^1$ are independent random variables, then $XY \in \mathcal{L}^1$. Find an example of two random variables $X, Y \in \mathcal{L}^1$ for which $XY \notin \mathcal{L}^1$.

Exercice. a) Given a sequence $p_n \in [0, 1]$ and a sequence X_n of IID random variables taking values in $\{-1, 1\}$ such that $\mathrm{P}[X_n = 1] = p_n$ and $\mathrm{P}[X_n = -1] = 1 - p_n$. Show that

$$\frac{1}{n} \sum_{k=1}^{n} (X_k - m_k) \to 0$$

in probability, where $m_k = 2p_k - 1$.
b) We assume the same set up like in a) but this time, the sequence p_n is dependent on a parameter. Given a sequence X_n of independent random variables taking values in $\{-1, 1\}$ such that $\mathrm{P}[X_n = 1] = p_n$ and $\mathrm{P}[X_n = -1] = 1 - p_n$ with $p_n = (1 + \cos[\theta + n\alpha])/2$, where θ is a parameter. Prove that $\frac{1}{n} \sum_n X_n \to 0$ in $\mathcal{L}^1$ for almost all θ. You can take for granted the fact that $\frac{1}{n} \sum_{k=1}^{n} p_k \to 1/2$ for almost all real parameters $\theta \in [0, 2\pi]$

Exercice. Prove that $X_n \to X$ in $\mathcal{L}^1$, then there exists of a subsequence $Y_n = X_{n_k}$ satisfying $Y_n \to X$ almost everywhere.

Exercice. Given a sequence of random variables X_n. Show that X_n converges to X in probability if and only if

$$\mathrm{E}[\frac{|X_n - X|}{1 + |X_n - X|}] \to 0$$

for $n \to \infty$.

Exercice. Give an example of a sequence of random variables X_n which converges almost everywhere, but not completely.

Exercice. Use the weak law of large numbers to verify that the volume of an n-dimensional ball of radius 1 satisfies $V_n \to 0$ for $n \to \infty$. Estimate, how fast the volume goes to 0. (See example (2.6))

2.7 The probability distribution function

Definition. The **law** of a random variable X is the probability measure μ on $\mathbb{R}$ defined by $\mu(B) = \mathrm{P}[X^{-1}(B)]$ for all B in the Borel σ-algebra of $\mathbb{R}$. The measure μ is also called the **push-forward measure** under the measurable map $X : \Omega \to \mathbb{R}$.

Definition. The **distribution function** of a random variable X is defined as

$$F_X(s) = \mu((-\infty, s]) = \mathrm{P}[X \leq s] .$$

The distribution function is sometimes also called **cumulative density function** (CDF) but we do not use this name here in order not to confuse it with the **probability density function** (PDF) $f_X(s) = F'_X(s)$ for continuous random variables.

Remark. The distribution function F is very useful. For example, if X is a continuous random variable with distribution function F, then $Y = F(X)$ has the uniform distribution on $[0, 1]$. We can reverse this. If we want to produce random variables with a distribution function F, just take a random variable Y with uniform distribution on $[0, 1]$ and define $X = F^{-1}(Y)$. This random variable has the distribution function F because $\{X \in [a, b]\ \} = \{F^{-1}(Y) \in [a, b]\ \} = \{Y \in F([a, b])\ \} = \{Y \in [F(a), F(b)]\} = F(b) - F(a)$. We see that we need only to have a random number generator which produces uniformly distributed random variables in $[0, 1]$ to produce random variables with a given continuous distribution.

Definition. A set of random variables is called **identically distributed**, if each random variable in the set has the same distribution function. It is called **independent and identically distributed** if the random variables are independent and identically distributed. A common abbreviation for independent identically distributed random variables is **IID**.

Example. Let $\Omega = [0,1]$ be the unit interval with the Lebesgue measure μ and let m be an integer. Define the random variable $X(x) = x^m$. One calls its distribution a **power distribution**. It is in $\mathcal{L}^1$ and has the expectation $\mathrm{E}[X] = 1/(m+1)$. The distribution function of X is $F_X(s) = s^{(1/m)}$ on $[0,1]$ and $F_X(s) = 0$ for $s < 0$ and $F_X(s) = 1$ for $s \geq 1$. The random variable is **continuous** in the sense that it has a **probability density function** $f_X(s) = F'_X(s) = s^{1/m-1}/m$ so that $F_X(s) = \int_{-\infty}^{s} f_X(t)\, dt$.

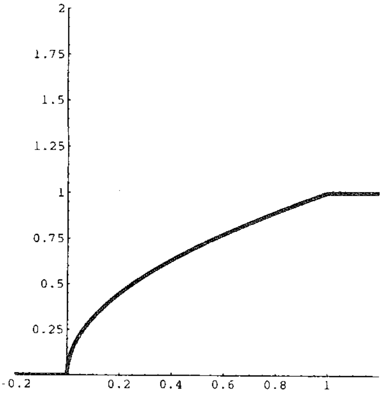


Figure. *The distribution function $F_X(s)$ of $X(x) = x^m$ in the case $m = 2$.*

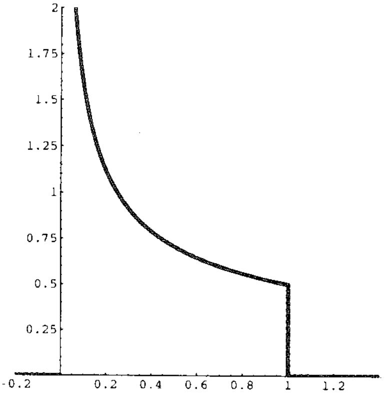


Figure. *The density function $f_X(s)$ of $X(x) = x^m$ in the case $m = 2$.*

Given two IID random variables X, Y with the m'th power distribution as above, we can look at the random variables $V = X+Y, W = X-Y$. One can realize V and W on the unit square $\Omega = [0,1] \times [0,1]$ by $V(x,y) = x^m + y^m$ and $W(x,y) = x^m - y^m$. The distribution functions $F_V(s) = \mathrm{P}[V \leq s]$ and $F_W(s) = \mathrm{P}[V \leq s]$ are the areas of the set $A(s) = \{(x,y) \mid x^m + y^m \leq s\}$ and $B(s) = \{(x,y) \mid x^m - y^m \leq s\}$.

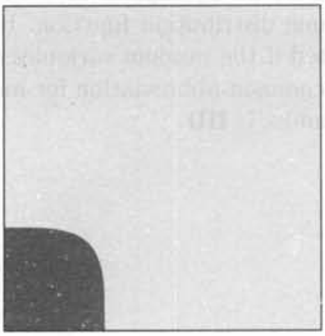


Figure. *$F_V(s)$ is the area of the set $A(s)$, shown here in the case $m = 4$.*

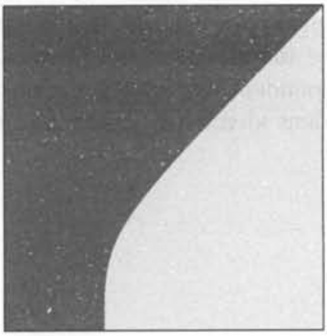


Figure. *$F_W(s)$ is the area of the set $B(s)$, shown here in the case $m = 4$.*

We will later see how to compute the distribution function of a sum of independent random variables algebraically from the probability distribution function F_X. From the area interpretation, we see in this case

$$F_V(s) = \left\{ \begin{array}{ll} \int_0^{s^{1/m}} (s - x^m)^{1/m} \, dx, & s \in [0,1] , \\ 1 - \int_{(s-1)^{1/m}}^{1} 1 - (s - x^m)^{1/m} \, dx, & s \in [1,2] \end{array} \right.$$

and

$$F_W(s) = \left\{ \begin{array}{ll} \int_0^{(s+1)^{1/m}} 1 - (x^m - s)^{1/m} \, dx \, , & s \in [-1,0] , \\ s^{1/m} + \int_{s^{1/m}}^{1} 1 - (x^m - s)^{1/m} \, dx, & s \in [0,1] \end{array} \right.$$

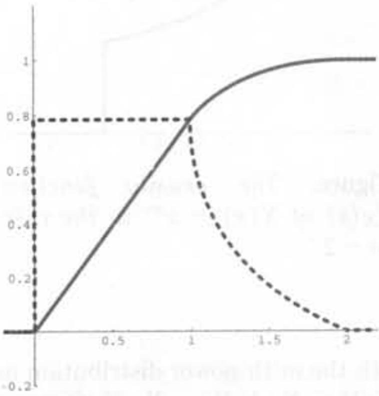


Figure. *The function $F_V(s)$ with density (dashed) $f_V(s)$ of the sum of two power distributed random variables with $m = 2$.*

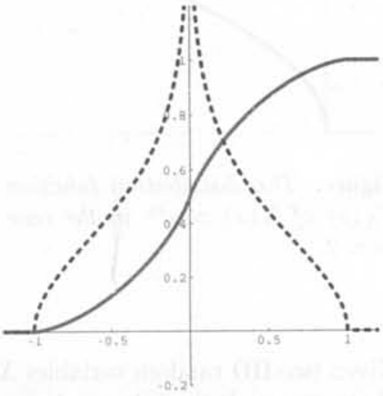


Figure. *The function $F_W(s)$ with density (dashed) $f_W(s)$ of the difference of two power distributed random variables with $m = 2$.*

Exercice. a) Verify that for $\theta > 0$ the **Maxwell distribution**

$$f(x) = \frac{4}{\sqrt{\pi}} \theta^{3/2} x^2 e^{-\theta x^2}$$

is a probability distribution on $\mathbb{R}^+ = [0, \infty)$. This distribution can model the speed distribution of molecules in thermal equilibrium.
a) Verify that for $\theta > 0$ the **Rayleigh distribution**

$$f(x) = 2\theta x e^{-\theta x^2}$$

is a probability distribution on $\mathbb{R}^+ = [0, \infty)$. This distribution can model the speed distribution $\sqrt{X^2 + Y^2}$ of a two dimensional wind velocity (X, Y), where both X, Y are normal random variables.

2.8 Convergence of random variables

In order to formulate the strong law of large numbers, we need some other notions of convergence.

Definition. A sequence of random variables X_n converges **in probability** to a random variable X, if

$$\mathrm{P}[|X_n - X| \geq \epsilon] \to 0$$

for all $\epsilon > 0$.

Definition. A sequence of random variables X_n converges **almost everywhere** or **almost surely** to a random variable X, if $\mathrm{P}[X_n \to X] = 1$.

Definition. A sequence of $\mathcal{L}^p$ random variables X_n **converges in $\mathcal{L}^p$** to a random variable X, if

$$||X_n - X||_p \to 0$$

for $n \to \infty$..

Definition. A sequence of random variables X_n converges **fast in probability**, or **completely** if

$$\sum_n \mathrm{P}[|X_n - X| \geq \epsilon] < \infty$$

for all $\epsilon > 0$.

We have so four notions of convergence of random variables $X_n \to X$, if the random variables are defined on the same probability space $(\Omega, \mathcal{A}, \mathrm{P})$. Later we will the two equivalent but weaker notions **convergence in distribution** and **weak convergence**, which not necessarily assume X_n and X to be defined on the same probability space. Lets add these two definitions also here. We will see later, in theorem (2.13.2) that the following definitions are equivalent:

Definition. A sequence of random variables X_n converges **in distribution**, if $F_{X_n}(x) \to F_X(x)$ for all points s, where F_X is continuous.

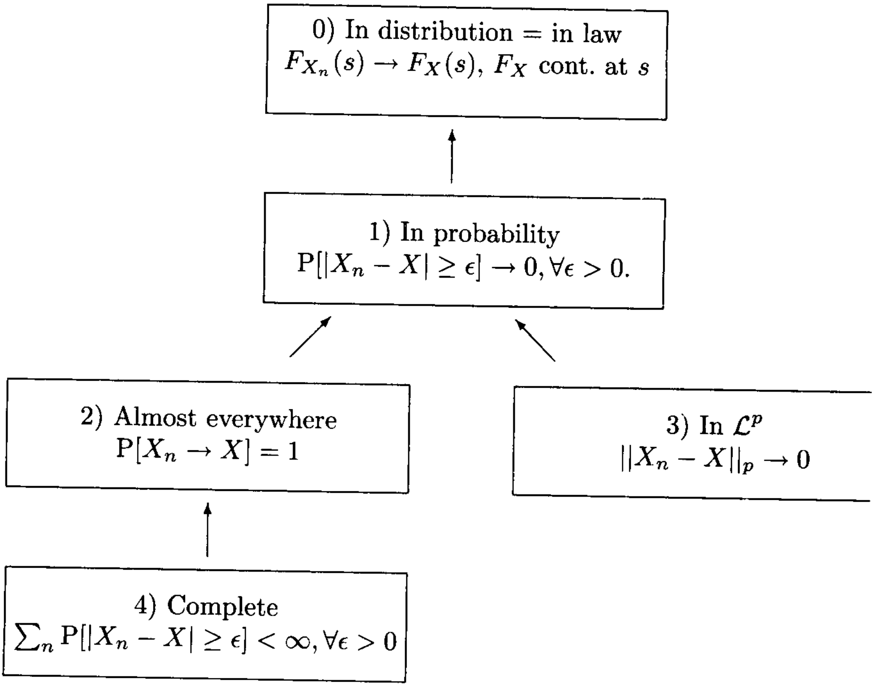
Example. Let $\Omega_n = \{1, 2, ..., n\}$ with the uniform distribution $\mathrm{P}[\{k\}] = 1/n$ and X_n the random variable $X_n(x) = x/n$. Let $X(x) = x$ on the probability space $[0, 1]$ with probability $\mathrm{P}[[a, b)] = b - a$. The random variables X_n and X are defined on a different probability spaces but X_n converges to X in distribution for $n \to \infty$.

Definition. A sequence of random variables X_n converges **in law** to a random variable X, if the laws μ_n of X_n converge weakly to the law μ of X.

Remark. In other words, X_n converges weakly to X if for every continuous function f on $\mathbb{R}$ of compact support, one has

$$\int f(x)\, d\mu_n(x) \to \int f(x)\, d\mu(x) .$$

Proposition 2.8.1. The next figure shows the relations between the different convergence types.


0) In distribution = in law
$F_{X_n}(s) \to F_X(s)$, F_X cont. at s
1) In probability
$\mathrm{P}[|X_n - X| \geq \epsilon] \to 0, \forall \epsilon > 0.$
2) Almost everywhere
$\mathrm{P}[X_n \to X] = 1$
3) In $\mathcal{L}^p$
$||X_n - X||_p \to 0$
4) Complete
$\sum_n \mathrm{P}[|X_n - X| \geq \epsilon] < \infty, \forall \epsilon > 0$

Proof. 2) ⇒ 1): Since

$$\{X_n \to X\} = \bigcap_k \bigcup_m \bigcap_{n \geq m} \{|X_n - X| \leq 1/k\}$$

"almost everywhere convergence" is equivalent to

$$1 = \mathrm{P}[\bigcup_m \bigcap_{n \geq m} \{|X_n - X| \leq \frac{1}{k}\}] = \lim_{n \to \infty} \mathrm{P}[\bigcap_{n \geq m} \{|X_n - X| \leq \frac{1}{k}\}]$$

for all k. Therefore

$$\mathrm{P}[|X_m - X| \geq \epsilon] \leq P[\bigcap_{n \geq m} \{|X_n - X| \geq \epsilon\}] \to 0$$

for all $\epsilon > 0$.
4) ⇒ 2): The first Borel-Cantelli lemma implies that for all $\epsilon > 0$

$$\mathrm{P}[|X_n - X| \geq \epsilon, \text{ infinitely often}] = 0\,.$$

We get so for $\epsilon_n \to 0$

$$\mathrm{P}[\bigcup_n |X_n - X| \geq \epsilon_k, \text{ infinitely often}] \leq \sum_n \mathrm{P}[|X_n - X| \geq \epsilon_k, \text{ infinitely often}] =$$

from which we obtain $\mathrm{P}[X_n \to X] = 1$.
3) ⇒ 1): Use the Chebychev-Markov inequality (2.5.4), to get

$$\mathrm{P}[|X_n - X| \geq \epsilon] \leq \frac{\mathrm{E}[|X_n - X|^p]}{\epsilon^p} .$$

□

Example. Here is an example of convergence in probability but not almost everywhere convergence. Let $([0,1], \mathcal{A}, \mathrm{P})$ be the Lebesgue measure space, where $\mathcal{A}$ is the Borel σ-algebra on $[0,1]$. Define the random variables

$$X_{n,k} = 1_{[k2^{-n},(k+1)2^{-n}]}, \; n = 1, 2, \dots, \; k = 0, \dots, 2^n - 1 .$$

By lexicographical ordering $X_1 = X_{1,1}, X_2 = X_{2,1}, X_3 = X_{2,2}, X_4 = X_{2,3}, \dots$ we get a sequence X_n satisfying

$$\liminf_{n \to \infty} X_n(\omega) = 0, \limsup_{n \to \infty} X_n(\omega) = 1$$

but $\mathrm{P}[|X_{n,k} \geq \epsilon] \leq 2^{-n}$.

Example. And here is an example of almost everywhere but not $\mathcal{L}^p$ convergence: the random variables

$$X_n = 2^n 1_{[0,2^{-n}]}$$

on the probability space $([0,1], \mathcal{A}, \mathrm{P})$ converge almost everywhere to the constant random variable $X = 0$ but not in $\mathcal{L}^p$ because $||X_n||_p = 2^{n(p-1)/p}$.

With more assumptions other implications can hold. We give two examples.

Proposition 2.8.2. Given a sequence $X_n \in \mathcal{L}^\infty$ with $||X_n||_\infty \leq K$ for all n, then $X_n \to X$ in probability if and only if $X_n \to X$ in $\mathcal{L}^1$.

Proof. (i) $\mathrm{P}[|X| \leq K) = 1$. Proof. For $k \in \mathbb{N}$,

$$\mathrm{P}[|X| > K + \frac{1}{k}] \leq \mathrm{P}[|X - X_n| > \frac{1}{k}] \to 0, n \to \infty$$

so that $\mathrm{P}[|X| > K + \frac{1}{k}] = 0$. Therefore

$$\mathrm{P}[|X| > K] = \mathrm{P}[\bigcup_k \{|X| > K + \frac{1}{k}\}] = 0 .$$

(ii) Given $\epsilon > 0$. Choose m such that for all $n > m$

$$\mathrm{P}[|X_n - X| > \frac{\epsilon}{3}] < \frac{\epsilon}{3K} \, .$$

Then, using (i)

$$\begin{aligned} \mathrm{E}[|X_n - X|) &= \mathrm{E}[(|X_n - X|; |X_n - X| > \epsilon] + \mathrm{E}[(|X_n - X|; |X_n - X| \leq \epsilon] \\ &\leq 2K\mathrm{P}[|X_n - X| > \frac{\epsilon}{3}] + \frac{\epsilon}{3} \leq \epsilon \, . \end{aligned}$$

□

Definition. A family $\mathcal{C} \subset \mathcal{L}^1$ of random variables is called **uniformly integrable**, if

$$\lim_{R \to \infty} \sup_{X \in \mathcal{C}} \mathrm{E}[1_{|X|>R}] \to 0$$

for all $X \in \mathcal{C}$. The next lemma was already been used in the proof of the weak law of large numbers for IID random variables.

Lemma 2.8.3. Given $X \in \mathcal{L}^1$ and $\epsilon > 0$. Then, there exists $K \geq 0$ with $\mathrm{E}[|X|; |X| > K] < \epsilon$.

Proof. Given $\epsilon > 0$. If $X \in \mathcal{L}^1$, we can find $\delta > 0$ such that if $\mathrm{P}[A] < \delta$, then $\mathrm{E}[|X|, A] < \epsilon$. Since $K\mathrm{P}[|X| > K] \leq \mathrm{E}[|X|]$, we can choose K such that $\mathrm{P}[|X| > K) < \delta$. Therefore $\mathrm{E}[|X|; |X| > K] < \epsilon$. □

The next proposition gives a necessary and sufficient condition for $\mathcal{L}^1$ convergence.

Proposition 2.8.4. Given a sequence random variables $X_n \in \mathcal{L}^1$. The following is equivalent:
a) X_n converges in probability to X and $\{X_n\}_{n \in \mathbb{N}}$ is uniformly integrable.
b) X_n converges in $\mathcal{L}^1$ to X.

Proof. *a)* $\Rightarrow$ *b)*. Define for $K \geq 0$ and a random variable X the bounded variable

$$X^{(K)} = X \cdot 1_{\{-K \leq X \leq K\}} + K \cdot 1_{\{X > K\}} - K \cdot 1_{\{X < -K\}} \, .$$

By the uniform integrability condition and the above lemma (2.8.3), we can choose K such that for all n,

$$\mathrm{E}[|X_n^{(K)} - X_n|] < \frac{\epsilon}{3}, \; \mathrm{E}[|X^{(K)} - X|] < \frac{\epsilon}{3} \, .$$

Since $|X_n^{(K)} - X^{(K)}| \leq |X_n - X|$, we have $X_n^{(K)} \to X^{(K)}$ in probability. We have so by the last proposition (2.8.2) $X_n^{(K)} \to X^{(K)}$ in $\mathcal{L}^1$ so that for $n > m$ $\mathrm{E}[|X_n^{(K)} - X^{(K)}|] \leq \epsilon/3$. Therefore, for $n > m$ also

$$\mathrm{E}[|X_n - X|] \leq \mathrm{E}[|X_n - X_n^{(K)}|] + \mathrm{E}[|X_n^{(K)} - X^{(K)}|] + \mathrm{E}[|X^{(K)} - X|] \leq \epsilon \, .$$

a) $\Rightarrow$ *b)*. We have seen already that $X_n \to X$ in probability if $||X_n - X||_1 \to 0$. We have to show that $X_n \to X$ in $\mathcal{L}^1$ implies that X_n is uniformly integrable.
Given $\epsilon > 0$. There exists m such that $\mathrm{E}[|X_n - X|] < \epsilon/2$ for $n > m$. By the absolutely continuity property, we can choose $\delta > 0$ such that $\mathrm{P}[A] < \epsilon$ implies

$$\mathrm{E}[|X_n|; A] < \epsilon, 1 \leq n \leq m, \mathrm{E}[|X|; A] < \epsilon/2 \, .$$

Because X_n is bounded in $\mathcal{L}^1$, we can choose K such that $K^{-1} \sup_n \mathrm{E}[|X_n|] < \delta$ which implies $\mathrm{E}[|X_n| > K] < \delta$. For $n \geq m$, we have therefore

$$\mathrm{E}[|X_n|; |X_n| > K] \leq \mathrm{E}[|X|; |X_n| > K] + \mathrm{E}[|X - X_n|] < \epsilon \, .$$

□

Exercice. a) $\mathrm{P}[\sup_{k \geq n} |X_k - X| > \epsilon] \to 0$ for $n \to \infty$ and all $\epsilon > 0$ if and only if $X_n \to X$ almost everywhere.
b) A sequence X_n converges almost surely if and only if

$$\lim_{n \to \infty} \mathrm{P}[\sup_{k \geq 1} |X_{n+k} - X_n| > \epsilon] = 0$$

for all $\epsilon > 0$.

2.9 The strong law of large numbers

The weak law of large numbers makes a statement about the stochastic convergence of sums

$$\frac{S_n}{n} = \frac{X_1 + \cdots + X_n}{n}$$

of random variables X_n. The strong laws of large numbers make analog statements about almost everywhere convergence.

The first version of the strong law does not assume the random variables to have the same distribution. They are assumed to have the same expectation and have to be bounded in $\mathcal{L}^4$.

Theorem 2.9.1 (Strong law for independent L^1-random variables). Assume X_n are independent random variables in $\mathcal{L}^4$ with common expectation $\mathrm{E}[X_n] = m$ and for which $M = \sup_n ||X_n||_4^4 < \infty$. Then $S_n/n \to m$ almost everywhere.

Proof. In the proof of theorem (2.6.3), we derived

$$\mathrm{P}[|\frac{S_n}{n} - m| \geq \epsilon] \leq 2M\frac{1}{\epsilon^4 n^2} .$$

This means that $S_n/n \to m$ converges completely. By proposition (2.8) we have almost everywhere convergence. □

Here is an application of the strong law:

Definition. A real number $x \in [0,1]$ is called **normal** to the base 10, if its decimal expansion $x = x_1 x_2 \dots$ has the property that each digit appears with the same frequency $1/10$.

Corollary 2.9.2. (Normality of numbers) On the probability space $([0,1], \mathcal{B}, \mathrm{Q} = dx)$, Lebesgue almost all numbers x are normal.

Proof. Define the random variables $X_n(x) = x_n$, where x_n is the n'th decimal digit. We have only to verify that X_n are IID random variables. The strong law of large numbers will assure that almost all x are normal. Let $\Omega = \{0, 1, \dots, 9\ \}^{\mathbb{N}}$ be the space of all infinite sequences $\omega = (\omega_1, \omega_2, \omega_3, \dots)$. Define on Ω the product σ-algebra $\mathcal{A}$ and the product probability measure P. Define the measurable map $S(\omega) = \sum_{n=1}^{\infty} \omega_k/10^k = x$ from Ω to $[0,1]$. It produces for every sequence in Ω a real number $x \in [0,1]$. The integers ω_k are just the **decimal digits** of x. The map S is measure preserving and can be inverted on a set of measure 1 because almost all real numbers have a unique decimal expansion.
Because $X_n(x) = X_n(S(\omega)) = Y_n(\omega) = \omega_n$, if $S(\omega) = x$. We see that X_n are the same random variables than Y_n. The later are by construction IID with uniform distribution on $\{0, 1, \dots, 9\ \}$. □

Remark. While almost all numbers are normal, it is difficult to decide normality for specific real numbers. One does not know for example whether $\pi - 3 = 0.1415926\dots$ of $\sqrt{2} - 1 = 0.41421\dots$ is normal.
The strong law for IID random variables was first proven by Kolmogorov in 1930. Only much later in 1981, it has been observed that the weaker notion of **pairwise independence** is sufficient [25].

Theorem 2.9.3 (Strong law for pairwise independent L^1 random variables). Assume $X_n \in \mathcal{L}^1$ are pairwise independent and identically distributed random variables. Then $S_n/n \to \mathrm{E}[X_1]$ almost everywhere.

Proof. We can assume without loss of generality that $X_n \geq 0$ (because we can split $X_n = X_n^+ + X_n^-$ into its positive $X_n^+ = X_n \vee 0 = \max(X_n, 0)$ and negative part $X^- = -X \vee 0 = \max(-X, 0)$. Knowing the result for $X_n^\pm$ implies the result for X_n.).

Define $f_R(t) = t \cdot 1_{[-R,R]}$, the random variables $X_n^{(R)} = f_R(X_n)$ and $Y_n = X_n^{(n)}$ as well as

$$S_n = \frac{1}{n}\sum_{i=1}^n X_i,\ T_n = \frac{1}{n}\sum_{i=1}^n Y_i\ .$$

(i) It is enough to show that $T_n - \mathrm{E}[T_n] \to 0$.
Proof. Since $\mathrm{E}[Y_n] \to \mathrm{E}[X_1] = m$, we get $\mathrm{E}[T_n] \to m$. Because

$$\begin{aligned}
\sum_{n\geq 1} \mathrm{P}[Y_n \neq X_n] &\leq \sum_{n\geq 1} \mathrm{P}[X_n \geq n] = \sum_{n\geq 1} \mathrm{P}[X_1 \geq n] \\
&= \sum_{n\geq 1}\sum_{k\geq n} \mathrm{P}[X_n \in [k, k+1]] \\
&= \sum_{k\geq 1} k \cdot \mathrm{P}[X_1 \in [k, k+1]] \leq \mathrm{E}[X_1] < \infty\ ,
\end{aligned}$$

we get by the first Borel-Cantelli lemma that $\mathrm{P}[Y_n \neq X_n,\ \text{infinitely often}] = 0$. This means $T_n - S_n \to 0$ almost everywhere, proving $\mathrm{E}[S_n] \to m$.
(ii) Fix a real number $\alpha > 1$ and define an exponentially growing subsequence $k_n = [\alpha^n]$ which is the **integer part** of α^n. Denote by μ the law of the random variables X_n. For every $\epsilon > 0$, we get using Chebychev inequality (2.5.5), pairwise independence for $k_n = [\alpha^n]$ and constants C which can vary from line to line:

$$\begin{aligned}
\sum_{n=1}^{\infty} \mathrm{P}[|T_{k_n} - \mathrm{E}[T_{k_n}]| \geq \epsilon] &\leq \sum_{n=1}^{\infty} \frac{\mathrm{Var}[T_{k_n}]}{\epsilon^2 k_n^2} \\
&= \frac{1}{\epsilon^2 k_n^2} \sum_{m=1}^{k_n} \mathrm{Var}[Y_m] \\
&= \frac{1}{\epsilon^2} \sum_{m=1}^{\infty} \mathrm{Var}[Y_m] \sum_{n:k_n\geq m} \frac{1}{k_n^2} \\
&\leq^{(1)} \frac{1}{\epsilon^2} \sum_{m=1}^{\infty} \mathrm{Var}[Y_m]\, \frac{C}{m^2} \\
&\leq C \sum_{m=1}^{\infty} \frac{1}{m^2} \mathrm{E}[Y_m^2]
\end{aligned}$$

In (1) we used that with $k_n = [\alpha^n]$ one has $\sum_{n:k_n\geq m} k_n^{-1} \leq C \cdot m^{-2}$.

Lets take some breath and continue, where we have just left off:

$$\begin{aligned}
\sum_{n=1}^{\infty} \mathrm{P}[|T_{k_n} - \mathrm{E}[T_{k_n}]| \geq \epsilon] &\leq C \sum_{m=1}^{\infty} \frac{1}{m^2} \mathrm{E}[Y_m^2] \\
&\leq C \sum_{m=1}^{\infty} \frac{1}{m^2} \sum_{l=0}^{m-1} \int_l^{l+1} x^2 \, d\mu(x) \\
&= C \sum_{l=0}^{\infty} \sum_{m=l+1}^{\infty} \frac{1}{m^2} \int_l^{l+1} x^2 \, d\mu(x) \\
&\leq C \sum_{l=0}^{\infty} \sum_{m=l+1}^{\infty} \frac{(l+1)}{m^2} \int_l^{l+1} x \, d\mu(x) \\
&\leq^{(2)} C \sum_{l=0}^{\infty} \int_l^{l+1} x \, d\mu(x) \\
&\leq C \cdot \mathrm{E}[X_1] < \infty \, .
\end{aligned}$$

In (2) we used that $\sum_{m=l+1}^{n} m^{-2} \leq C \cdot (l+1)^{-1}$.

We have now proved complete (=fast stochastic) convergence. This implies the almost everywhere convergence of $T_{k_n} - \mathrm{E}[T_{k_n}] \to 0$.

(iii) So far, the convergence has only be verified along a subsequence k_n. Because we assumed $X_n \geq 0$, the sequence $U_n = \sum_{i=1}^{n} Y_n = nT_n$ is monotonically increasing. For $k \in [k_m, k_{m+1}]$, we get therefore

$$\frac{k_m}{k_{m+1}} \frac{U_{k_m}}{k_m} = \frac{U_{k_m}}{k_{m+1}} \leq \frac{U_n}{n} \leq \frac{U_{k_{m+1}}}{k_m} = \frac{k_{m+1}}{k_m} \frac{U_{k_{m+1}}}{k_{m+1}}$$

and from $\lim_{n\to\infty} T_n = \mathrm{E}[X_1]$ almost everywhere,

$$\frac{1}{\alpha} \mathrm{E}[X_1] \leq \liminf_n T_n \leq \limsup_n T_n \leq \alpha \mathrm{E}[X_1]$$

follows. □

Remark. The strong law of large numbers can be interpreted as a statement about the growth of the sequence $\sum_{k=1}^{n} X_n$. For $\mathrm{E}[X_1] = 0$, the convergence $\frac{1}{n} \sum_{k=1}^{n} X_n \to 0$ means that for all $\epsilon > 0$ there exists m such that for $n > m$

$$|\sum_{k=1}^{n} X_n| \leq \epsilon n \, .$$

This means that the trajectory $\sum_{k=1}^{n} X_n$ is finally contained in any arbitrary small cone. In other words, it grows slower than linear. The exact description for the growth of $\sum_{k=1}^{n} X_n$ is given by the **law of the iterated logarithm of Khinchin** which says that a sequence of IID random variables X_n with $\mathrm{E}[X_n] = m$ and $\sigma(X_n) = \sigma \neq 0$ satisfies

$$\limsup_{n\to\infty} \frac{S_n}{\Lambda_n} = +1, \liminf_{n\to\infty} \frac{S_n}{\Lambda_n} = -1 \, ,$$

with $\Lambda_n = \sqrt{2\sigma^2 n \log\log n}$.

Remark. The IID assumption on the random variables can not be weakened without further restrictions. Take for example a sequence X_n of random variables satisfying $\mathrm{P}[X_n = \pm 2^n] = 1/2$. Then $\mathrm{E}[X_n] = 0$ but even S_n/n does not converge.

Exercice. Let X_i be IID random variables in $\mathcal{L}^2$. Define $Y_k = \frac{1}{k}\sum_{i=1}^k X_i$. What can you say about $S_n = \frac{1}{n}\sum_{k=1}^n Y_k$?

2.10 Birkhoff's ergodic theorem

In this section we fix a probability space $(\Omega, \mathcal{A}, \mathrm{P})$ and consider sequences of random variables X_n which are defined dynamically by a map T on Ω by

$$X_n(\omega) = X(T^n(\omega)) ,$$

where $T^n(\omega) = T(T(\dots T(\omega)))$ is the n'th iterate of ω. This can include as a special case the situation that the random variables are independent, but it can be much more general. Similarly as martingale theory covered later in these notes, ergodic theory is not only a generalization of classical probability theory, it is a considerable extension of it, both by language as by scope.

Definition. A measurable map $T : \Omega \to \Omega$ from the probability space onto itself is called **measure preserving**, if $\mathrm{P}[T^{-1}(A)] = \mathrm{P}[A]$ for all $A \in \mathcal{A}$. The map T is called **ergodic** if $T(A) = A$ implies $\mathrm{P}[A] = 0$ or $\mathrm{P}[A] = 1$. A measure preserving map T is called **invertible**, if there exists a measurable, measure preserving inverse T^{-1} of T. An invertible an measure preserving map T is also called an **automorphism** of the probability space.

Example. Let $\Omega = \{|z| = 1\} \subset \mathbb{C}$ be the unit circle in the complex plane with the measure $\mathrm{P}[\mathrm{Arg}(z) \in [a,b]] = (b-a)/(2\pi)$ for $0 < a < b < 2\pi$ and the Borel σ-algebra $\mathcal{A}$. If $w = e^{2\pi i \alpha}$ is a complex number of length 1, then the rotation $T(z) = wz$ defines a measure preserving transformation on $(\Omega, \mathcal{B}, \mathrm{P})$. It is invertible with inverse $T^{-1}(z) = z/w$.

Example. The transformation $T(z) = z^2$ on the same probability space as in the previous example is also measure preserving. Note that $\mathrm{P}[T(A)] = 2\mathrm{P}[A]$ but $\mathrm{P}[T^{-1}(A)] = \mathrm{P}[A]$ for all $A \in \mathcal{B}$. The map is measure preserving but it is not invertible.

Remark. T is ergodic if and only if for any $X \in \mathcal{L}^2$ the condition $X(T) = X$ implies that X is constant almost everywhere.

Example. The rotation on the circle is ergodic if α is irrational. Proof: with $z = e^{2\pi i x}$ one can write a random variable X on Ω as a Fourier series $f(z) = \sum_{n=-\infty}^{\infty} a_n z^n$ which is the sum $f_0+f_++f_-$, where $f_+ = \sum_{n=1}^{\infty} a_n z^n$ is analytic in $|z| < 1$ and $f_- = \sum_{n=1}^{\infty} a_n z^{-n}$ is analytic in $|z| > 1$ and f_0 is constant. By doing the same decomposition for $f(T(z)) = \sum_{n=-\infty}^{\infty} a_n w^n z^n$, we see that $f_+ = \sum_{n=1}^{\infty} a_n z^n = \sum_{n=1}^{\infty} a_n w^n z^n$. But these are the Taylor expansions of $f_+ = f_+(T)$ and so $a_n = a_n w^n$. Because $w^n \neq 1$ for irrational α, we deduce $a_n = 0$ for $n \geq 1$. Similarly, one derives $a_n = 0$ for $n \leq -1$. Therefore $f(z) = a_0$ is constant.

Example. Also the non-invertible squaring transformation $T(x) = x^2$ on the circle is ergodic as a Fourier argument shows again: T preserves again the decomposition of f into three analytic functions $f = f_- + f_0 + f_+$ so that $f(T(z)) = \sum_{n=-\infty}^{\infty} a_n z^{2n} = \sum_{n=-\infty}^{\infty} a_n z^n$ implies $\sum_{n=1}^{\infty} a_n z^{2n} = \sum_{n=1}^{\infty} a_n z^n$. Comparing Taylor coefficients of this identity for analytic functions shows $a_n = 0$ for odd n because the left hand side has zero Taylor coefficients for odd powers of z. But because for even $n = 2^l k$ with odd k, we have $a_n = a_{2^l k} = a_{2^{l-1} k} = \cdots = a_k = 0$, all coefficients $a_k = 0$ for $k \geq 1$. Similarly, one sees $a_k = 0$ for $k \leq -1$.

Definition. Given a random variable $X \in \mathcal{L}$, one obtains a sequence of random variables $X_n = X(T^n) \in \mathcal{L}$ by $X(T^n)(\omega) = X(T^n \omega)$. Define $S_0 = 0$ and $S_n = \sum_{k=0}^{n} X(T^k)$.

Theorem 2.10.1 (Maximal ergodic theorem of Hopf). Given $X \in \mathcal{L}^1$, the event $A = \{\sup_n S_n > 0 \}$ satisfies

$$\mathrm{E}[X; A] = \mathrm{E}[1_A X] \geq 0 .$$

Proof. Define $Z_n = \max_{0 \leq k \leq n} S_k$ and the sets $A_n = \{Z_n > 0\} \subset A_{n+1}$. Then $A = \bigcup_n A_n$. Clearly $Z_n \in \mathcal{L}^1$. For $0 \leq k \leq n$, we have $Z_n \geq S_k$ and so $Z_n(T) \geq S_k(T)$ and hence

$$Z_n(T) + X \geq S_{k+1} .$$

By taking the maxima on both sides over $0 \leq k \leq n$, we get

$$Z_n(T) + X \geq \max_{1 \leq k \leq n+1} S_k .$$

On $A_n = \{Z_n > 0\}$, we can extend this to $Z_n(T) + X \geq \max_{1 \leq k \leq n+1} S_k \geq \max_{0 \leq k \leq n+1} S_k = Z_{n+1} \geq Z_n$ so that on A_n

$$X \geq Z_n - Z_n(T) .$$

Integration over the set A_n gives

$$\mathrm{E}[X; A_n] \geq \mathrm{E}[Z_n; A_n] - \mathrm{E}[Z_n(T); A_n] .$$

Using (1) this inequality, the fact (2) that $Z_n = 0$ on $X \setminus A_n$, the (3) inequality $Z_n(T) \geq S_n(T) \geq 0$ on A_n and finally that T is measure preserving (4), leads to

$$\begin{array}{rcl} \mathrm{E}[X; A_n] & \geq_{(1)} & \mathrm{E}[Z_n; A_n] - \mathrm{E}[Z_n(T); A_n] \\ & =_{(2)} & \mathrm{E}[Z_n] - \mathrm{E}[Z_n(T); A_n] \\ & \geq_{(3)} & \mathrm{E}[Z_n - Z_n(T)] =_{(4)} 0 \end{array}$$

for every n and so to $\mathrm{E}[X; A] \geq 0$. □

Theorem 2.10.2 (Ergodic theorem of Birkhoff, 1931). For any $X \in \mathcal{L}^1$ the time average

$$\overline{S}_n = \frac{S_n}{n} = \frac{1}{n} \sum_{i=0}^{n-1} X(T^i x)$$

converges almost everywhere to a T-invariant random variable $\overline{X}$ satisfying $\mathrm{E}[X] = \mathrm{E}[\overline{X}]$. Especially, if T is ergodic, then S_n/n converges to $\mathrm{E}[X]$.

Proof. Define $\overline{X} = \limsup_{n\to\infty} \overline{S}_n$, $\underline{X} = \liminf_{n\to\infty} \overline{S}_n$. We get $\overline{X} = \overline{X}(T)$ and $\underline{X} = \underline{X}(T)$ because

$$\frac{n+1}{n}\overline{S}_n - \overline{S}_n(T) = \frac{X}{n} .$$

(i) $\overline{X} = \underline{X}$.
Define for $\alpha < \beta \in \mathbb{R}$ the sets $A_{\alpha,\beta} = \{\underline{X} < \beta, \alpha < \overline{X}\}$. Because $\{\underline{X} < \overline{X}\} = \bigcup_{\alpha<\beta,\alpha,\beta\in\mathbb{Q}} A_{\alpha,\beta}$, it is enough to show that $\mathrm{P}[A_{\alpha,\beta}] = 0$ for rational $\alpha < \beta$. Define

$$A = \{\sup_n(S_n - n\alpha) > 0\} = \{\sup_n(\overline{S}_n - \alpha) > 0\} .$$

Because $A_{\alpha,\beta} \subset A$ and $A_{\alpha,\beta}$ is T-invariant, we get from the maximal ergodic theorem $\mathrm{E}[X - \alpha, A_{\alpha,\beta}] \geq 0$ and so

$$\mathrm{E}[X, A_{\alpha,\beta}] \geq \alpha \cdot \mathrm{P}[A_{\alpha,\beta}] .$$

Replacing X, α, β with $-X, -\beta, -\alpha$ and using $\overline{-X} = -\underline{X}$, $\underline{-X} = -\overline{X}$ gives $\mathrm{E}[X; A_{\alpha,\beta}] \leq \beta \cdot \mathrm{P}[A_{\alpha,\beta}]$ and because $\beta < \alpha$, the claim $\mathrm{P}[A_{\alpha,\beta}] = 0$ follows.

(ii) $\overline{X} \in \mathcal{L}^1$.
$|\overline{S}_n| \leq |X|$, and $\overline{S}_n$ converges point wise to $\overline{X} = \underline{X}$ and $X \in \mathcal{L}^1$. Lebesgue's

dominated convergence theorem gives $\overline{X} \in \mathcal{L}$.

(iii) $\mathrm{E}[X] = \mathrm{E}[\overline{X}]$.
Define the sets $B_{k,n} = \{\overline{X} \in [\frac{k}{n}, \frac{k+1}{n})\}$ for $k \in \mathbb{Z}, n \geq 1$. Define for $\epsilon > 0$, $Y = X - \frac{k}{n} + \epsilon$. Using the maximal ergodic theorem, we get $\mathrm{E}[Y; B_{k,n}] \geq 0$. Because $\epsilon > 0$ was arbitrary,

$$\mathrm{E}[X; B_{k,n}] \geq \frac{k}{n} \cdot$$

With this inequality

$$\mathrm{E}[\overline{X}, B_{k,n}] \leq \frac{k+1}{n}\mathrm{P}[B_{k,n}] \leq \frac{1}{n}\mathrm{P}[B_{k,n}] + \mathrm{E}[X; B_{k,n}] \, .$$

Summing over k gives

$$\mathrm{E}[\overline{X}] \leq \frac{1}{n} + \mathrm{E}[X]$$

and because n was arbitrary, $\mathrm{E}[\overline{X}] \leq \mathrm{E}[X]$. Doing the same with $-X$ and using (i), we end with

$$\mathrm{E}[-\overline{X}] = \mathrm{E}[\underline{-X}] \leq \mathrm{E}[\overline{-X}] \leq \mathrm{E}[-X] \, .$$

□

Corollary 2.10.3. The strong law of large numbers holds for IID random variables $X_n \in \mathcal{L}^1$.

Proof. Given a sequence of IID random variables $X_n \in \mathcal{L}^1$. Let μ be the law of X_n. Define the probability space $\Omega = (\mathbb{R}^{\mathbb{Z}}, \mathcal{A}, \mathrm{P})$, where $\mathrm{P} = \mu^{\mathbb{Z}}$ is the product measure. If $T : \Omega \to \Omega$, $T(\omega)_n = \omega_{n+1}$ denotes the shift on Ω, then $X_n = X(T^n)$ with with $X(\omega) = \omega_0$. Since every T-invariant function is constant almost everywhere, we must have $\overline{X} = \mathrm{E}[X]$ almost everywhere, so that $S_n/n \to \mathrm{E}[X]$ almost everywhere. □

Remark. While ergodic theory is closely related to probability theory, the **notation** in the two fields is different. The reason is that the origin of the theories are different. One usually writes $(X, \mathcal{A}, m)$ for a probability space. An example of different language is also that ergodic theorists do not use the word "random variables" X but speak of "functions" f. Good introductions to ergodic theory are [36, 13, 8, 77, 54, 107].

2.11 More convergence results

We mention now some results about the almost everywhere convergence of sums of random variables in contrast to the weak and strong laws which were dealing with averaged sums.

Theorem 2.11.1 (Kolmogorov's inequalities). a) Assume $X_k \in \mathcal{L}^2$ are independent random variables. Then

$$\mathrm{P}[\sup_{1 \leq k \leq n} |S_k - \mathrm{E}[S_k]| \geq \epsilon] \leq \frac{1}{\epsilon^2} \mathrm{Var}[S_n] .$$

b) Assume $X_k \in \mathcal{L}^\infty$ are independent random variables and $||X_n||_\infty \leq R$. Then

$$\mathrm{P}[\sup_{1 \leq k \leq n} |S_k - \mathrm{E}[S_k]| \geq \epsilon] \geq 1 - \frac{(R+\epsilon)^2}{\sum_{k=1}^n \mathrm{Var}[X_k]} .$$

Proof. We can assume $\mathrm{E}[X_k] = 0$ without loss of generality.
a) For $1 \leq k \leq n$ we have

$$S_n^2 - S_k^2 = (S_n - S_k)^2 + 2(S_n - S_k)S_k = 2(S_n - S_k)S_k$$

and therefore $\mathrm{E}[S_n^2; A_k] \geq \mathrm{E}[S_k^2; A_k]$ for all $A_k \in \sigma(X_1, \dots, X_k)$ by the independence of $S_n - S_k$ and S_k. The sets $A_1 = \{|S_1| \geq \epsilon\}$, $A_{k+1} = \{|S_{k+1}| \geq \epsilon, \max_{i \leq l \leq k} |S_l| < \epsilon\}$ are mutually disjoint. We have to estimate the probability of the events

$$B_n = \{\max_{1 \leq k \leq n} |S_k| \geq \epsilon\} = \bigcup_{k=1}^n A_k .$$

We get

$$\mathrm{E}[S_n^2] \geq \mathrm{E}[S_n^2; B_n] = \sum_{k=1}^n \mathrm{E}[S_n^2; A_k] \geq \sum_{k=1}^n \mathrm{E}[S_k^2; A_k] \geq \epsilon^2 \sum_{k=1}^n \mathrm{P}[A_k] = \epsilon^2 \mathrm{P}[B_n] .$$

b)

$$\mathrm{E}[S_k^2; B_n] = \mathrm{E}[S_k^2] - \mathrm{E}[S_k^2; B_n^c] \geq \mathrm{E}[S_k^2] - \epsilon^2(1 - \mathrm{P}[B_n]) .$$

On A_k, $|S_{k-1}| \leq \epsilon$ and $|S_k| \leq |S_{k-1}| + |X_k| \leq \epsilon + R$ holds. We use that in the estimate

$$\begin{aligned} \mathrm{E}[S_n^2; B_n] &= \sum_{k=1}^{n} \mathrm{E}[S_k^2 + (S_n - S_k)^2; A_k] \\ &= \sum_{k=1}^{n} \mathrm{E}[S_k^2; A_k] + \sum_{k=1}^{n} \mathrm{E}[(S_n - S_k)^2; A_k] \\ &\leq (R+\epsilon)^2 \sum_{k=1}^{n} \mathrm{P}[A_k] + \sum_{k=1}^{n} \mathrm{P}[A_k] \sum_{j=k+1}^{n} \mathrm{Var}[X_j] \\ &\leq \mathrm{P}[B_n]((\epsilon + R)^2 + \mathrm{E}[S_n^2]) \end{aligned}$$

so that

$$\mathrm{E}[S_n^2] \leq \mathrm{P}[B_n]((\epsilon + R)^2 + \mathrm{E}[S_n^2]) + \epsilon^2 - \epsilon^2 \mathrm{P}[B_n] \,.$$

and so

$$\mathrm{P}[B_n] \geq \frac{\mathrm{E}[S_n^2] - \epsilon^2}{(\epsilon + R)^2 + \mathrm{E}[S_n] - \epsilon^2} \geq 1 - \frac{(\epsilon + R)^2}{(\epsilon + R)^2 + E[S_n^2] - \epsilon^2} \geq 1 - \frac{(\epsilon + R)^2}{\mathrm{E}[S_n^2]} \,.$$

□

Remark. The inequalities remain true in the limit $n \to \infty$. The first inequality is then

$$\mathrm{P}[\sup_k |S_k - \mathrm{E}[S_k]| \geq \epsilon] \leq \frac{1}{\epsilon^2} \sum_{k=1}^{\infty} \mathrm{Var}[X_k] \,.$$

Of course, the statement in *a*) is void, if the right hand side is infinite. In this case, however, the inequality in *b*) states that $\sup_k |S_k - \mathrm{E}[S_k]| \geq \epsilon$ almost surely for every $\epsilon > 0$.

Remark. For $n = 1$, Kolmogorov's inequality reduces to Chebychev's inequality (2.5.5)

Lemma 2.11.2. A sequence X_n of random variables converges almost everywhere, if and only if

$$\lim_{n \to \infty} \mathrm{P}[\sup_{k \geq 1} |X_{n+k} - X_n| > \epsilon] = 0$$

for all $\epsilon > 0$.

Proof. This is an exercise. □

Theorem 2.11.3 (Kolmogorov). Assume $X_n \in \mathcal{L}^2$ are independent and $\sum_{n=1}^\infty \mathrm{Var}[X_n] < \infty$. Then

$$\sum_{n=1}^{\infty}(X_n - \mathrm{E}[X_n])$$

converges almost everywhere.

Proof. Define $Y_n = X_n - \mathrm{E}[X_n]$ and $S_n = \sum_{k=1}^n Y_k$. Given $m \in \mathbb{N}$. Apply Kolmogorov's inequality to the sequence Y_{m+k} to get

$$\mathrm{P}[\sup_{n \geq m} |S_n - S_m| \geq \epsilon] \leq \frac{1}{\epsilon^2} \sum_{k=m+1}^{\infty} \mathrm{E}[X_k^2] \to 0$$

for $m \to \infty$. The above lemma implies that $S_n(\omega)$ converges. □

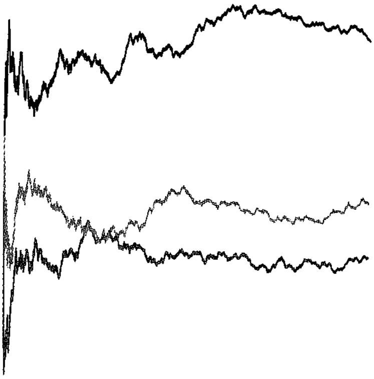


Figure. *We sum up independent random variables X_k which take values $\frac{\pm 1}{k^\alpha}$ with equal probability. According to theorem (2.11.3),the process*

$$S_n = \sum_{k=1}^{n}(X_k - \mathrm{E}[X_k]) = \sum_{k=1}^{n} X_k$$

converges if

$$\sum_{k=1}^{\infty} \mathrm{E}[X_k^2] = \sum_{k=1}^{\infty} \frac{1}{k^{2\alpha}}$$

converges. This is the case if $\alpha > 1/2$. The picture shows some experiments in the case $\alpha = 0.6$.

The following theorem gives a necessary and sufficient condition that a sum $S_n = \sum_{k=1}^n X_k$ converges for a sequence X_n of independent random variables.

Definition. Given $R \in \mathbb{R}$ and a random variable X, we define the bounded random variable

$$X^{(R)} = 1_{|X|<R} X .$$

Theorem 2.11.4 (Three series theorem). Assume $X_n \in \mathcal{L}$ be independent. Then $\sum_{n=1}^\infty X_n$ converges almost everywhere if and only if for some $R > 0$ all of the following three series converge:

$$\sum_{k=1}^\infty \mathrm{P}[|X_k| > R] \;<\; \infty\,, \tag{2.4}$$

$$\sum_{k=1}^\infty |\mathrm{E}[X_k^{(R)}]| \;<\; \infty\,, \tag{2.5}$$

$$\sum_{k=1}^\infty \mathrm{Var}[X_k^{(R)}] \;<\; \infty\,. \tag{2.6}$$

Proof. "$\Rightarrow$" Assume first that the three series all converge. By (3) and Kolmogorov's theorem, we know that $\sum_{k=1}^\infty X_k^{(R)} - \mathrm{E}[X_k^{(R)}]$ converges almost surely. Therefore, by (2), $\sum_{k=1}^\infty X_k^{(R)}$ converges almost surely. By (1) and Borel-Cantelli, $\mathrm{P}[X_k \neq X_k^{(R)}$ infinitely often$) = 0$. Since for almost all ω, $X_k^{(R)}(\omega) = X_k(\omega)$ for sufficiently large k and for almost all ω, $\sum_{k=1}^\infty X_k^{(R)}(\omega)$ converges, we get a set of measure one, where $\sum_{k=1}^\infty X_k$ converges.

"$\Leftarrow$" Assume now that $\sum_{n=1}^\infty X_n$ converges almost everywhere. Then $X_k \to 0$ almost everywhere and $\mathrm{P}[|X_k| > R$, infinitely often$) = 0$ for every $R > 0$. By the second Borel-Cantelli lemma, the sum (1) converges.

The almost sure convergence of $\sum_{n=1}^\infty X_n$ implies the almost sure convergence of $\sum_{n=1}^\infty X_n^{(R)}$ since $\mathrm{P}[|X_k| > R$, infinitely often$) = 0$.

Let $R > 0$ be fixed. Let Y_k be a sequence of independent random variables such that Y_k and $X_k^{(R)}$ have the same distribution and that all the random variables $X_k^{(R)}, Y_k$ are independent. The almost sure convergence of $\sum_{n=1}^\infty X_n^{(R)}$ implies that of $\sum_{n=1}^\infty X_n^{(R)} - Y_k$. Since $\mathrm{E}[X_k^{(R)} - Y_k] = 0$ and $\mathrm{P}[|X_k^{(R)} - Y_k| \leq 2R) = 1$, by Kolmogorov inequality b), the series $T_n = \sum_{k=1}^n X_k^{(R)} - Y_k$ satisfies for all $\epsilon > 0$

$$\mathrm{P}[\sup_{k\geq 1} |T_{n+k} - T_n| > \epsilon] \geq 1 - \frac{(R+\epsilon)^2}{\sum_{k=n}^\infty \mathrm{Var}[X_k^{(R)} - Y_k]}\,.$$

Claim: $\sum_{k=1}^\infty \mathrm{Var}[X_k^{(R)} - Y_k] < \infty$.

Assume, the sum is infinite. Then the above inequality gives $\mathrm{P}[\sup_{k\geq} |T_{n+k} - T_n| \geq \epsilon] = 1$. But this contradicts the almost sure convergence of $\sum_{k=1}^\infty X_k^{(R)} - Y_k$ because the latter implies by Kolmogorov inequality that $\mathrm{P}[\sup_{k\geq 1} |S_{n+k} - S_n| > \epsilon] < 1/2$ for large enough n. Having shown that $\sum_{k=1}^\infty (\mathrm{Var}[X_k^{(R)} - Y_k)] < \infty$, we are done because then by Kolmogorov's theorem (2.11.3), the sum $\sum_{k=1}^\infty X_k^{(R)} - \mathrm{E}[X_k^{(R)}]$ converges, so that (2) holds.

□

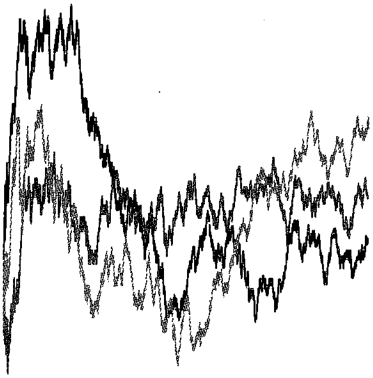


Figure. *A special case of the three series theorem is when X_k are uniformly bounded $X_k \leq R$ and have zero expectation* $\mathrm{E}[X_k] = 0$. *In that case, almost everywhere convergence of $S_n = \sum_{k=1}^{n} X_k$ is equivalent to the convergence of $\sum_{k=1}^{\infty} \mathrm{Var}[X_k]$. For example, in the case*

$$X_k = \left\{ \begin{array}{c} \frac{1}{k^\alpha} \\ -\frac{1}{k^\alpha} \end{array} \right. ,$$

and $\alpha = 1/2$, we do not have almost everywhere convergence of S_n, because $\sum_{k=1}^{\infty} \mathrm{Var}[X_k] = \sum_{k=1}^{\infty} \frac{1}{k} = \infty$.

Definition. A real number $\alpha \in \mathbb{R}$ is called a **median** of $X \in \mathcal{L}$ if $\mathrm{P}[X \leq \alpha] \geq 1/2$ and $\mathrm{P}[X \geq \alpha] \geq 1/2$. We denote by $\mathrm{med}(X)$ the set of medians of X.

Remark. The median is not unique and in general different from the mean. It is also defined for random variables for which the mean does not exist.

Proposition 2.11.5. (Comparing median and mean) For $Y \in \mathcal{L}^2$. Then every $\alpha \in \mathrm{med}(Y)$ satisfies

$$|\alpha - \mathrm{E}[Y]| \leq \sqrt{2}\sigma[Y] \, .$$

Proof. For every $\beta \in \mathbb{R}$, one has

$$\frac{|\alpha - \beta|^2}{2} \leq |\alpha - \beta|^2 \min(\mathrm{P}[Y \geq \alpha], \mathrm{P}[Y \leq \alpha]) \leq \mathrm{E}[(Y - \beta)^2] \, .$$

Now put $\beta = \mathrm{E}[Y]$. □

Theorem 2.11.6 (Lévy). Given a sequence $X_n \in \mathcal{L}$ which is independent. Choose $\alpha_{l,k} \in \mathrm{med}(S_l - S_k)$. Then, for all $n \in \mathbb{N}$ and all $\epsilon > 0$

$$\mathrm{P}[\max_{1 \leq k \leq n} |S_n + \alpha_{n,k}| \geq \epsilon] \leq 2\mathrm{P}[S_n \geq \epsilon] \, .$$

Proof. Fix $n \in \mathbb{N}$ and $\epsilon > 0$. The sets

$$A_1 = \{S_1 + \alpha_{n,1} \geq \epsilon\}, A_{k+1} = \{\max_{1\leq l\leq k}(S_n + \alpha_{n,l}) < \epsilon, S_{k+1} + \alpha_{n,k+1} \geq \epsilon\}$$

for $1 \leq k \leq n$ are disjoint and $\bigcup_{k=1}^n A_k = \{\max_{1\leq k\leq n}(S_k + \alpha_{n,k}) \geq \epsilon\}$. Because $\{S_n \geq \epsilon\}$ contains all the sets A_k as well as $\{S_n - S_k \geq \alpha_{n,k}\}$ for $1 \leq k \leq n$, we have using the independence of $\sigma(A_k)$ and $\sigma(S_n - S_k)$

$$\begin{aligned} \mathrm{P}[S_n \geq \epsilon] &\geq \sum_{k=1}^n \mathrm{P}[\{S_n - S_k \geq \alpha_{n,k}\} \cap A_k] \\ &= \sum_{k=1}^n \mathrm{P}[\{S_n - S_k \geq \alpha_{n,k}\}]\mathrm{P}[A_k] \\ &\geq \frac{1}{2}\sum_{k=1}^n \mathrm{P}[A_k] \\ &= \frac{1}{2}\mathrm{P}[\bigcup_{k=1}^n A_k] \\ &= \frac{1}{2}\mathrm{P}[\max_{1\leq k\leq n}(S_n + \alpha_{n,k}) \geq \epsilon] \,. \end{aligned}$$

Applying this inequality to $-X_n$, we get also $\mathrm{P}[-S_n - \alpha_{n,m} \geq -\epsilon] \geq 2\mathrm{P}[-S_n \geq -\epsilon]$ and so

$$\mathrm{P}[\max_{1\leq k\leq n}|S_n + \alpha_{n,k}| \geq \epsilon] \leq 2\mathrm{P}[S_n \geq \epsilon] \,.$$

□

Corollary 2.11.7. (Lévy) Given a sequence $X_n \in \mathcal{L}$ of independent random variables. If the partial sums S_n converge in probability to S, then S_n converges almost everywhere to S.

Proof. Take $\alpha_{l,k} \in \mathrm{med}(S_l - S_k)$. Since S_n converges in probability, there exists $m_1 \in \mathbb{N}$ such that $|\alpha_{l,k}| \leq \epsilon/2$ for all $m_1 \leq k \leq l$. In addition, there exists $m_2 \in \mathbb{N}$ such that $\sup_{n\geq 1}\mathrm{P}[|S_{n+m} - S_m| \geq \epsilon/2] < \epsilon/2$ for all $m \geq m_2$. For $m = \max\{m_1, m_2\}$, we have for $n \geq 1$

$$\mathrm{P}[\max_{1\leq l\leq n}|S_{l+m} - S_m| \geq \epsilon] \leq \mathrm{P}[\max_{1\leq l\leq n}|S_{l+m} - S_m + \alpha_{n+m,l+m}| \geq \epsilon/2] \,.$$

The right hand side can be estimated by theorem (2.11.6) applied to X_{n+m} with

$$\leq 2\mathrm{P}[|S_{n+m} - S_m| \geq \frac{\epsilon}{2}] < \epsilon \,.$$

Now apply the convergence lemma (2.11.2). □

Exercice. Prove the strong law of large numbers of independent but not necessarily identically distributed random variables: Given a sequence of independent random variables $X_n \in \mathcal{L}^2$ satisfying $\mathrm{E}[X_n] = m$. If

$$\sum_{k=1}^{\infty} \mathrm{Var}[X_k]/k^2 < \infty ,$$

then $S_n/n \to m$ almost everywhere.
Hint: Use Kolmogorov's theorem for $Y_k = X_k/k$.

Exercice. Let X_n be an IID sequence of random variables with uniform distribution on $[0, 1]$. Prove that almost surely

$$\sum_{n=1}^{\infty} \prod_{i=1}^{n} X_i < \infty .$$

Hint: Use $\mathrm{Var}[\prod_i X_i] = \prod \mathrm{E}[X_i^2] - \prod \mathrm{E}[X_i]^2$ and use the three series theorem.

2.12 Classes of random variables

The **probability distribution function** $F_X : \mathbb{R} \to [0, 1]$ of a random variable X was defined as

$$F_X(x) = \mathrm{P}[X \leq x] ,$$

where $\mathrm{P}[X \leq x]$ is a short hand notation for $\mathrm{P}[\{\omega \in \Omega \mid X(\omega) \leq x \}$. With the law $\mu_X = X^* P$ of X on $\mathbb{R}$ has $F_X(x) = \int_{-\infty}^{x} d\mu(x)$ so that F is the anti-derivative of μ. One reason to introduce distribution functions is that one can replace integrals on the probability space Ω by integrals on the real line $\mathbb{R}$ which is more convenient.

Remark. The distribution function F_X determines the law μ_X because the measure $\nu((-\infty, a]) = F_X(a)$ on the π-system $\mathcal{I}$ given by the intervals $\{(-\infty, a]\}$ determines a unique measure on $\mathbb{R}$. Of course, the distribution function does not determine the random variable itself. There are many different random variables defined on different probability spaces, which have the same distribution.

Proposition 2.12.1. The distribution function F_X of a random variable is

a) non-decreasing,
b) $F_X(-\infty) = 0, F_X(\infty) = 1$
c) continuous from the right: $F_X(x + h) = F_X$.

Furthermore, given a function F with the properties $a), b), c)$, there exists a random variable X on the probability space $(\Omega, \mathcal{A}, \mathrm{P})$ which satisfies $F_X = F$.

Proof. a) follows from $\{X \leq x\} \subset \{X \leq y\}$ for $x \leq y$. b) $\mathrm{P}[\{X \leq -n\}] \to 0$ and $\mathrm{P}[\{X \leq n\}] \to 1$. c) $F_X(x + h) - F_X = \mathrm{P}[x < X \leq x + h] \to 0$ for $h \to 0$.
Given F, define $\Omega = \mathbb{R}$ and $\mathcal{A}$ as the Borel σ-algebra on $\mathbb{R}$. The measure $\mathrm{P}[(-\infty, a]] = F[a]$ on the π-system $\mathcal{I}$ defines a unique measure on $(\Omega, \mathcal{A})$. □

Remark. Every Borel probability measure μ on $\mathbb{R}$ determines a distribution function F_X of some random variable X by

$$\int_{-\infty}^{x} d\mu(x) = F(x) .$$

The proposition tells also that one can define a class of **distribution functions**, the set of real functions F which satisfy properties $a), b), c)$.

Example. Bertrands paradox mentioned in the introduction shows that the choice of the distribution functions is important. In any of the three cases, there is a distribution function $f(x, y)$ which is radially symmetric. The constant distribution $f(x, y) = 1/\pi$ is obtained when we throw the center of the line into the disc. The disc A_r of radius r has probability $\mathrm{P}[A_r] = r^2/\pi$. The density in the r direction is $2r/\pi$. The distribution $f(x, y) = 1/r = 1/\sqrt{x^2 + y^2}$ is obtained when throwing parallel lines. This will put more weight to center. The probability $\mathrm{P}[A_r] = r/\pi$ is bigger than the area of the disc. The radial density is $1/\pi$. $f(x, y)$ is the distribution when we rotate the line around a point on the boundary. The disc A_r of radius r has probability $\arcsin(r)$. The density in the r direction is $1/\sqrt{1 - r^2}$.

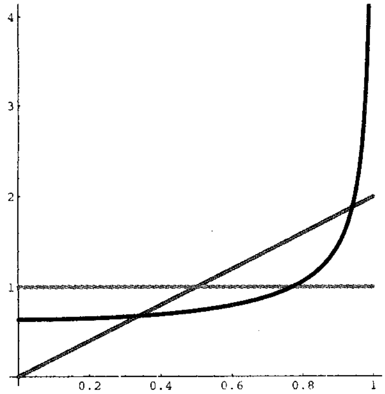


Figure. *A plot of the radial density function $f(r)$ for the three different interpretation of the Bertrand paradox.*

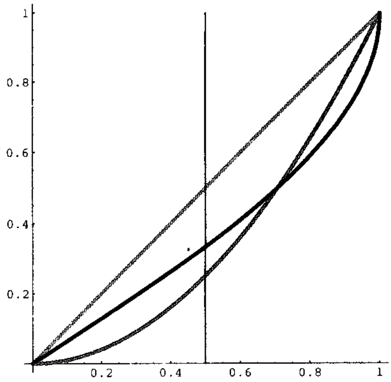


Figure. *A plot of the radial distribution function $F(r) = \mathrm{P}[A_r]$ There are different values at $F(1/2)$.*

So, what happens, if we really do an experiment and throw randomly lines onto a disc? The punch line of the story is that the outcome of the experiment very much depends on how the experiment will be performed. If we would do the experiment by hand, we would probably try to throw the center of the stick into the middle of the disc. Since we would aim to the center, the distribution would be different from any of the three solutions given in Bertrand's paradox.

Definition. A distribution function F is called **absolutely continuous** (ac), if there exists a Borel measurable function f satisfying $F(x) = \int_{-\infty}^{x} f(x)\, dx$. One calls a random variable with an absolutely continuous distribution function a **continuous random variable**.

Definition. A distribution function is called **pure point** (pp) or **atomic** if there exists a countable sequence of real numbers x_n and a sequence of positive numbers $p_n, \sum_n p_n = 1$ such that $F(x) = \sum_{n,x_n \leq x} p_n$. One calls a random variable with a discrete distribution function a **discrete random variable**.

Definition. A distribution function F is called **singular continuous** (sc) if F is continuous and if there exists a Borel set S of zero Lebesgue measure such that $\mu_F(S) = 1$. One calls a random variable with a singular continuous distribution function a **singular continuous random variable**.

Remark. The definition of (ac),(pp) and (sc) distribution functions is compatible for the definition of (ac),(pp) and (sc) Borel measures on $\mathbb{R}$. A Borel measure is (pp), if $\mu(A) = \sum_{x \in A} \mu(\{a\})$. It is continuous, if it contains no **atoms**, points with positive measure. It is (ac), if there exists a measurable function f such that $\mu = f\,dx$. It is (sc), if it is continuous and if $\mu(S) = 1$ for some Borel set S of zero Lebesgue measure.

The following decomposition theorem shows that these three classes are natural:

Theorem 2.12.2 (Lebesgue decomposition theorem). Every Borel measure μ on $(\mathbb{R}, \mathcal{B})$ can be decomposed in a unique way as $\mu = \mu_{pp} + \mu_{ac} + \mu_{sc}$, where μ_{pp} is pure point, μ_{sc} is singular continuous and μ_{ac} is absolutely continuous with respect to the Lebesgue measure λ.

Proof. Denote by λ the Lebesgue measure on $(\mathbb{R}, \mathcal{B})$ for which $\lambda([a,b]) = b-a$. We first show that any measure μ can be decomposed as $\mu = \mu_{ac}+\mu_s$, where μ_{ac} is absolutely continuous with respect to λ and μ_s is singular. The decomposition is unique: $\mu = \mu_{ac}^{(1)} + \mu_s^{(2)} = \mu_{ac}^{(2)} + \mu_s^{(2)}$ implies that $\mu_{ac}^{(1)} - \mu_{ac}^{(2)} = \mu_s^{(2)} - \mu_s^{(2)}$ is both absolutely continuous and singular continuous with respect to μ which is only possible, if they are zero. To get the existence of the decomposition, define $c = \sup_{A \in \mathcal{A}, \lambda(A)=0} \mu(A)$. If $c = 0$, then μ is absolutely continuous and we are done. If $c > 0$, take an increasing sequence $A_n \in \mathcal{B}$ with $\mu(A_n) \to c$. Define $A = \bigcup_{n \geq 1} A_n$ and μ_{ac} as $\mu_{ac}(B) = \mu(A \cap B)$. To split the singular part μ_s into a singular continuous and pure point part, we again have uniqueness because $\mu_s = \mu_{sc}^{(1)} + \mu_{sc}^{(2)} = \mu_{pp}^{(2)} + \mu_{pp}^{(2)}$ implies that $\nu = \mu_{sc}^{(1)} - \mu_{sc}^{(2)} = \mu_{pp}^{(2)} - \mu_{pp}^{(2)}$ are both singular continuous and pure point which implies that $\nu = 0$. To get existence, define the finite or countable set $A = \{\omega \mid \mu(\omega) > 0\}$ and define $\mu_{pp}(B) = \mu(A \cap B)$. □

Definition. The **Gamma function** is defined for $x > 0$ as

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t}\, dt \, .$$

It satisfies $\Gamma(n) = (n-1)!$ for $n \in \mathbb{N}$. Define also

$$B(p,q) = \int_0^1 x^{p-1}(1-x)^{q-1}\, dx \, ,$$

the **Beta function**.

Here are some examples of absolutely continuous distributions:

ac1) The **normal distribution** $N(m, \sigma^2)$ on $\Omega = \mathbb{R}$ has the probability density function

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-m)^2}{2\sigma^2}} \, .$$

ac2) The **Cauchy distribution** on $\Omega = \mathbb{R}$ has the probability density function

$$f(x) = \frac{1}{\pi} \frac{b}{b^2 + (x-m)^2} .$$

ac3) The **uniform distribution** on $\Omega = [a, b]$ has the probability density function

$$f(x) = \frac{1}{b-a} .$$

ac4) The **exponential distribution** $\lambda > 0$ on $\Omega = [0, \infty)$ has the probability density function

$$f(x) = \lambda e^{-\lambda x} .$$

ac5) The **log normal distribution** on $\Omega = [0, \infty)$ has the density function

$$f(x) = \frac{1}{\sqrt{2\pi x^2 \sigma^2}} e^{-\frac{(\log(x)-m)^2}{2\sigma^2}} .$$

ac6) The **beta distribution** on $\Omega = [0, 1]$ with $p > 1, q > 1$ has the density

$$f(x) = \frac{x^{p-1}(1-x)^{q-1}}{B(p,q)} .$$

ac7) The **Gamma distribution** on $\Omega = [0, \infty)$ with parameters $\alpha > 0, \beta > 0$

$$f(x) = \frac{x^{\alpha-1}\beta^{\alpha}e^{-x/\beta}}{\Gamma(\alpha)} .$$

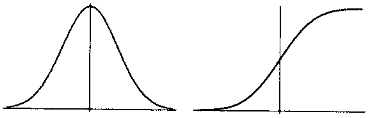


Figure. *The probability density and the CDF of the normal distribution.*

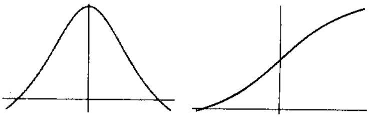


Figure. *The probability density and the CDF of the Cauchy distribution.*

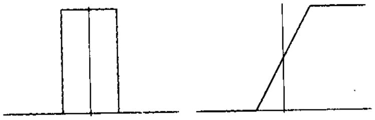


Figure. *The probability density and the CDF of the uniform distribution.*

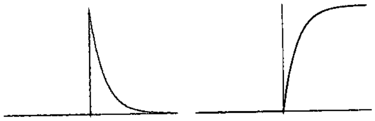


Figure. *The probability density and the CDF of the exponential distribution.*

Definition. We use the notation

$$\left(\begin{array}{c} n \\ k \end{array} \right) = \frac{n!}{(n-k)!k!}$$

for the **Binomial coefficient**, where $k! = k(k-1)(k-2)\cdots 2\cdot 1$ is the **factorial** of k with the convention $0! = 1$. For example,

$$\left(\begin{array}{c} 10 \\ 3 \end{array} \right) = \frac{10!}{7!3!} = 10 * 9 * 8/6 = 120 \, .$$

Examples of discrete distributions:

pp1) The **binomial distribution** on $\Omega = \{1, \dots, n\}$

$$\mathrm{P}[X = k] = \left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k}$$

pp2) The **Poisson distribution** on $\Omega = \mathbb{N}$

$$\mathrm{P}[X = k] = e^{-\lambda} \frac{\lambda^k}{k!}$$

pp3) The **Discrete uniform distribution** on $\Omega = \{1, .., n\}$

$$\mathrm{P}[X = k] = \frac{1}{n}$$

pp4) The **geometric distribution** on $\Omega = \mathbb{N} = \{0, 1, 2, 3, \dots\}$

$$\mathrm{P}[X = k] = p(1-p)^k$$

pp5) The distribution of **first success** on $\Omega = \mathbb{N} \setminus \{0\} = \{1, 2, 3, \dots\}$

$$\mathrm{P}[X = k] = p(1-p)^{k-1}$$

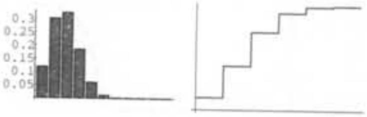


Figure. *The probabilities and the CDF of the binomial distribution.*

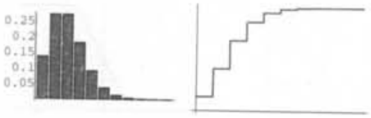


Figure. *The probabilities and the CDF of the Poisson distribution.*

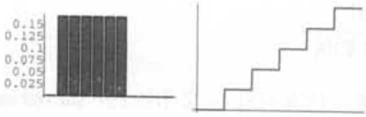


Figure. *The probabilities and the CDF of the uniform distribution.*

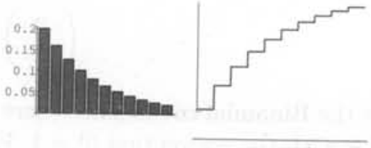


Figure. *The probabilities and the CDF of the geometric distribution.*

An example of a singular continuous distribution:

sc1) The **Cantor distribution.** Let $C = \bigcap_{n=0}^{\infty} \mathrm{E}_n$ be the Cantor set, where $E_0 = [0,1]$, $E_1 = [0,1/3] \cup [2/3,1]$ and E_n is inductively obtained by cutting away the middle third of each interval in E_{n-1}. Define

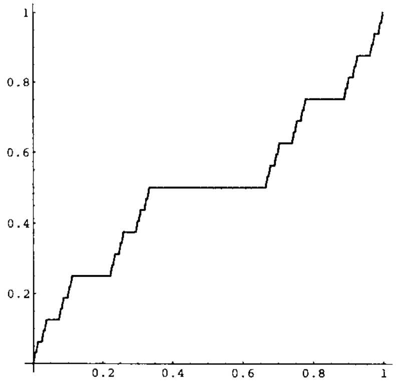
$$F(x) = \lim_{n \to \infty} F_n(x)$$

where $F_n(x)$ has the density $(3/2)^n \cdot 1_{\mathrm{E}_n}$. One can realize a random variable with the Cantor distribution as a sum of IID random variables as follows:

$$X = \sum_{n=1}^{\infty} \frac{X_n}{3^n} ,$$

where X_n take values 0 and 2 with probability $1/2$ each.

Figure. *The CDF of the Cantor distribution is continuous but not absolutely continuous. The function $F_X(x)$ is in this case called the* **Cantor function**. *Its graph is also called a* **Devils staircase**



Lemma 2.12.3. Given $X \in \mathcal{L}$ with law μ. For any measurable map $h : \mathbb{R}^1 \to [0, \infty)$ for which $h(X) \in \mathcal{L}^1$, one has $\mathrm{E}[h(X)] = \int_{\mathbb{R}} h(x) \, d\mu(x)$. Especially, if $\mu = \mu_{ac} = f \, dx$ then

$$\mathrm{E}[h(X)] = \int_{\mathbb{R}} h(x) f(x) \, dx \, .$$

If $\mu = \mu_{pp}$, then

$$\mathrm{E}[h(X)] = \sum_{x, \mu(\{x\}) \neq 0} h(x) \mu(\{x\}) \, .$$

Proof. If the function h is nonnegative, prove it first for $X = c1_{x \in A}$, then for step functions $X \in \mathcal{S}$ and then by the monotone convergence theorem for any $X \in \mathcal{L}$ for which $h(x) \in \mathcal{L}^1$. If $h(X)$ is integrable, then $\mathrm{E}[h(X)] = \mathrm{E}[h^+(X)] - \mathrm{E}[h^-(X)]$. □

Proposition 2.12.4.

Distribution	Parameters	Mean	Variance
ac1) Normal	$m \in \mathbb{R}$, $\sigma^2 > 0$	m	σ^2
ac2) Cauchy	$m \in \mathbb{R}$, $b > 0$	"m"	∞
ac3) Uniform	$a < b$	$(a+b)/2$	$(b-a)^2/12$
ac4) Exponential	$\lambda > 0$	$1/\lambda$	$1/\lambda^2$
ac5) Log-Normal	$m \in \mathbb{R}$, $\sigma^2 > 0$	$e^{\mu+\sigma^2/2}$	$(e^{\sigma^2} - 1)e^{2m+\sigma^2}$
ac6) Beta	$p, q > 0$	$p/(p+q)$	$\frac{pq}{(p+q)^2(p+q+1)}$
ac7) Gamma	$\alpha, \beta > 0$	$\alpha\beta$	$\alpha\beta^2$

Proposition 2.12.5.

pp1) Bernoulli	$n \in \mathbb{N}, p \in [0,1]$	np	$np(1-p)$
pp2) Poisson	$\lambda > 0$	λ	λ
pp3) Uniform	$n \in \mathbb{N}$	$(1+n)/2$	$(n^2-1)/12$
pp4) Geometric	$p \in (0,1)$	$(1-p)/p$	$(1-p)/p^2$
pp5) First Success	$p \in (0,1)$	$1/p$	$(1-p)/p^2$
sc1) Cantor	-	$1/2$	$1/8$

Proof. These are direct computations, which we do in some of the examples: **Exponential distribution**:

$$\mathrm{E}[X^p] = \int_0^\infty x^p \lambda e^{-\lambda x}\, dx = \frac{p}{\lambda}\mathrm{E}[X^{p-1}] = \frac{p!}{\alpha^p} \,.$$

Poisson distribution:

$$\mathrm{E}[X] = \sum_{k=0}^{\infty} k e^{-\lambda}\frac{\lambda^k}{k!} = \lambda e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!} = \lambda \,.$$

For calculating higher moments, one can also use the **probability generating function**

$$\mathrm{E}[z^X] = \sum_{k=0}^{\infty} e^{-\lambda}\frac{(\lambda z)^k}{k!} = e^{-\lambda(1-z)}$$

and then differentiate this identity with respect to z at the place $z = 0$. We get then

$$\mathrm{E}[X] = \lambda, \mathrm{E}[X(X-1)] = \lambda^2, \mathrm{E}[X^3] = \mathrm{E}[X(X-1)(X-2)], \dots$$

so that $\mathrm{E}[X^2] = \lambda + \lambda^2$ and $\mathrm{Var}[X] = \lambda$.

Geometric distribution. Differentiating the identity for the **geometric series**

$$\sum_{k=0}^{\infty} x^k = \frac{1}{1-x}$$

gives

$$\sum_{k=0}^{\infty} k x^{k-1} = \frac{1}{(1-x)^2} \,.$$

Therefore

$$\mathrm{E}[X_p] = \sum_{k=0}^{\infty} k(1-p)^k p = \sum_{k=0}^{\infty} k(1-p)^k p = p\sum_{k=1}^{\infty} k(1-p)^k = \frac{p}{p^2} = \frac{1}{p} \,.$$

For calculating the higher moments one can proceed as in the Poisson case or use the moment generating function.

Cantor distribution: because one can realize a random variable with the

Cantor distribution as $X = \sum_{n=1}^{\infty} X_n/3^n$, where the IID random variables X_n take the values 0 and 2 with probability $p = 1/2$ each, we have

$$\mathrm{E}[X] = \sum_{n=1}^{\infty} \frac{\mathrm{E}[X_n]}{3^n} = \sum_{n=1}^{\infty} \frac{1}{3^n} = \frac{1}{1-1/3} - 1 = \frac{1}{2}$$

and

$$\mathrm{Var}[X] = \sum_{n=1}^{\infty} \frac{\mathrm{Var}[X_n]}{3^n} = \sum_{n=1}^{\infty} \frac{\mathrm{Var}[X_n]}{9^n} = \sum_{n=1}^{\infty} \frac{1}{9^n} = \frac{1}{1-1/9} - 1 = \frac{9}{8} - 1 = \frac{1}{8} .$$

See also corollary (3.1.6) for an other computation. □

Computations can sometimes be done in an elegant way using **characteristic functions** $\phi_X(t) = \mathrm{E}[e^{itX}]$ or **moment generating functions** $M_X(t) = \mathrm{E}[e^{tX}]$. With the moment generating function one can get the moments with the **moment formula**

$$\mathrm{E}[X^n] = \int_{\mathbb{R}} x^n \, d\mu = \frac{d^n M_X}{dt^n}(t)|_{t=0} .$$

For the characteristic function one obtains

$$\mathrm{E}[X^n] = \int_{\mathbb{R}} x^n \, d\mu = (-i)^n \frac{d^n \phi_X}{dt^n}(t)|_{t=0} .$$

Example. The random variable $X(x) = x$ has the uniform distribution on $[0,1]$. Its moment generating function is $M_X(t) = \int_0^1 e^{tx} \, dx = (e^t - 1)/t = 1 + t/2! + t^2/3! + \dots$. A comparison of coefficients gives the moments $\mathrm{E}[X^m] = 1/(m+1)$, which agrees with the moment formula.

Example. A random variable X which has the Normal distribution $N(m, \sigma)$ has the moment generating function $M_X(t) = e^{tm + \sigma^2 t^2/2}$. All the **moments** can be obtained with the moment formula. For example, $\mathrm{E}[X] = M_X'(0) = m$, $\mathrm{E}[X^2] = M_X''(0) = m^2 + \sigma^2$.

Example. For a Poisson distributed random variable X on $\Omega = \mathbb{N} = \{0, 1, 2, 3, \dots\}$ with $\mathrm{P}[X = k] = e^{-\lambda}\frac{\lambda^k}{k!}$, the moment generating function is

$$M_X(t) = \sum_{k=0}^{\infty} \mathrm{P}[X = k] e^{tk} = e^{\lambda(1-e^t)} .$$

Example. A random variable X on $\Omega = \mathbb{N} = \{0, 1, 2, 3, \dots\}$ with the geometric distribution $\mathrm{P}[X = k] = p(1-p)^k$ has the moment generating function

$$M_X(t) = \sum_{k=0}^{\infty} e^{kt} p(1-p)^k = p \sum_{k=0}^{\infty} ((1-p)e^t)^k = \frac{p}{1-(1-p)e^t} .$$

A random variable X on $\Omega = \{1, 2, 3, \dots\}$ with the distribution of first success $\mathrm{P}[X = k] = p(1-p)^{k-1}$, has the moment generating function

$$M_X(t) = \sum_{k=1}^{\infty} e^{kt} p(1-p)^{k-1} = e^t p \sum_{k=0}^{\infty} ((1-p)e^t)^k = \frac{pe^t}{1-(1-p)e^t} .$$

Exercice. Compute the mean and variance of the **Erlang distribution**

$$f(x) = \frac{\lambda^k t^{k-1}}{(k-1)!} e^{-\lambda x}$$

on the positive real line $\Omega = [0, \infty)$ with the help of the moment generating function. If k is allowed to be an arbitrary positive real number, then the Erlang distribution is called the **Gamma distribution.**

Lemma 2.12.6. If X, Y are independent random variables, then their moment generating functions satisfy

$$M_{X+Y}(t) = M_X(t) \cdot M_Y(t) .$$

Proof. If X and Y are independent, then also e^{tX} and e^{tY} are independent. Therefore,

$$\mathrm{E}[e^{t(X+Y)}] = \mathrm{E}[e^{tX}e^{tY}] = \mathrm{E}[e^{tX}]\mathrm{E}[e^{tY}] = M_X(t) \cdot M_Y(t) .$$

□

Example. The lemma can be used to compute the moment generating function of the binomial distribution. A random variable X with binomial distribution can be written as a sum of IID random variables X_i taking values 0 and 1 with probability $1-p$ and p. Because for $n = 1$, we have $M_{X_i}(t) = (1-p) + pe^t$, the moment generating function of X is $M_X(t) = [(1-p) + pe^t]^n$. The moment formula allows us to compute moments $\mathrm{E}[X^n]$ and central moments $\mathrm{E}[(X - \mathrm{E}[X])^n]$ of X. Examples:

$$\begin{aligned}
\mathrm{E}[X] &= np \\
\mathrm{E}[X^2] &= np(1-p+np) \\
\mathrm{Var}[X] &= \mathrm{E}[(X-\mathrm{E}[X])^2] = \mathrm{E}[X^2] - \mathrm{E}[X]^2 = np(1-p) \\
\mathrm{E}[X^3] &= np(1+3(n-1)p+(2-3n+n^2)p^2) \\
\mathrm{E}[X^4] &= np(1+7(n-1)p+6(2-3n \\
&\quad +n^2)p^2+(-6+11n-6n^2+n^3)p^3) \\
\mathrm{E}[(X-E[X])^4] &= \mathrm{E}[X^4] - 8\mathrm{E}[X]\mathrm{E}[X^3] + 6\mathrm{E}[X^2]^2 + \mathrm{E}[X]^4 \\
&= np(1-p)(1+(5n-6)p-(-6+n+6n^2)p^2)
\end{aligned}$$

Example. The sum $X+Y$ of a Poisson distributed random variable X with parameter λ and a Poisson distributed random variable Y with parameter μ is Poisson distributed with parameter $\lambda+\mu$ as can be seen by multiplying their moment generating functions.

Definition. An interesting quantity for a random variable with a continuous distribution with probability density f_X is the **Shannon entropy** or simply **entropy**

$$H(X) = -\int_R f(x) \log(f(x)) \, dx \; .$$

Without restricting the class of functions, $H(X)$ is allowed to be $-\infty$ or ∞. The entropy allows to distinguish several distributions from others by asking for the distribution with the largest entropy. For example, among all distribution functions on the positive real line $[0,\infty)$ with fixed expectation $m = 1/\lambda$, the **exponential distribution** $\lambda e^{-\lambda}$ is the one with maximal entropy. We will return to these interesting entropy extremization questions later.

Example. Let us compute the entropy of the random variable $X(x) = x^m$ on $([0,1], \mathcal{B}, dx)$. We have seen earlier that the density of X is $f_X(x) = x^{1/m-1}/m$ so that

$$H(X) = -\int_0^1 (x^{1/m-1}/m) \log(x^{1/m-1}/m) \, dx \; .$$

To compute this integral, note first that $f(x) = x^a \log(x^a) = a x^a \log(x)$ has the antiderivative $ax^{1+a}((1+a)\log(x)-1)/(1+a)^2$ so that $\int_0^1 x^a \log(x^a) \, dx = -a/(1+a^2)$ and $H(X) = (1-m+\log(m))$. Because $\frac{d}{dm} H(X_m) = (1/m) - 1$ and $\frac{d^2}{dm^2} H(X_m) = -1/m^2$, the entropy has its maximum at $m=1$, where the density is uniform. The entropy decreases for $m \to \infty$. Among all random variables $X(x) = x^m$, the random variable $X(x) = x$ has maximal entropy.

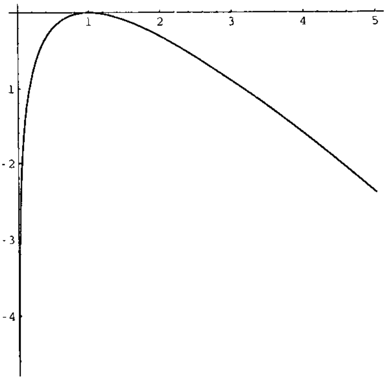


Figure. *The entropy of the random variables $X(x) = x^m$ on $[0,1]$ as a function of m. The maximum is attained for $m=1$, which is the* **uniform distribution**

2.13 Weak convergence

Definition. Denote by $C_b(\mathbb{R})$ the vector space of bounded continuous functions on $\mathbb{R}$. This means that $||f||_\infty = \sum_{x \in \mathbb{R}} |f(x)| < \infty$ for every $f \in C_b(\mathbb{R})$. A sequence of Borel probability measures μ_n on $\mathbb{R}$ **converges weakly** to a probability measure μ on $\mathbb{R}$ if for every $f \in C_b(\mathbb{R})$ one has

$$\int_{\mathbb{R}} f \, d\mu_n \to \int_X f \, d\mu$$

in the limit $n \to \infty$.

Remark. For weak convergence, it is enough to test $\int_{\mathbb{R}} f \, d\mu_n \to \int_X f \, d\mu$ for a dense set in $C_b(\mathbb{R})$. This dense set can consist of the space $P(\mathbb{R})$ of polynomials or the space $C_b^\infty(\mathbb{R})$ of bounded, smooth functions.
An important fact is that a sequence of random variables X_n converges in distribution to X if and only if $\mathrm{E}[h(X_n)] \to \mathrm{E}[h(X)]$ for all smooth functions h on the real line. This will be used the proof of the central limit theorem.

Weak convergence defines a topology on the set $M_1(\mathbb{R})$ of all Borel probability measures on $\mathbb{R}$. Similarly, one has a topology for $M_1([a,b])$.

Lemma 2.13.1. The set $M_1(I)$ of all probability measures on an interval $I = [a,b]$ is a compact topological space.

Proof. We need to show that any sequence μ_n of probability measures on I has an accumulation point. The set of functions $f_k(x) = x^k$ on $[a,b]$ span all polynomials and so a dense set in $C_b([a,b])$. The sequence μ_n converges to μ if and only if all the moments $\int_a^b x^k \, d\mu_n$ converge for $n \to \infty$ and for all $k \in \mathbb{N}$. In other words, the compactness of $M_1([a,b])$ is equivalent to the compactness of the product space $I^{\mathbb{N}}$ with the product topology, which is **Tychonovs theorem.** □

Remark. In functional analysis, a more general theorem called **Banach-Alaoglu theorem** is known: a closed and bounded set in the dual space X^* of a Banach space X is compact with respect to the **weak-* topology**, where the functionals μ_n converge to μ if and only if $\mu_n(f)$ converges to $\mu(f)$ for all $f \in X$. In the present case, $X = C_b[a,b]$ and the dual space X^* is the space of all **signed measures** on $[a,b]$ (see [7]).

Remark. The compactness of probability measures can also be seen by looking at the distribution functions $F_\mu(s) = \mu((-\infty, s])$. Given a sequence F_n of monotonically increasing functions, there is a subsequence F_{n_k} which converges to an other monotonically increasing function F, which is again a distribution function. This fact generalizes to distribution functions on the line where the limiting function F is still a right-continuous and nondecreasing function **Helly's selection theorem** but the function F does not need to be a distribution function any more, if the interval $[a, b]$ is replaced by the real line $\mathbb{R}$.

Definition. A sequence of random variables X_n converges **weakly** or **in law** to a random variable X, if the laws μ_{X_n} of X_n converge weakly to the law μ_X of X.

Definition. Given a distribution function F, we denote by $\text{Cont}(F)$ the set of continuity points of F.

Remark. Because F is nondecreasing and takes values in $[0, 1]$, the only possible discontinuity is a **jump discontinuity**. They happen at points t_i, where $a_i = \mu(\{t_i\}) > 0$. There can be only countably many such discontinuities, because for every rational number $p/q > 0$, there are only finitely many a_i with $a_i > p/q$ because $\sum_i a_i \leq 1$.

Definition. We say that a sequence of random variables X_n converges **in distribution** to a random variable X, if $F_{X_n}(x) \to F_X(x)$ point wise for all $x \in \text{Cont}(F)$.

Theorem 2.13.2 (Weak convergence = convergence in distribution). A sequence X_n of random variables converges in law to a random variable X if and only if X_n converges in distribution to X.

Proof. (i) Assume we have convergence in law. We want to show that we have convergence in distribution. Given $s \in \text{Cont}(f)$ and $\delta > 0$. Define a continuous function $1_{(-\infty,s]} \leq f \leq 1_{(-\infty,s+\delta]}$. Then

$$F_n(s) = \int_{\mathbb{R}} 1_{(-\infty,s]} \, d\mu_n \leq \int_{\mathbb{R}} f \, d\mu_n \leq \int_{\mathbb{R}} 1_{(-\infty,s+\delta]} \, d\mu_n = F_n(s+\delta) \; .$$

This gives

$$\limsup_{n\to\infty} F_n(s) \leq \lim_{n\to\infty} \int f \, d\mu_n = \int f \, d\mu \leq F(x+\delta) \; .$$

Similarly, we obtain with a function $1_{(-\infty,s-\delta]} \leq f \leq 1_{(-\infty,s]}$

$$\liminf_{n\to\infty} F_n(s) \geq \lim_{n\to\infty} \int f \, d\mu_n = \int f \, d\mu \geq F(s-\delta) \; .$$

Since F is continuous at x we have for $\delta \to 0$:

$$F(s) = \lim_{\delta\to 0} F(s-\delta) \leq \liminf_{n\to\infty} F_n(s) \leq \limsup_{n\to\infty} F_n(s) \leq F(s) \; .$$

That is we have established convergence in distribution.
(ii) Assume now we have no convergence in law. There exists then a continuous function f so that $\int f \; d\mu_n$ to $\int f \; d\mu$ fails. That is, there is a subsequence and $\epsilon > 0$ such that $|\int f \; d\mu_{n_k} - \int f \; d\mu| \geq \epsilon > 0$. There exists a compact interval I such that $|\int_I f \; d\mu_{n_k} - \int_I f \; d\mu| \geq \epsilon/2 > 0$ and we can assume that μ_{n_k} and μ have support on I. The set of all probability measures on I is compact in the weak topology. Therefore, a subsequence of $d\mu_{n_k}$ converges weakly to a measure ν and $|\nu(f) - \mu(f)| \geq \epsilon/2$. Define the π-system $\mathcal{I}$ of all intervals $\{(-\infty, s] \mid s$ continuity point of $F\ \}$. We have $\mu_n((-\infty, s]) = F_{X_n}(s) \to F_X(s) = \mu(-\infty, s])$. Using (i) we see $\mu_{n_k}((-\infty, s]) \to \nu(-\infty, s]$ also, so that μ and ν agree on the π system $\mathcal{I}$. If μ and ν agree on $\mathcal{I}$, they agree on the π-system of all intervals $\{(-\infty, s]\}$. By lemma (2.1.4), we know that $\mu = \nu$ on the Borel σ-algebra and so $\mu = \nu$. This contradicts $|\nu(f) - \mu(f)| \geq \epsilon/2$. So, the initial assumption of having no convergence in law was wrong. □

2.14 The central limit theorem

Definition. For any random variable X with non-zero variance, we denote by

$$X^* = \frac{(X - \mathrm{E}[X])}{\sigma(X)}$$

the **normalized random variable**, which has mean $\mathrm{E}[X^*] = 0$ and variance $\sigma(X^*) = \sqrt{\mathrm{Var}[X^*]} = 1$. Given a sequence of random variables X_k, we again use the notation $S_n = \sum_{k=1}^n X_k$.

Theorem 2.14.1 (Central limit theorem for independent L^3 random variables). Assume $X_i \in \mathcal{L}^3$ are independent and satisfy

$$M = \sup_i ||X_i||_3 < \infty, \; \delta = \liminf_{n \to \infty} \frac{1}{n} \sum_{i=1}^n \mathrm{Var}[X_i] > 0 \; .$$

Then S_n^* converges in distribution to a random variable with standard normal distribution $N(0,1)$:

$$\lim_{n \to \infty} \mathrm{P}[S_n^* \leq x] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-y^2/2} \, dy, \;\; \forall x \in \mathbb{R} \; .$$

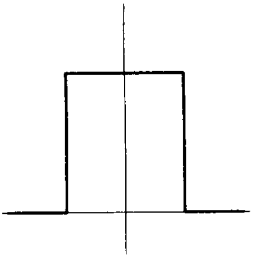


Figure. *The probability density function $f_{S_1^*}$ of the random variable $X(x) = x$ on $[-1,1]$.*

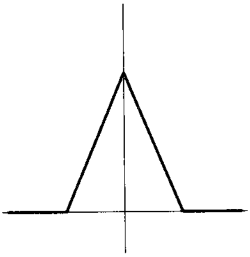


Figure. *The probability density function $f_{S_2^*}$ of the random variable $X(x) = x$ on $[-1,1]$.*

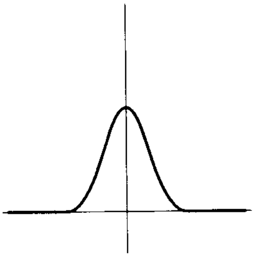


Figure. *The probability density function $F_{S_3^*}$ of the random variable $X(x) = x$ on $[-1,1]$.*

Lemma 2.14.2. A $N(0,\sigma^2)$ distributed random variable X satisfies

$$\mathrm{E}[|X|^p] = \frac{1}{\sqrt{\pi}} 2^{p/2} \sigma^p \Gamma(\frac{1}{2}(p+1)) \; .$$

Especially $\mathrm{E}[|X|^3] = \sqrt{\frac{8}{\pi}}\sigma^3$.

Proof. With the density function $f(x) = (2\pi\sigma^2)^{-1/2} e^{-\frac{x^2}{2\sigma^2}}$, we have $\mathrm{E}[|X|^p] = 2\int_0^\infty x^p f(x)\, dx$ which is after a substitution $z = x^2/(2\sigma^2)$ equal to

$$\frac{1}{\sqrt{\pi}} 2^{p/2} \sigma^p \int_0^\infty x^{\frac{1}{2}(p+1)-1} e^{-x} \, dx \; .$$

The integral to the right is by definition equal to $\Gamma(\frac{1}{2}(p+1))$. □

After this preliminary computation, we turn to the proof of the central limit theorem.

Proof. Define for fixed $n \geq 1$ the random variables

$$Y_i = \frac{(X_i - \mathrm{E}[X_i])}{\sigma(S_n)}, \; 1 \leq i \leq n$$

so that $S_n^* = \sum_{i=1}^n Y_i$. Define $N(0,\sigma)$-distributed random variables $\tilde{Y}_i$ having the property that the set of random variables

$$\{Y_1, \dots, Y_n, \tilde{Y}_1, \dots \tilde{Y}_n \}$$

are independent. The distribution of $\tilde{S}_n = \sum_{i=1}^n \tilde{Y}_i$ is just the normal distribution $N(0,1)$. In order to show the theorem, we have to prove $\mathrm{E}[f(S_n^*)] -$

$\mathrm{E}[f(\tilde{S}_n)] \to 0$ for any $f \in C_b(\mathbb{R})$. It is enough to verify it for smooth f. Define

$$Z_k = \tilde{Y}_1 + \ldots \tilde{Y}_{k-1} + Y_{k+1} + \cdots + Y_n \; .$$

Note that $Z_1 + Y_1 = S_n^*$ and $Z_n + \tilde{Y}_n = \tilde{S}_n$. Using first a telescopic sum and then Taylor's theorem, we can write

$$\begin{aligned} f(S_n^*) - f(\tilde{S}_n) &= \sum_{k=1}^{n} [f(Z_k + Y_k) - f(Z_k + \tilde{Y}_k)] \\ &= \sum_{k=1}^{n} [f'(Z_k)(Y_k - \tilde{Y}_k)] + \sum_{k=1}^{n} [\frac{1}{2} f''(Z_k)(Y_k^2 - \tilde{Y}_k^2)] \\ &\quad + \sum_{k=1}^{n} [R(Z_k, Y_k) + R(Z_k, \tilde{Y}_k)] \end{aligned}$$

with a Taylor rest term $R(Z, Y)$, which can depend on f. We get therefore

$$|E[f(S_n^*)] - \mathrm{E}[f(\tilde{S}_n)]| \leq \sum_{k=1}^{n} \mathrm{E}[|R(Z_k, Y_k)|] + \mathrm{E}[|R(Z_k, \tilde{Y}_k)|] \; . \qquad (2.7)$$

Because $\tilde{Y}_k$ are $N(0, \sigma^2)$-distributed, we get by lemma (2.14.2) and the Jensen inequality (2.5.1)

$$\mathrm{E}[|\tilde{Y}_k|^3] = \sqrt{\frac{8}{\pi}} \sigma^3 = \sqrt{\frac{8}{\pi}} \mathrm{E}[|Y_k|^2]^{3/2} \leq \sqrt{\frac{8}{\pi}} \mathrm{E}[|Y_k|^3] \; .$$

Taylor's theorem gives $|R(Z_k, Y_k)| \leq \mathrm{const} \cdot |Y_k|^3$ so that

$$\begin{aligned} \sum_{k=1}^{n} \mathrm{E}[|R(Z_k, Y_k)|] + \mathrm{E}[|R(Z_k, \tilde{Y}_k)|] &\leq \mathrm{const} \cdot \sum_{k=1}^{n} \mathrm{E}[|Y_k|^3] \\ &\leq \mathrm{const} \cdot n \cdot \sup_i ||X_i||_3 / \mathrm{Var}[S_n]^{3/2} \\ &= \mathrm{const} \cdot \frac{\sup_i ||X_i||_3}{(\mathrm{Var}[S_n]/n)^{3/2}} \cdot \frac{1}{\sqrt{n}} \\ &= \frac{M}{\delta^{3/2}} \frac{1}{\sqrt{n}} = \frac{C(f)}{\sqrt{n}} \to 0 \; . \end{aligned}$$

We have seen that for every smooth $f \in C_b(\mathbb{R})$ there exists a constant $C(f)$ such that $|E[f(S_n^*)] - \mathrm{E}[f(\tilde{S}_n)]| \leq C(f)/\sqrt{n}$. □

if we assume the X_i to be identically distributed, we can relax the condition $X_i \in \mathcal{L}^3$ to $X_i \in \mathcal{L}^2$:

Theorem 2.14.3 (Central limit theorem for IID L^2 random variables). If $X_i \in \mathcal{L}^2$ are IID and satisfy $0 < \mathrm{Var}[X_i]$, then S_n^* converges weakly to a random variable with standard normal distribution $N(0, 1)$.

Proof. The same proof gives equation (5.4). We change the estimation of Taylor $|R(z,y)| \leq \delta(y) \cdot y^2$ with $\delta(y) \to 0$ for $|y| \to 0$. Using the IID property and using dominated convergence we can estimate the rest term

$$R = \sum_{k=1}^{n} \mathrm{E}[|R(Z_k, Y_k)|] + \mathrm{E}[|R(Z_k, \tilde{Y}_k)|]$$

as follows:

$$\begin{aligned}
R &\leq \sum_{k=1}^{n} \mathrm{E}[\delta(Y_k)Y_k^2] + \mathrm{E}[\delta(\tilde{Y}_k)\tilde{Y}_k^2] \\
&= n \cdot \mathrm{E}[\delta(\frac{X_1}{\sigma\sqrt{n}})\frac{X_1^2}{\sigma^2 n}] + n \cdot \mathrm{E}[\delta(\frac{\tilde{X}_1}{\sigma\sqrt{n}})\frac{\tilde{X}_1^2}{\sigma^2 n}] \\
&\leq n \cdot \mathrm{E}[\delta(\frac{X_1}{\sigma\sqrt{n}})]\mathrm{E}[\frac{X_1^2}{\sigma^2 n}] + n \cdot \mathrm{E}[\delta(\frac{\tilde{X}_1}{\sigma\sqrt{n}})]\mathrm{E}[\frac{\tilde{X}_1^2}{\sigma^2 n}] \\
&= \mathrm{E}[\delta(\frac{X_1}{\sigma\sqrt{n}})]\mathrm{E}[\frac{X_1^2}{\sigma^2}] + \mathrm{E}[\delta(\frac{\tilde{X}_1}{\sigma\sqrt{n}})]\mathrm{E}[\frac{\tilde{X}_1^2}{\sigma^2}] \\
&\leq \mathrm{E}[\delta(\frac{X_1}{\sigma\sqrt{n}})]C \to 0
\end{aligned}$$

□

The central limit theorem can be interpreted as a solution to a fixed point problem:

Definition. Let $\mathcal{P}_{0,1}$ be the space of probability measure μ on $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ which have the properties that $\int_{\mathbb{R}} x^2 \, d\mu(x) = 1$, $\int_{\mathbb{R}} x \, d\mu(x) = 0$. Define the map

$$T\mu(A) = \int_{\mathbb{R}} \int_{\mathbb{R}} 1_A(\frac{x+y}{\sqrt{2}}) \mu(dx) \, \mu(dy)$$

on $\mathcal{P}_{0,1}$.

Corollary 2.14.4. The only attracting fixed point of T on $\mathcal{P}_{0,1}$ is the law of the standard normal distribution.

Proof. If μ is the law of a random variable X with $\mathrm{Var}[X] = 1$ and $\mathrm{E}[X] = 0$. Then $T(\mu)$ is the law of the normalized random variable $(X+X)/\sqrt{2}$ because the independent random variables X, Y can be realized on the probability space $(\mathbb{R}^2, \mathcal{B}, \mu \times \mu)$ as coordinate functions $X((x,y)) = x, Y((x,y)) = y$. Then $T(\mu)$ is obviously the law of $(X+Y)/\sqrt{2}$. Now use that $T^n(X) = (S_{2^n})^*$ converges in distribution to $N(0,1)$. □

For independent $0-1$ experiments with win probability $p \in (0,1)$, the central limit theorem is quite old. In this case

$$\lim_{n\to\infty} \mathrm{P}[\frac{(S_n-np)}{\sqrt{np(1-p)}} \leq x] = \frac{1}{2\pi}\int_{-\infty}^{x} e^{-y^2/2}\, dy$$

as had been shown by de Moivre in 1730 in the case $p = 1/2$ and for general $p \in (0,1)$ by Laplace in 1812. It is a direct consequence of the central limit theorem:

Corollary 2.14.5. (DeMoivre-Laplace limit theorem) The distribution of X_n^* converges to the normal distribution if X_n has the binomial distribution $B(n,p)$.

For more general versions of the central limit theorem, see [105]. The next limit theorem for discrete random variables illustrates, why Poisson distribution on $\mathbb{N}$ is natural. Denote by $B(n,p)$ the binomial distribution on $\{1,\dots,n\}$ and with P_α the Poisson distribution on $\mathbb{N} \setminus \{0\}$.

Theorem 2.14.6 (Poisson limit theorem). Let X_n be a $B(n,p_n)$-distributed and suppose $np_n \to \alpha$. Then X_n converges in distribution to a random variable X with Poisson distribution with parameter α.

Proof. We have to show that $\mathrm{P}[X_n = k] \to \mathrm{P}[X = k]$ for each fixed $k \in \mathbb{N}$.

$$\begin{aligned} \mathrm{P}[X_n = k] &= \binom{n}{k} p_n^k (1-p_n)^{n-k} \\ &= \frac{n(n-1)(n-2)\dots(n-k+1)}{k!} p_n^k (1-p_n)^{n-k} \\ &\sim \frac{1}{k!}(np_n)^k (1-\frac{np_n}{n})^{n-k} \to \frac{\alpha^k}{k!} e^{-\alpha} . \end{aligned}$$

□

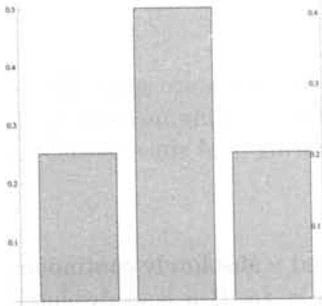


Figure. *The binomial distribution $B(2,1/2)$ has its support on $\{0,1,2\}$.*

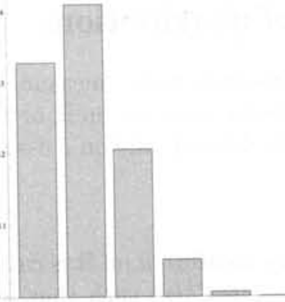


Figure. *The binomial distribution $B(5,1/5)$ has its support on $\{0,1,2,3,4,5\}$.*

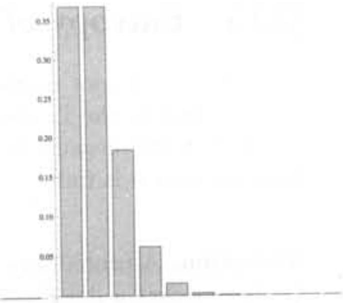


Figure. *The Poisson distribution with $\alpha = 1$ on $\mathbb{N} = \{0,1,2,3,\dots\}$.*

Exercice. It is custom to use the notation

$$\Phi(s) = F_X(s) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{s} e^{-y^2/2}\, dy$$

for the distribution function of a random variable X which has the standard normal distribution $N(0,1)$. Given a sequence of IID random variables X_n with this distribution.
a) Justify that one can estimate for large n probabilities

$$\mathrm{P}[a \le S_n^* \le b] \sim \Phi(b) - \Phi(a) .$$

b) Assume X_i are all uniformly distributed random variables in $[0,1]$. Estimate for large n

$$\mathrm{P}[|S_n/n - 0.5| \ge \epsilon]$$

in terms of Φ, ϵ and n.
c) Compare the result in b) with the estimate obtained in the weak law of large numbers.

Exercice. Define for $\lambda > 0$ the transformation

$$T_\lambda(\mu)(A) = \int_{\mathbb{R}} \int_{\mathbb{R}} 1_A(\frac{x+y}{\lambda})\, d\mu(x)\, d\mu(y)$$

in $\mathcal{P} = M_1(\mathbb{R})$, the set of all Borel probability measures on $\mathbb{R}$. For which λ can you describe the limit?

2.15 Entropy of distributions

Denote by ν a (not necessarily finite) measure on a measure space $(\Omega, \mathcal{A})$. An example is the Lebesgue measure on $\mathbb{R}$ or the counting measure on $\mathbb{N}$. Note that the measure is defined only on a δ-subring of $\mathcal{A}$ since we did not assume that ν is finite.

Definition. A probability measure μ on $\mathbb{R}$ is called ν **absolutely continuous**, if there exists a density $f \in \mathcal{L}^1(\nu)$ such that $\mu = f\nu$. If μ is ν-absolutely continuous, one writes $\mu \ll \nu$. Call $\mathcal{P}(\nu)$ the set of all ν absolutely continuous measures. The set $\mathcal{P}(\nu)$ is the set of functions $f \in \mathcal{L}^1(\nu)$ satisfying $f \geq 0$ and $\int f(x)\, d\nu(x) = 1$.

Remark. The fact that $\mu \ll \nu$ defined earlier is equivalent to this is called the Radon-Nykodym theorem ([?]). The function f is therefore called the **Radon-Nykodym derivative** of μ with respect to ν.

Example. If ν is the counting measure $\mathbb{N} = \{0, 1, 2, \dots\}$ and ν is the law of the geometric distribution with parameter p, then the density is $f(k) = p(1-p)^k$.

Example. If ν is the Lebesgue measure on $(-\infty, \infty)$ and μ is the law of the standard normal distribution, then the density is $f(x) = e^{-x^2/2}/\sqrt{2\pi}$. There is a multi-variable calculus trick using polar coordinates, which immediately shows that f is a density:

$$\int\int_{\mathbb{R}^2} e^{-(x^2+y^2)/2}\, dxdy = \int_0^\infty \int_0^{2\pi} e^{-r^2/2}\, rd\theta dr = 2\pi\ .$$

Definition. For any probability measure $\mu \in \mathcal{P}(\nu)$ define the **entropy**

$$H(\mu) = \int_\Omega -f(\omega) \log(f(\omega))\, d\nu(\omega)\ .$$

It generalizes the earlier defined Shannon entropy, where the assumption had been $d\nu = dx$.

Example. Let ν be the counting measure on a countable set Ω, where $\mathcal{A}$ is the σ-algebra of all subsets of Ω and let the measure ν is defined on the δ-ring of all finite subsets of Ω. In this case,

$$H(\mu) = \sum_{\omega \in \Omega} -f(\omega) \log(f(\omega))\ .$$

For example, for $\Omega = \mathbb{N} = \{0, 1, 2, 3, \dots\}$ with counting measure ν, the **geometric distribution** $\mathrm{P}[\{k\}] = p(1-p)^k$ has the entropy

$$\sum_{k=0}^\infty -(1-p)^k p \log((1-p)^k p) = \log(\frac{1-p}{p}) - \frac{\log(1-p)}{p}\ .$$

Example. Let ν be the Lebesgue measure on $\mathbb{R}$. If $\mu = f dx$ has a density function f, we have

$$H(\mu) = \int_{\mathbb{R}} -f(x)\log(f(x))\, dx\ .$$

For example, for the standard normal distribution μ with probability density function $f(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$, the entropy is $H(f) = (1 + \log(2\pi))/2$.

Example. If ν is the Lebesgue measure dx on $\Omega = \mathbb{R}^+ = [0, \infty)$. A random variable on Ω with probability density function $f(x) = \lambda e^{-\lambda x}$ is called the **exponential distribution**. It has the mean $1/\lambda$. The entropy of this distribution is $(\log(\lambda) - 1)/\lambda$.

Example. If ν is a probability measure on $\mathbb{R}$, f a density and

$$\mathcal{A} = \{A_1, \dots, .A_n\ \}$$

is a partition on $\mathbb{R}$. For the step function

$$\tilde{f} = \sum_{i=1}^n (\int_{A_i} f\ d\nu) 1_{A_i} \in \mathcal{S}(\nu)\ ,$$

the entropy $H(\tilde{f}\nu)$ is equal to

$$H(\{A_i\}) = \sum_i -\nu(A_i)\log(\nu(A_i))$$

which is called the **entropy of the partition** $\{A_i\}$. The approximation of the density f by a step functions $\tilde{f}$ is called **coarse graining** and the entropy of $\tilde{f}$ is called the **coarse grained entropy**. It has first been considered by Gibbs in 1902.

Remark. In **ergodic theory**, where one studies measure preserving transformations T of probability spaces, one is interested in the growth rate of the entropy of a partition generated by $\mathcal{A}, T(\mathcal{A}), .., T^n(\mathcal{A})$. This leads to the notion of an **entropy of a measure preserving transformation** called **Kolmogorov-Sinai entropy**.

Interpretation. Assume that Ω is finite and that ν the counting measure and $\mu(\{\omega\}) = f(\omega)$ the probability distribution of random variable describing the measurement of an experiment. If the event $\{\omega\}$ happens, then $-\log(f(\omega))$ is a measure for the **information or "surprise"** that the event happens. The averaged information or surprise is

$$H(\mu) = \sum_\omega -f(\omega)\log(f(\omega))\ .$$

If f takes only the values 0 or 1, which means that μ is **deterministic**, then $H(\mu) = 0$. There is no surprise and what the measurements show, is the reality. On the other hand, if f is the uniform distribution on Ω, then $H(\mu) = \log(|\Omega|)$. We will see in a moment that this is the maximal entropy.

Definition. Given two probability measures $\mu = f\nu$ and $\tilde{\mu} = \tilde{f}\nu$ which are both absolutely continuous with respect to ν. Define the **relative entropy**

$$H(\tilde{\mu}|\mu) = \int \tilde{f}(\omega) \log(\frac{\tilde{f}(\omega)}{f(\omega)}) \, d\nu(x) \quad \in [0, \infty] .$$

It is the expectation $\mathrm{E}_{\tilde{\mu}}[l]$ of the **Likelihood coefficient** $l = \log(\frac{\tilde{f}(x)}{f(x)})$. The negative relative entropy $-H(\tilde{\mu}|\mu)$ is also called the **conditional entropy**. One writes also $H(f|\tilde{f})$ instead of $H(\tilde{\mu}|\mu)$.

Theorem 2.15.1 (Gibbs inequality). $0 \leq H(\tilde{\mu}|\mu) \leq +\infty$ and $H(\tilde{\mu}|\mu) = 0$ if and only if $\mu = \tilde{\mu}$.

Proof. We can assume $H(\tilde{\mu}|\mu) < \infty$. The function $u(x) = x \log(x)$ is convex on $\mathbb{R}^+ = [0, \infty)$ and satisfies $u(x) \geq x - 1$.

$$H(\tilde{\mu}|\mu) = \int_\Omega \tilde{f}(\omega) u(\frac{\tilde{f}(x)}{f(x)}) \, d\nu \geq \int_\Omega f(\omega)(\frac{\tilde{f}(x)}{f(x)} - 1) \, d\nu = 0 .$$

If $\mu = \tilde{\mu}$, then $f = \tilde{f}$ almost everywhere and $H(\tilde{\mu}|\mu) = 0$.
On the other hand, if $H(\tilde{\mu}|\mu) = 0$, then by the Jensen inequality (2.5.1)

$$0 = \mathrm{E}_\mu[u(\frac{\tilde{f}}{f})] \geq u(\mathrm{E}_\mu[\frac{\tilde{f}}{f}]) = u(1) = 0 .$$

Therefore, $\mathrm{E}_\mu[u(\frac{\tilde{f}}{f})] = u(\mathrm{E}_\mu[\frac{\tilde{f}}{f}])$. The strict convexity of u implies that $\frac{\tilde{f}}{f}$ must be a constant. Since both f and $\tilde{f}$ are densities, we have $f = \tilde{f}$. □

Remark. The relative entropy can be used to measure the distance between two distributions. It is not a metric although. The relative entropy is also known under the name **Kullback-Leibler divergence** or **Kullback-Leibler metric**, if $\nu = dx$ [85].

Theorem 2.15.2 (Distributions with maximal entropy). The following distributions have maximal entropy.
a) If Ω is finite with counting measure ν. The **uniform distribution** on Ω has maximal entropy among all distributions on Ω. It is unique with this property.
b) $\Omega = \mathbb{N}$ with counting measure ν. The **geometric distribution** with parameter $p = c^{-1}$ has maximal entropy among all distributions on $\mathbb{N} = \{0, 1, 2, 3, \dots\}$ with fixed mean c. It is unique with this property.
c) $\Omega = \{0, 1\}^N$ with counting measure ν. The **product distribution** η^N, where $\eta(1) = p, \eta(0) = 1 - p$ with $p = c/N$ has maximal entropy among all distributions satisfying $\mathrm{E}[S_N] = c$, where $S_N(\omega) = \sum_{i=1}^N \omega_i$. It is unique with this property.
d) $\Omega = [0, \infty)$ with Lebesgue measure ν. The **exponential distribution** with density $f(x) = \lambda e^{-\lambda x}$ with parameter λ on Ω has the maximal entropy among all distributions with fixed mean $c = 1/\lambda$. It is unique with this property.
e) $\Omega = \mathbb{R}$ with Lebesgue measure ν. The **normal distribution** $N(m, \sigma^2)$ has maximal entropy among all distributions with fixed mean m and fixed variance σ^2. It is unique with this property.
f) **Finite measures.** Let $(\Omega, \mathcal{A})$ be an arbitrary measure space for which $0 < \nu(\Omega) < \infty$. Then the measure ν with uniform distribution $f = 1/\nu(\Omega)$ has maximal entropy among all other measures on Ω. It is unique with this property.

Proof. Let $\mu = f\nu$ be the measure of the distribution from which we want to prove maximal entropy and let $\tilde{\mu} = \tilde{f}\nu$ be any other measure. The aim is to show $H(\tilde{\mu}|\mu) = H(\mu) - H(\tilde{\mu})$ which implies the maximality since by the Gibbs inequality lemma (2.15.1) $H(\tilde{\mu}|\mu) \geq 0$.
In general,

$$H(\tilde{\mu}|\mu) = -H(\tilde{\mu}) - \int_\Omega \tilde{f}(\omega) \log(f(\omega)) \, d\nu$$

so that in each case, we have to show

$$H(\mu) = -\int_\Omega \tilde{f}(\omega) \log(f(\omega)) \, d\nu \, . \tag{2.8}$$

With

$$H(\tilde{\mu}|\mu) = H(\mu) - H(\tilde{\mu})$$

we also have uniqueness: if two measures $\tilde{\mu}, \mu$ have maximal entropy, then $H(\tilde{\mu}|\mu) = 0$ so that by the Gibbs inequality lemma (2.15.1) $\mu = \tilde{\mu}$.

a) The density $f = 1/|\Omega|$ is constant. Therefore $H(\mu) = \log(|\Omega|)$ and equation (2.8) holds.

b) The geometric distribution on $\mathbf{N} = \{0, 1, 2, \dots\}$ satisfies $\mathrm{P}[\{k\}] = f(k) = p(1-p)^k$, so that

$$\begin{aligned}\int_\Omega \tilde{f}(\omega) \log(f(\omega))\, d\nu &= \log(p) + \int \tilde{f}(\omega) k \log(1-p)\, d\nu \\ &= \log(p) - \log(1-p) \int_\Omega \tilde{f}(\omega)\, d\nu(\omega) \\ &= \log(\frac{p}{(1-p)})\end{aligned}$$

which is also the entropy of μ.

c) The discrete density is $f(\omega) = p^{S_N}(1-p)^{N-S_N}$ so that

$$\log(f(k)) = S_N \log(p) + (N - S_N)\log(1-p)$$

and

$$\sum_k \tilde{f}(k) \log(f(k)) = \mathrm{E}[S_N]\log(p) + (N - \mathrm{E}[S_N])\log(1-p) \; .$$

The claim follows since we fixed $\mathrm{E}[S_N]$.

d) The density is $f(x) = \alpha e^{-\alpha x}$, so that $\log(f(x)) = \log(\alpha) - \alpha x$. The claim follows since we fixed $\mathrm{E}[X] = \int x\, d\tilde{\mu}(x)$ was assumed to be fixed for all distributions.

e) For the normal distribution $\log(f(x)) = a + b(x-m)^2$ with two real number a, b depending only on m and σ. The claim follows since we fixed $\mathrm{Var}[X] = \mathrm{E}[(x-m)^2]$ for all distributions.

f) The density $f = 1$ is constant. Therefore $H(\mu) = 0$ which is also on the right hand side of equation (2.8). □

Remark. This result has relations to the foundations of **thermodynamics**, where one considers the phase space of N particles moving in a finite region in Euclidean space. The energy surface is then a compact surface Ω and the motion on this surface leaves a measure ν invariant which is induced from the flow invariant Lebesgue measure. The measure ν is called the **micro-canonical ensemble**. According to $f)$ in the above, it is the measure which maximizes entropy.

Remark. Let us try to get the maximal distribution using **calculus of variations**. In order to find the maximum of the **functional**

$$H(f) = -\int f \log(f)\, d\nu$$

on $\mathcal{L}^1(\nu)$ under the constraints

$$F(f) = \int_\Omega f\, d\nu = 1, \; G(f) = \int_\Omega X f\, d\nu = c \, ,$$

we have to find the critical points of $\tilde{H} = H - \lambda F - \mu G$ In infinite dimensions, **constrained critical points** are points, where the **Lagrange equations**

$$\begin{aligned} \frac{\partial}{\partial f}H(f) &= \lambda\frac{\partial}{\partial f}F(f) + \mu\frac{\partial}{\partial f}G(f) \\ F(f) &= 1 \\ G(f) &= c \end{aligned}$$

are satisfied. The derivative $\partial/\partial f$ is the **functional derivative** and λ, μ are the **Lagrange multipliers**. We find (f, λ, ν) as a solution of the system of equations

$$\begin{aligned} -1 - \log(f(x)) &= \lambda + \mu x, \\ \int_\Omega f(x)\, d\nu(x) &= 1, \\ \int_\Omega x f(x)\, d\nu(x) &= c \end{aligned}$$

by solving the first equation for f:

$$\begin{aligned} f &= e^{-\lambda-\mu x+1} \\ \int e^{-\lambda-\mu x+1}\, d\nu(x) &= 1 \\ \int x e^{-\lambda-\mu x+1}\, d\nu(x) &= c \end{aligned}$$

dividing the third equation by the second, so that we can get μ from the equation $\int x e^{-\mu x} x\, d\nu(x) = c \int e^{-\mu(x)}\, d\nu(x)$ and λ from the third equation $e^{1+\lambda} = \int e^{-\mu x}\, d\nu(x)$. This variational approach produces critical points of the entropy. Because the Hessian $D^2(H) = -1/f$ is negative definite, it is also negative definite when restricted to the surface in $\mathcal{L}^1$ determined by the restrictions $F = 1, G = c$. This indicates that we have found a **global maximum**.

Example. For $\Omega = \mathbb{R}$, $X(x) = x^2$, we get the normal distribution $N(0,1)$.

Example. For $\Omega = \mathbb{N}$, $X(n) = \epsilon_n$, we get $f(n) = e^{-\epsilon_n \lambda_1}/Z(f)$ with $Z(f) = \sum_n e^{-\epsilon_n \lambda_1}$ and where λ_1 is determined by $\sum_n \epsilon_n e^{-\epsilon_n \lambda_1} = c$. This is called the **discrete Maxwell-Boltzmann distribution**. In physics, one writes $\lambda^{-1} = kT$ with the **Boltzmann constant** k, determining T, the **temperature**.

Here is a dictionary matching some notions in probability theory with corresponding terms in statistical physics. The statistical physics jargon is often more intuitive.

Probability theory	Statistical mechanics
Set Ω	Phase space
Measure space	Thermodynamic system
Random variable	Observable (for example energy)
Probability density	Thermodynamic state
Entropy	Boltzmann-Gibbs entropy
Densities of maximal entropy	Thermodynamic equilibria
Central limit theorem	Maximal entropy principle

Distributions, which maximize the entropy possibly under some constraint are mathematically natural because they are critical points of a variational principle. Physically, they are natural, because nature prefers them. From the statistical mechanical point of view, the extremal properties of entropy offer insight into thermodynamics, where large systems are modeled with statistical methods. Thermodyanamic equilibria often extremize variational problems in a given set of measures.

Definition. Given a measure space $(\Omega, \mathcal{A})$ with a not necessarily finite measure ν and a random variable $X \in \mathcal{L}$. Given $f \in \mathcal{L}^1$ leading to the probability measure $\mu = f\nu$. Consider the moment generating function $Z(\lambda) = \mathrm{E}_\mu[e^{\lambda X}]$ and define the interval $\Lambda = \{\lambda \in \mathbb{R} \mid Z(\lambda) < \infty \}$ in $\mathbb{R}$. For every $\lambda \in \Lambda$ we can define a new probability measure

$$\mu_\lambda = f_\lambda \nu = \frac{e^{\lambda X}}{Z(\lambda)} \mu$$

on Ω. The set

$$\{\mu_\lambda \mid \lambda \in \Lambda \}$$

of measures on $(\Omega, \mathcal{A})$ is called the **exponential family** defined by ν and X.

Theorem 2.15.3 (Minimizing relative entropy). For all probability measures $\tilde{\mu}$ which are absolutely continuous with respect to ν, we have for all $\lambda \in \Lambda$

$$H(\tilde{\mu}|\mu) - \lambda \mathrm{E}_{\tilde{\mu}}[X] \geq -\log Z(\lambda) .$$

The minimum $-\log Z(\lambda)$ is obtained for μ_λ.

Proof. For every $\tilde{\mu} = \tilde{f}\nu$, we have

$$\begin{aligned} H(\tilde{\mu}|\mu) &= \int_\Omega \tilde{f} \log(\frac{\tilde{f}}{\tilde{f}_\lambda}) \cdot \frac{\tilde{f}_\lambda}{f} \, d\nu \\ &= -H(\tilde{\mu}|\mu_\lambda) + (-\log(Z(\lambda)) + \lambda \mathrm{E}_{\tilde{\mu}}[X]) . \end{aligned}$$

For $\tilde{\mu} = \mu_\lambda$, we have

$$H(\mu_\lambda|\mu) = -\log(Z(\lambda)) + \lambda \mathrm{E}_{\mu_\lambda}[X] .$$

Therefore

$$H(\tilde{\mu}|\mu) - \lambda \mathrm{E}_{\tilde{\mu}}[X] = H(\tilde{\mu}|\mu_\lambda) - \log(Z(\lambda)) \geq -\log Z(\lambda) .$$

The minimum is obtained for $\tilde{\mu} = \mu_\lambda$. □

Corollary 2.15.4. (Minimizers for relative entropy)
a) μ_λ minimizes the relative entropy $\tilde{\mu} \mapsto H(\tilde{\mu}|\mu)$ among all ν-absolutely continuous measures $\tilde{\mu}$ with fixed $\mathrm{E}_{\tilde{\mu}}[X]$.
b) If we fix λ by requiring $\mathrm{E}_{\mu_\lambda}[X] = c$, then μ_λ maximizes the entropy $H(\tilde{\mu})$ among all measures $\tilde{\mu}$ satisfying $\mathrm{E}_{\tilde{\mu}}[X] = c$.

Proof. a) Minimizing $\tilde{\mu} \mapsto H(\tilde{\mu}|\mu)$ under the constraint $\mathrm{E}_{\tilde{\mu}}[X] = c$ is equivalent to minimize

$$H(\tilde{\mu}|\mu) - \lambda \mathrm{E}_{\tilde{\mu}}[X],$$

and to determine the Lagrange multiplier λ by $\mathrm{E}_{\mu_\lambda}[X] = c$. The above theorem shows that μ_λ is minimizing that.
b) If $\mu = f\nu, \mu_\lambda = e^{-\lambda X} f/Z$, then

$$0 \leq H(\tilde{\mu}, \mu_\lambda) = -H(\tilde{\mu}) + (-\log(Z)) - \lambda \mathrm{E}_\mu[X] = -H(\tilde{\mu}) + H(\mu_\lambda) .$$

□

Corollary 2.15.5. If $\nu = \mu$ is a probability measure, then μ_λ maximizes

$$F(\mu) = H(\mu) + \lambda \mathrm{E}_\mu[X]$$

among all measures $\tilde{\mu}$ which are absolutely continuous with respect to μ.

Proof. Take $\mu = \nu$. Since then $f = 1$, $H(\tilde{\mu}|\mu) = -H(\tilde{\mu})$. The claim follows from the theorem since a minimum of $H(\tilde{\mu}|\mu) - \lambda \mathrm{E}_{\tilde{\mu}}[X]$ corresponds to a maximum of $F(\mu)$. □

This corollary can also be proved by calculus of variations, namely by finding the minimum of $F(f) = \int f \log(f) + Xf \, d\nu$ under the constraint $\int f \, d\nu = 1$.

Remark. In statistical mechanics, the measure μ_λ is called the **Gibbs distribution** or **Gibbs canonical ensemble** for the observable X and $Z(\lambda)$ is called the **partition function**. In physics, one uses the notation $\lambda = -(kT)^{-1}$, where T is the temperature. Maximizing $H(\mu) - (kT)^{-1}\mathrm{E}_\mu[X]$ is the same as minimizing $\mathrm{E}_\mu[X] - kTH(\mu)$ which is called the **free energy** if X is the **Hamiltonian** and $\mathrm{E}_\mu[X]$ is the **energy**. The measure μ is the **a priori model**, the **micro canonical ensemble**. Adding the restriction that X has a specific expectation value $c = \mathrm{E}_\mu[X]$ leads to the probability measure μ_λ, the **canonical ensemble**. We illustrated two physical principles: nature maximizes entropy when the energy is fixed and minimizes the free energy, when energy is not fixed.

Example. Take on the real line the Hamiltonian $X(x) = x^2$ and a measure $\mu = f dx$, we get the energy $\int x^2 \, d\mu$. Among all symmetric distributions fixing the energy, the Gaussian distribution maximizes the entropy.

Example. Let $\Omega = \mathbb{N} = \{0, 1, 2, \dots \}$ and $X(k) = k$ and let ν be the counting measure on Ω and μ the Poisson measure with parameter 1. The partition function is

$$Z(\lambda) = \sum_k e^{\lambda k} \frac{e^{-1}}{k!} = \exp(e^\lambda - 1)$$

so that $\Lambda = \mathbb{R}$ and μ_λ is given by the weights

$$\mu_\lambda(k) = \exp(e^{-\lambda} + 1) e^{\lambda k} \frac{e^{-1}}{k!} = e^{-\alpha} \frac{\alpha^k}{k!} ,$$

where $\alpha = e^\lambda = 0$. The exponential family of the Poisson measure is the family of all Poisson measures.

Example. The geometric distribution on $\mathbb{N} = \{0, 1, 2, 3, \dots \}$ is an exponential family.

Example. The product measure on $\Omega = \{0, 1 \}^N$ with win probability p is an exponential family with respect to $X(k) = k$.

Example. $\Omega = \{1, \dots, N\}$, ν the counting measure and let μ_p be the binomial distribution with p. Take $\mu = \mu_{1/2}$ and $X(k) = k$. Since

$$\begin{aligned} 0 \quad &\leq \quad H(\tilde{\mu}|\mu) = H(\tilde{\mu}|\mu_p) + \log(p)\mathrm{E}[X] + \log(1-p)\mathrm{E}[(N - \mathrm{E}[X])] \\ &= \quad -H(\tilde{\mu}|\mu_p) + H(\mu_p) , \end{aligned}$$

μ_p is an exponential family.

Remark. There is an obvious generalization of the maximum entropy principle to the case, when we have finitely many random variables $\{X_i\}_{i=1}^n$. Given $\mu = f\nu$ we define the (n-dimensional) exponential family

$$\mu_\lambda = f_\lambda \nu = \frac{e^{\sum_{i=1}^n \lambda_i X_i}}{Z(\lambda)} \mu ,$$

where

$$Z(\lambda) = \mathrm{E}_\mu[e^{\sum_{i=1}^n \lambda_i X_i}]$$

is the partition function defined on a subset Λ of $\mathbb{R}^n$.

Theorem 2.15.6. For all probability measures $\tilde{\mu}$ which are absolutely continuous with respect to ν, we have for all $\lambda \in \Lambda$

$$H(\tilde{\mu}|\mu) - \sum_i \lambda_i \mathrm{E}_{\tilde{\mu}}[X_i] \geq -\log Z(\lambda) .$$

The minimum $-\log Z(\lambda)$ is obtained for μ_λ. If we fix λ_i by requiring $\mathrm{E}_{\mu_\lambda}[X_i] = c_i$, then μ_λ maximizes the entropy $H(\tilde{\mu})$ among all measures $\tilde{\mu}$ satisfying $\mathrm{E}_{\tilde{\mu}}[X_i] = c_i$.
Assume $\nu = \mu$ is a probability measure. The measure μ_λ maximizes

$$F(\tilde{\mu}) = H(\tilde{\mu}) + \lambda \mathrm{E}_{\tilde{\mu}}[X] .$$

Proof. Take the same proofs as before by replacing λX with $\lambda \cdot X = \sum_i \lambda_i X_i$. □

2.16 Markov operators

Definition. Given a not necessarily finite probability space $(\Omega, \mathcal{A}, \nu)$. A linear operator $P : \mathcal{L}^1(\Omega) \to \mathcal{L}^1(\Omega)$ is called a **Markov operator**, if

$$\boxed{\begin{array}{l} f \geq 0 \Rightarrow Pf \geq 0, \\ f \geq 0 \Rightarrow ||Pf||_1 = ||f||_1. \end{array}}$$

Remark. In other words, a Markov operator P has to leave the closed **positive cone** invariant $\mathcal{L}^1_+ = \{f \in \mathcal{L}^1 \mid f \geq 0\}$ and preserve the norm on that cone.

Remark. A Markov operator on $(\Omega, \mathcal{A}, \nu)$ leaves invariant the set $\mathcal{D}(\nu) = \{f \in \mathcal{L}^1 \mid f \geq 0, ||f||_1 = 1\}$ of **probability densities**. They correspond bijectively to the set $\mathcal{P}(\nu)$ of probability measures which are absolutely continuous with respect to ν. A Markov operator is therefore also called a **stochastic operator**.

Example. Let T be a measure preserving transformation on $(\Omega, \mathcal{A}, \nu)$. It is called **nonsingular** if $T^*\nu$ is absolutely continuous with respect to ν. The unique operator $P : \mathcal{L}^1 \to \mathcal{L}^1$ satisfying

$$\int_A Pf \, d\nu = \int_{T^{-1}A} f \, d\nu$$

is called the **Perron-Frobenius operator** associated to T. It is a Markov operator. Closely related is the operator $Pf(x) = f(Tx)$ for measure preserving invertible transformations. This **Koopman operator** is often studied on $\mathcal{L}^2$, but it becomes a Markov operator when considered as a transformation on $\mathcal{L}^1$.

Exercice. Assume $\Omega = [0,1]$ with Lebesgue measure μ. Verify that the Perron-Frobenius operator for the tent map

$$T(x) = \begin{cases} 2x & ,x \in [0,1/2] \\ 2(1-x) & ,x \in [1/2,1] \end{cases}$$

is $Pf(x) = \frac{1}{2}(f(\frac{1}{2}x) + f(1 - \frac{1}{2}x))$.

Here is an abstract version of the Jensen inequality (2.5.1). It is due to M. Kuczma. See [61].

Theorem 2.16.1 (Jensen inequality for positive operators). Given a convex function u and an operator $P : \mathcal{L}^1 \to \mathcal{L}^1$ mapping positive functions into positive functions satisfying $P1 = 1$, then

$$u(Pf) \leq Pu(f)$$

for all $f \in \mathcal{L}^1_+$ for which $Pu(f)$ exists.

Proof. We have to show $u(Pf)(\omega) \leq Pu(f)(\omega)$ for almost all $\omega \in \Omega$. Given $x = (Pf)(\omega)$, there exists by definition of convexity a linear function $y \mapsto ay + b$ such that $u(x) = ax + b$ and $u(y) \geq ay + b$ for all $y \in \mathbb{R}$. Therefore, since $af + b \leq u(f)$ and P is positive

$$u(Pf)(\omega) = a(Pf)(\omega) + b = P(af + b)(\omega) \leq P(u(f))(\omega) .$$

□

The following theorem states that relative entropy does not increase along orbits of Markov operators. The assumption that $\{f > 0 \}$ is mapped into itself is actually not necessary, but simplifies the proof.

Theorem 2.16.2 (Voigt, 1981). Given a Markov operator $\mathcal{P}$ which maps $\{f > 0\}$ into itself. For all $f, g \in \mathcal{L}^1_+$,

$$H(\mathcal{P}f|\mathcal{P}g) \leq H(f|g) .$$

Proof. We can assume that $\{g(\omega) = 0\} \subset A = \{f(\omega) = 0\}$ because nothing is to show in the case $H(f|g) = \infty$. By restriction to the measure space space $(A^c, \mathcal{A} \cap A^c, \nu(\cdot \cap A))$, we can assume $f > 0, g > 0$ so that by our assumption also $Pf > 0$ and $Pg > 0$.

(i) Assume first $(f/g)(\omega) \leq c$ for some constant $c \in \mathbb{R}$.
For fixed g, the linear operator $Rh = P(hg)/P(g)$ maps positive functions into positive functions. Take the convex function $u(x) = x\log(x)$ and put $h = f/g$. Using Jensen's inequality, we get

$$\frac{Pf}{Pg}\log\frac{Pf}{Pg} = u(Rh) \leq Ru(h) = \frac{P(f\log(f/g))}{Pg}$$

which is equivalent to $Pf\log\frac{Pf}{Pg} \leq P(f\log(f/g))$. Integration gives

$$\begin{aligned} H(Pf|Pg) &= \int Pf\log\frac{Pf}{Pg}\,d\nu \\ &\leq \int P(f\log(f/g))\,d\nu = \int f\log(f/g)\,d\nu = H(f|g)\,. \end{aligned}$$

(ii) Define $f_k = \inf(f, kg)$ so that $f_k/g \leq k$. We have $f_k \subset f_{k+1}$ and $f_k \to f$ in $\mathcal{L}^1$. From (i) we know that $H(Pf_k|Pg) \leq H(f_k|g)$. We can assume $H(f|g) < \infty$ because the result is trivially true in the other case. Define $B = \{f \leq g\}$. On B, we have $f_k\log(f_k/g) = f\log(f/g)$ and on $\Omega\setminus B$ we have

$$f_k\log(f_k/g) \leq f_{k+1}\log(f_{k+1}/g)u \to f\log(f/g)$$

so that by Lebesgue dominated convergence theorem,

$$H(f|g) = \lim_{k\to\infty} H(f_k|g)\,.$$

As an increasing sequence, Pf_k converges to Pf almost everywhere. The elementary inequality $x\log(x) - x \geq x\log(y) - y$ for all $x \geq y \geq 0$ gives

$$(Pf_k)\log(Pf_k) - (Pf_k)\log(Pg) - (Pf_k) + (Pg) \geq 0\,.$$

Integration gives with Fatou's lemma (2.4.2)

$$H(Pf|Pg) - ||Pf|| + ||Pg|| \leq \liminf_{k\to\infty} H(Pf_k|Pg) - ||Pf_k|| + ||Pg||$$

and so $H(Pf|Pg) \leq \liminf_{k\to\infty} H(Pf_k|Pg)$. □

Corollary 2.16.3. For an **invertible** Markov operator $\mathcal{P}$, the relative entropy is constant: $H(\mathcal{P}f|\mathcal{P}g) = H(f|g)$.

Proof. Because $\mathcal{P}$ and $\mathcal{P}^{-1}$ are both Markov operators,

$$H(f|g) = H(\mathcal{P}\mathcal{P}^{-1}f|\mathcal{P}\mathcal{P}^{-1}g) \leq H(\mathcal{P}^{-1}f|\mathcal{P}^{-1}g) \leq H(f|g)\,.$$

□

Example. If a measure preserving transformation T is invertible, then the corresponding Koopman operator and Perron-Frobenius operators preserve relative entropy.

Corollary 2.16.4. The operator $T(\mu)(A) = \int_{\mathbb{R}^2} 1_A(\frac{x+y}{\sqrt{2}})\ d\mu(x)\ d\mu(y)$ does not decrease entropy.

Proof. Denote by X_μ a random variable having the law μ and with $\mu(X)$ the law of a random variable. For a fixed random variable Y, we define the Markov operator

$$P_Y(\mu) = \mu(\frac{X_\mu + Y}{\sqrt{2}}) .$$

Because the entropy is nondecreasing for each P_Y, we have this property also for the nonlinear map $T(\mu) = P_{X_\mu}(\mu)$. □

We have shown as a corollary of the central limit theorem that T has a unique fixed point attracting all of $\mathcal{P}_{0,1}$. The entropy is also strictly increasing at infinitely many points of the orbit $T^n(\mu)$ since it converges to the fixed point with maximal entropy. It follows that T is not invertible.

More generally: given a sequence X_n of IID random variables. For every n, the map P_n which maps the law of S_n^* into the law of S_{n+1}^* is a Markov operator which does not increase entropy. We can summarize: summing up IID random variables tends to increase the entropy of the distributions.
A fixed point of a Markov operator is called a **stationary state** or in more physical language a **thermodynamic equilibrium**. Important questions are: is there a thermodynamic equilibrium for a given Markov operator $\mathcal{P}$ and if yes, how many are there?

2.17 Characteristic functions

Distribution functions are in general not so easy to deal with, as for example, when summing up independent random variables. It is therefore convenient to deal with its Fourier transforms, the characteristic functions. It is an important topic by itself [60].

Definition. Given a random variable X, its **characteristic function** is a real-valued function on $\mathbb{R}$ defined as

$$\phi_X(u) = \mathrm{E}[e^{iuX}] .$$

If F_X is the distribution function of X and μ_X its law, the characteristic function of X is the Fourier-Stieltjes transform

$$\phi_X(t) = \int_{\mathbb{R}} e^{itx}\ dF_X(x) = \int_{\mathbb{R}} e^{itx}\ \mu_X(dx) .$$

Remark. If F_X is a continuous distribution function $dF_X(x) = f_X(x)\,dx$, then ϕ_X is the **Fourier transform** of the density function f_X:

$$\int_{\mathbb{R}} e^{itx}\, f_X(x)\, dx\, .$$

Remark. By definition, characteristic functions are Fourier transforms of probability measures: if μ is the law of X, then $\phi_X = \hat{\mu}$.

Example. For a random variable with density $f_X(x) = x^m/(m+1)$ on $\Omega = [0,1]$ the characteristic function is

$$\phi_X(t) = \int_0^1 e^{itx} x^m\, dx/(m+1) = \frac{m!(1-e^{it}e_m(-it))}{(-it)^{1+m}(m+1)}\, ,$$

where $e_n(x) = \sum_{k=0}^n x^k/(k!)$ is the n'th **partial exponential function.**

Theorem 2.17.1 (Lèvy formula). The characteristic function ϕ_X determines the distribution of X. If a, b are points of continuity of F, then

$$F_X(b) - F_X(a) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{e^{-ita} - e^{-itb}}{it}\phi_X(t)\, dt\, . \tag{2.9}$$

In general, one has

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{e^{-ita} - e^{-itb}}{it}\phi_X(t)\, dt = \mu[(a,b)] + \frac{1}{2}\mu[\{a\}] + \frac{1}{2}\mu[\{b\}]\, .$$

Proof. Because a distribution function F has only countably many points of discontinuities, it is enough to determine $F(b) - F(a)$ in terms of ϕ if a and b are continuity points of F. The verification of the **Lévy formula** is then a computation. For continuous distributions with density $F_X' = f_X$ is the inverse formula for the Fourier transform: $f_X(a) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ita}\phi_X(t)\, dt$ so that $F_X(a) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{e^{-ita}}{-it}\phi_X(t)\, dt$. This proves the inversion formula if a and b are points of continuity.

The general formula needs only to be verified when μ is a point measure at the boundary of the interval. By linearity, one can assume μ is located on a single point b with $p = \mathrm{P}[X = b] > 0$. The Fourier transform of the Dirac measure $p\delta_b$ is $\phi_X(t) = pe^{itb}$. The claim reduces to

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{e^{-ita} - e^{-itb}}{it}\, pe^{itb}\, dt = \frac{p}{2}$$

which is equivalent to the claim $\lim_{R\to\infty}\int_{-R}^{R} \frac{e^{itc}-1}{it}\, dt = \pi$ for $c > 0$. Because the imaginary part is zero for every R by symmetry, only

$$\lim_{R\to\infty}\int_{-R}^{R} \frac{\sin(tc)}{t}\, dt = \pi$$

remains. The verification of this integral is a prototype computation in residue calculus. □

Theorem 2.17.2 (Characterization of weak convergence). A sequence X_n of random variables converges weakly to X if and only if its characteristic functions converge point wise:

$$\phi_{X_n}(x) \to \phi_X \; .$$

Proof. Because the exponential function e^{itx} is continuous for each t, it follows from the definition that weak convergence implies the point wise convergence of the characteristic functions. From formula (2.9) follows that if the characteristic functions converge point wise, then convergence in distribution takes place. We have learned in lemma (2.13.2) that weak convergence is equivalent to convergence in distribution. □

Example. Here is a table of characteristic functions (CF) $\phi_X(t) = \mathrm{E}[e^{itX}]$ and moment generating functions (MGF) $M_X(t) = \mathrm{E}[e^{tX}]$ for some familiar random variables:

Distribution	Parameter	CF	MGF
Normal	$m \in \mathbb{R}$, $\sigma^2 > 0$	$e^{mit-\sigma^2t^2/2}$	$e^{mt+\sigma^2t^2/2}$
$N(0,1)$		$e^{-t^2/2}$	$e^{t^2/2}$
Uniform	$[-a,a]$	$\sin(at)/(at)$	$\sinh(at)/(at)$
Exponential	$\lambda > 0$	$\lambda/(\lambda - it)$	$\lambda/(\lambda - t)$
binomial	$n \geq 1$, $p \in [0,1]$	$(1-p+pe^{it})^n$	$(1-p+pe^t)^n$
Poisson	$\lambda > 0$, λ	$e^{\lambda(e^{it}-1)}$	$e^{\lambda(e^t-1)}$
Geometric	$p \in (0,1)$	$\frac{p}{(1-(1-p)e^{it}}$	$\frac{p}{(1-(1-p)e^t}$
first success	$p \in (0,1)$	$\frac{pe^{it}}{(1-(1-p)e^{it}}$	$\frac{pe^t}{(1-(1-p)e^t}$
Cauchy	$m \in \mathbb{R}$, $b > 0$	$e^{imt-\|t\|}$	$e^{mt-\|t\|}$

Definition. Let F and G be two probability distribution functions. Their **convolution** $F \star G$ is defined as

$$F \star G(x) = \int_{\mathbb{R}} F(x-y) \, dG(y) \; .$$

Lemma 2.17.3. If F and G are distribution functions, then $F \star G$ is again a distribution function.

Proof. We have to verify the three properties which characterize distribution functions among real-valued functions as in proposition (2.12.1).
a) Since F is nondecreasing, also $F \star G$ is nondecreasing.
b) Because $F(-\infty) = 0$ we have also $F \star G(-\infty) = 0$. Since $F(\infty) = 1$ and dG is a probability measure, also $F \star G(\infty) = 1$.
c) Given a sequence $h_n \to 0$. Define $F_n(x) = F(x + h_n)$. Because F is continuous from the right, $F_n(x)$ converges point wise to $F(x)$. The Lebesgue dominated convergence theorem implies that $F_n \star G(x) = F \star G(x + h_n)$ converges to $F \star G(x)$. □

Example. Given two discrete distributions

$$F(x) = \sum_{n \leq x} p_n, \; G(x) = \sum_{n \leq x} q_n \; .$$

Then $F \star G(x) = \sum_{n \leq x} (p \star q)_n$, where $p \star q$ is the convolution of the sequences p, q defined by $(p \star q)_n = \sum_{k=0}^{n} p_k q_{n-k}$. We see that the convolution of discrete distributions gives again a discrete distribution.

Example. Given two continuous distributions F, G with densities h and k. Then the distribution of $F \star G$ is given by the convolution

$$h \star k(x) = \int_{\mathbb{R}} h(x - y) k(y) \, dy$$

because

$$(F \star G)'(x) = \frac{d}{dx} \int_{\mathbb{R}} F(x - y) k(y) \, dy = \int h(x - y) k(y) \, dy \; .$$

Lemma 2.17.4. If F and G are distribution functions with characteristic functions ϕ and ψ, then $F \star G$ has the characteristic function $\phi \cdot \psi$.

Proof. While one can deduce this fact directly from Fourier theory, we prove it by hand: use an approximation of the integral by step functions:

$$\begin{aligned}
& \int_{\mathbb{R}} e^{iux} d(F \star G)(x) \\
= \; & \lim_{N,n \to \infty} \sum_{k=-N2^n+1}^{N2^n} e^{iuk2^{-n}} \int_{\mathbb{R}} [F(\frac{k}{2^n} - y) - F(\frac{k-1}{2^n} - y)] \, dG(y) \\
= \; & \lim_{N,n \to \infty} \sum_{k=-N2^n+1}^{N2^n} \int_{\mathbb{R}} e^{iu\frac{k}{2^n} - y} [F(\frac{k}{2^n} - y) - F(\frac{k-1}{2^n} - y)] \cdot e^{iuy} \, dG(y) \\
= \; & \int_{\mathbb{R}} [\lim_{N \to \infty} \int_{-N-y}^{N-y} e^{iux} \, dF(x)] e^{iuy} \, dG(y) = \int_{\mathbb{R}} \phi(u) e^{iuy} \, dG(y) \\
= \; & \phi(u) \psi(u) \; .
\end{aligned}$$

□

It follows that the set of distribution functions forms an **associative commutative group** with respect to the convolution multiplication. The reason is that the characteristic functions have this property with point wise multiplication.

Characteristic functions become especially useful, if one deals with independent random variables. Their characteristic functions multiply:

Proposition 2.17.5. Given a finite set of independent random variables $X_j, j = 1, \dots, n$ with characteristic functions ϕ_j. The characteristic function of $\sum_{j=1}^n X_j$ is $\phi = \prod_{j=1}^n \phi_j$.

Proof. Since X_j are independent, we get for any set of complex valued measurable functions g_j, for which $\mathrm{E}[g_j(X_j)]$ exists:

$$\mathrm{E}[\prod_{j=1}^{n} g_j(X_j)] = \prod_{j=1}^{n} \mathrm{E}[g_j(X_j)] \, .$$

Proof: This follows almost immediately from the definition of independence since one can check it first for functions $g_j = 1_{A_j}$, where A_j are $\sigma(X_j$ measurable functions for which $g_j(X_j)g_k(X_k) = 1_{A_j \cap A_k}$ and

$$\mathrm{E}[g_j(X_j)g_k(X_k)] = m(A_j)m(A_k) = \mathrm{E}[g_j(X_j)]\mathrm{E}[g_k(X_k)] \, ,$$

then for step functions by linearity and then for arbitrary measurable functions.

If we put $g_j(x) = \exp(ix)$, the proposition is proved. □

Example. If X_n are IID random variables which take the values 0 and 2 with probability 1/2 each, the random variable $X = \sum_{n=1}^{\infty} X_n/3^n$ is a random variable with the Cantor distribution. Because the characteristic function of X_n is $\phi_{X_n/3^n}(t) = \mathrm{E}[e^{itX_n/3^n}] = \frac{e^{i2/3^n}-1}{2}$, we see that the characteristic function of X is

$$\phi_X(t) = \prod_{i=1}^{\infty} \frac{e^{i2/3^n}-1}{2} \, .$$

The centered random variable $Y = X - 1/2$ can be written as $Y = \sum_{n=1}^{\infty} Y_n/3^n$, where Y_n takes values $-1, 1$ with probability 1/2. So

$$\phi_Y(t) = \prod_{n} \mathrm{E}[e^{itY_n/3^n}] = \prod_{n} \frac{e^{i/3^n} + e^{-i/3^n}}{2} = \prod_{n=1}^{\infty} \cos(\frac{t}{3^n}) \, .$$

This formula for the Fourier transform of a singular continuous measure μ has already been derived by Wiener. The Fourier theory of fractal measures has been developed much more since then.

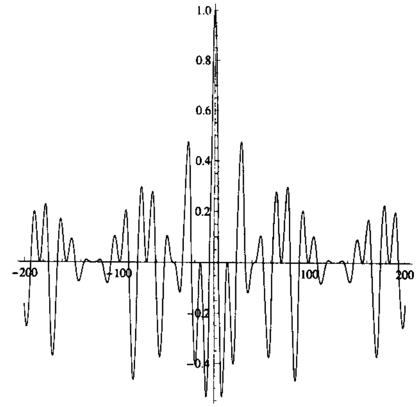


Figure. *The characteristic function $\phi_Y(t)$ of a random variable Y with a centered Cantor distribution supported on $[-1/2, 1/2]$ has an explicit formula $\phi_Y(t) = \prod_{n=1}^{\infty} \cos(\frac{t}{3^n})$ and already been derived by Wiener in the early 20'th century. The formula can also be used to compute moments of Y with the moment formula $\mathrm{E}[X^m] = (-i)^m \frac{d^m}{dt^m} \phi_X(t)|_{t=0}$.*

Corollary 2.17.6. The probability density of the sum of independent random variables $\sum_{j=1}^{n} X_j$ is $f_1 \star f_2 \star \cdots \star f_n$, if X_j has the density f_j.

Proof. This follows immediately from proposition (2.17.5) and the algebraic isomorphisms between the algebra of characteristic functions with convolution product and the algebra of distribution functions with point wise multiplication. □

Example. Let Y_k be IID random variables and let $X_k = \lambda^k Y_k$ with $0 < \lambda < 1$. The process $S_n = \sum_{k=1}^{n} X_k$ is called the **random walk with variable step size** or the **branching random walk** with exponentially decreasing steps. Let μ be the law of the random sum $X = \sum_{k=1}^{\infty} X_k$. If $\phi_Y(t)$ is the characteristic function of Y, then the characteristic function of X is

$$\phi_X(t) = \prod_{n=1}^{\infty} \phi_X(t\lambda^n) .$$

For example, if the random Y_n take values $-1, 1$ with probability $1/2$, where $\phi_Y(t) = \cos(t)$, then

$$\phi_X(t) = \prod_{n=1}^{\infty} \cos(t\lambda^n) .$$

The measure μ is then called a **Bernoulli convolution**. For example, for $\lambda = 1/3$, the measure is supported on the **Cantor set** as we have seen above. For more information on this stochastic process and the properties of the measure μ which in a subtle way depends on λ, see [41].

Exercice. Show that $X_n \to X$ in distribution if and only if the distribution functions satisfy $\phi_{X_n}(t) \to \phi_X(t)$ for all $t \in \mathbb{R}$.

Exercice. The **characteristic function** of a vector valued random variable $X = (X_1, \dots, X_k)$ is the real-valued function

$$\phi_X(t) = \mathrm{E}[e^{it \cdot X}]$$

on $\mathbb{R}^k$, where we wrote $t = (t_1, \dots, t_k)$. Two such random variables X, Y are **independent**, if the σ-algebras $X^{-1}(\mathcal{B})$ and $Y^{-1}(\mathcal{B})$ are independent, where $\mathcal{B}$ is the Borel σ-algebra on $\mathbb{R}^k$.
a) Show that if X and Y are independent then $\phi_{X+Y} = \phi_X \cdot \phi_Y$.
b) Given a real nonsingular $k \times k$ matrix A called the **covariance matrix** and a vector $m = (m_1, \dots, m_k)$ called the **mean** of X. We say, a vector valued random variable X has a **Gaussian distribution with covariance** A **and mean** m, if

$$\phi_X(t) = e^{im \cdot t - \frac{1}{2}(t \cdot At)} .$$

Show that the sum $X + Y$ of two Gaussian distributed random variables is again Gaussian distributed.
c) Find the probability density of a Gaussian distributed random variable X with covariance matrix A and mean m.

Exercice. The **Laplace transform** of a positive random variable $X \geq 0$ is defined as $l_X(t) = \mathrm{E}[e^{-tX}]$. The **moment generating function** is defined as $M(t) = \mathrm{E}[e^{tX}]$ provided that the expectation exists in a neighborhood of 0. The **generating function** of an integer-valued random variable is defined as $\zeta(X) = \mathrm{E}[u^X]$ for $u \in (0, 1)$. What does independence of two random variables X, Y mean in terms of (i) the Laplace transform, (ii) the moment generating function or (iii) the generating function?

Exercice. Let $(\Omega, \mathcal{A}, \mu)$ be a probability space and let $U, V \in \mathcal{X}$ be random variables (describing the energy density and the mass density of a thermodynamical system). We have seen that the **Helmholtz free energy**

$$\mathrm{E}_{\tilde{\mu}}[U] - kTH[\tilde{\mu}]$$

(k is a physical constant), T is the temperature, is taking its minimum for the exponential family. Find the measure minimizing the **free enthalpy** or **Gibbs potential**

$$\mathrm{E}_{\tilde{\mu}}[U] - kTH[\tilde{\mu}] - p\mathrm{E}_{\mu}[V] ,$$

where p is the pressure.

Exercice. Let $(\Omega, \mathcal{A}, \mu)$ be a probability space and $X_i \in \mathcal{L}$ random variables. Compute $\mathrm{E}_\mu[X_i]$ and the entropy of μ_λ in terms of the partition function $Z(\lambda)$.

Exercice. a) Given the discrete measure space $(\Omega = \{\epsilon_0 + n\delta\}, \nu)$, with $\epsilon_0 \in \mathbb{R}$ and $\delta > 0$ and where ν is the counting measure and let $X(k) = k$. Find the distribution f maximizing the entropy $H(f)$ among all measures $\tilde{\mu} = f\nu$ fixing $\mathrm{E}_{\tilde{\mu}}[X] = \epsilon$.
b) The physical interpretation is as follows: Ω is the discrete set of energies of a harmonic oscillator, ϵ_0 is the ground state energy, $\delta = \hbar\omega$ is the incremental energy, where ω is the frequency of the oscillation and $\hbar$ is Planck's constant. $X(k) = k$ is the Hamiltonian and $\mathrm{E}[X]$ is the energy. Put $\lambda = 1/kT$, where T is the temperature (in the answer of a), there appears a parameter λ, the Lagrange multiplier of the variational problem). Since can fix also the temperature T instead of the energy ϵ, the distribution in a) maximizing the entropy is determined by ω and T. Compute the spectrum $\epsilon(\omega, T)$ of the blackbody radiation defined by

$$\epsilon(\omega, T) = (E[X] - \epsilon_0)\frac{\omega^2}{\pi^2 c^3}$$

where c is the velocity of light. You have deduced then **Planck's blackbody radiation formula.**

2.18 The law of the iterated logarithm

We will give only a proof of the law of iterated logarithm in the special case, when the random variables X_n are independent and have all the standard normal distribution. The proof of the theorem for general IID random variables X_n can be found for example in [105]. The central limit theorem makes the general result plausible from the special case.

Definition. A random variable $X \in \mathcal{L}$ is called **symmetric** if its law μ_X satisfies:

$$\mu([-b, -a)) = \mu([a, b))$$

for all $a < b$. A symmetric random variable $X \in \mathcal{L}^1$ has zero mean. We again use the notation $S_n = \sum_{k=1}^n X_k$ in this section.

Lemma 2.18.1. Let X_n by symmetric and independent. For every $\epsilon > 0$

$$\mathrm{P}[\max_{1 \leq k \leq n} S_k > \epsilon] \leq 2\mathrm{P}[(S_n > \epsilon] \, .$$

Proof. This is a direct consequence of Lévy's theorem (2.11.6) because we can take $m = 0$ as the median of a symmetric distribution. □

Definition. Define for $n \geq 2$ the constants $\Lambda_n = \sqrt{2n \log\log n}$. It grows only slightly faster than $\sqrt{2n}$. For example, in order that the factor $\sqrt{\log\log n}$ is 3, we already have $n = \exp(\exp(9)) > 1.33 \cdot 10^{3519}$.

Theorem 2.18.2 (Law of iterated logarithm for $N(0,1)$). Let X_n be a sequence of IID $N(0,1)$-distributed random variables. Then

$$\limsup_{n\to\infty} \frac{S_n}{\Lambda_n} = 1, \quad \liminf_{n\to\infty} \frac{S_n}{\Lambda_n} = -1 .$$

Proof. We follow [47]. Because the second statement follows obviously from the first one by replacing X_n by $-X_n$, we have only to prove

$$\limsup_{n\to\infty} S_n/\Lambda_n = 1 .$$

(i) $\mathrm{P}[S_n > (1+\epsilon)\Lambda_n,$ infinitely often$] = 0$ for all $\epsilon > 0$.

Define $n_k = [(1+\epsilon)^k] \in \mathbb{N}$, where $[x]$ is the integer part of x and the events

$$A_k = \{S_n > (1+\epsilon)\Lambda_n, \text{for some } n \in (n_k, n_{k+1}]\ \}.$$

Clearly $\limsup_k A_k = \{S_n > (1+\epsilon)\Lambda_n,$ infinitely often$\}$. By the first Borel-Cantelli lemma (2.2.2), it is enough to show that $\sum_k \mathrm{P}[A_k] < \infty$. For each k, we get with the above lemma

$$\begin{aligned} \mathrm{P}[A_k] &\leq \mathrm{P}[\max_{n_k < n \leq n_{k+1}} S_n > (1+\epsilon)\Lambda_k] \\ &\leq \mathrm{P}[\max_{1 \leq n \leq n_{k+1}} S_n > (1+\epsilon)\Lambda_k] \\ &\leq 2\mathrm{P}[S_{n_{k+1}} > (1+\epsilon)\Lambda_k] . \end{aligned}$$

The right-hand side can be estimated further using that $S_{n_{k+1}}/\sqrt{n_{k+1}}$ is $N(0,1)$-distributed and that for a $N(0,1)$-distributed random variable $\mathrm{P}[X > t) \leq \text{const} \cdot e^{-t^2/2}$

$$\begin{aligned} 2\mathrm{P}[S_{n_{k+1}} > \Lambda_k] &= 2\mathrm{P}[\left(\frac{S_{n_{k+1}}}{\sqrt{n_{k+1}}} > (1+\epsilon)\frac{\sqrt{2n_k \log\log n_k}}{\sqrt{n_{k+1}}}\right)] \\ &\leq C\exp(-\frac{1}{2}(1+\epsilon)^2)\frac{2n_k \log\log n_k}{n_{k+1}}) \\ &\leq C_1 \exp(-(1+\epsilon)\log\log(n_k)) \\ &= C_1 \log(n_k)^{-(1+\epsilon)} \leq C_2 k^{-(1+\epsilon)} . \end{aligned}$$

Having shown that $\mathrm{P}[A_k] \leq \text{const} \cdot k^{-(1+\epsilon)}$ proves the claim $\sum_k \mathrm{P}[A_k] < \infty$.

(ii) $\mathrm{P}[S_n > (1-\epsilon)\Lambda_n$, infinitely often$] = 1$ for all $\epsilon > 0$.

It suffices to show, that for all $\epsilon > 0$, there exists a subsequence n_k

$$\mathrm{P}[S_{n_k} > (1-\epsilon)\Lambda_{n_k}, \text{ infinitely often}] = 1 \; .$$

Given $\epsilon > 0$. Choose $N > 1$ large enough and $c < 1$ near enough to 1 such that

$$c\sqrt{1 - 1/N} - 2/\sqrt{N} > 1 - \epsilon \; . \tag{2.10}$$

Define $n_k = N^k$ and $\Delta n_k = n_k - n_{k-1}$. The sets

$$A_k = \{S_{n_k} - S_{n_{k-1}} > c\sqrt{2\Delta n_k \log\log \Delta n_k}\}$$

are independent. In the following estimate, we use the fact that $\int_t^\infty e^{-x^2/2}\, dx \geq C \cdot e^{-t^2/2}$ for some constant C.

$$\begin{aligned}
\mathrm{P}[A_k] &= \mathrm{P}[\{S_{n_k} - S_{n_{k-1}} > c\sqrt{2\Delta n_k \log\log \Delta n_k}\}] \\
&= \mathrm{P}[\{\frac{S_{n_k} - S_{n_{k-1}}}{\sqrt{\Delta n_k}} > c\frac{\sqrt{2\Delta n_k \log\log \Delta n_k}}{\sqrt{\Delta n_k}}\}] \\
&\geq C \cdot \exp(-c^2 \log\log \Delta n_k) = C \cdot \exp(-c^2 \log(k \log N)) \\
&= C_1 \cdot \exp(-c^2 \log k) = C_1 k^{-c^2}
\end{aligned}$$

so that $\sum_k \mathrm{P}[A_k] = \infty$. We have therefore by Borel-Cantelli a set A of full measure so that for $\omega \in A$

$$S_{n_k} - S_{n_{k-1}} > c\sqrt{2\Delta n_k \log\log \Delta n_k}$$

for infinitely many k. From (i), we know that

$$S_{n_k} > -2\sqrt{2n_k \log\log n_k}$$

for sufficiently large k. Both inequalities hold therefore for infinitely many values of k. For such k,

$$\begin{aligned}
S_{n_k}(\omega) &> S_{n_{k-1}}(\omega) + c\sqrt{2\Delta n_k \log\log \Delta n_k} \\
&\geq -2\sqrt{2n_{k-1} \log\log n_{k-1}} + c\sqrt{2\Delta n_k \log\log \Delta n_k} \\
&\geq (-2/\sqrt{N} + c\sqrt{1 - 1/N})\sqrt{2n_k \log\log n_k} \\
&\geq (1-\epsilon)\sqrt{2n_k \log\log n_k} \; ,
\end{aligned}$$

where we have used assumption (2.10) in the last inequality. □

We know that $N(0,1)$ is the unique fixed point of the map T by the central limit theorem. The law of iterated logarithm is true for $T(X)$ implies that it is true for X. This shows that it would be enough to prove the theorem in the case when X has distribution in an arbitrary small neighborhood of

$N(0,1)$. We would need however sharper estimates.

We present a second proof of the central limit theorem in the IID case, to illustrate the use of characteristic functions.

Theorem 2.18.3 (Central limit theorem for IID random variables). Given $X_n \in \mathcal{L}^2$ which are IID with mean 0 and finite variance σ^2. Then $S_n/(\sigma\sqrt{n}) \to N(0,1)$ in distribution.

Proof. The characteristic function of $N(0,1)$ is $\phi(t) = e^{-t^2/2}$. We have to show that for all $t \in \mathbb{R}$

$$\mathrm{E}[e^{it\frac{S_n}{\sigma\sqrt{n}}}] \to e^{-t^2/2} .$$

Denote by ϕ_{X_n} the characteristic function of X_n. Since by assumption $\mathrm{E}[X_n] = 0$ and $\mathrm{E}[X_n^2] = \sigma^2$, we have

$$\phi_{X_n}(t) = 1 - \frac{\sigma^2}{2}t^2 + o(t^2) .$$

Therefore

$$\begin{aligned} \mathrm{E}[e^{it\frac{S_n}{\sigma\sqrt{n}}}] &= \phi_{X_n}(\frac{t}{\sigma\sqrt{n}})^n \\ &= (1 - \frac{1}{2}\frac{t^2}{n} + o(\frac{1}{n}))^n \\ &= e^{-t^2/2} + o(1) . \end{aligned}$$

□

This method can be adapted to other situations as the following example shows.

Proposition 2.18.4. Given a sequence of independent events $A_n \subset \Omega$ with $\mathrm{P}[A_n] = 1/n$. Define the random variables $X_n = 1_{A_n}$ and $S_n = \sum_{k=1}^n X_k$. Then

$$T_n = \frac{S_n - \log(n)}{\sqrt{\log(n)}}$$

converges to $N(0,1)$ in distribution.

Proof.

$$\mathrm{E}[S_n] = \sum_{k=1}^n \frac{1}{k} = \log(n) + \gamma + o(1) ,$$

where $\gamma = \lim_{n\to\infty} \sum_{k=1}^{n} \frac{1}{k} - \log(n)$ is the **Euler constant.**

$$\text{Var}[S_n] = \sum_{k=1}^{n} \frac{1}{k}(1 - \frac{1}{k}) = \log(n) + \gamma - \frac{\pi^2}{6} + o(1) .$$

satisfy $\text{E}[T_n] \to 0$ and $\text{Var}[T_n] \to 1$. Compute $\phi_{X_n} = 1 - \frac{1}{n} + \frac{e^{it}}{n}$ so that $\phi_{S_n}(t) = \prod_{k=1}^{n}(1 - \frac{1}{k} + \frac{e^{it}}{k})$ and $\phi_{T_n}(t) = \phi_{S_n}(s(t))e^{-is\log(n)}$, where $s = t/\sqrt{\log(n)}$. For $n \to \infty$, we compute

$$\begin{aligned}
\log\phi_{T_n}(t) &= -it\sqrt{\log(n)} + \sum_{k=1}^{n} \log(1 + \frac{1}{k}(e^{is} - 1)) \\
&= -it\sqrt{\log(n)} + \sum_{k=1}^{n} \log(1 + \frac{1}{k}(is - \frac{1}{2}s^2 + o(s^2))) \\
&= -it\sqrt{\log(n)} + \sum_{k=1}^{n} \frac{1}{k}(is + \frac{1}{2}s^2 + o(s^2)) + O(\sum_{k=1}^{n} \frac{s^2}{k^2}) \\
&= -it\sqrt{\log(n)} + (is - \frac{1}{2}s^2 + o(s^2))(\log(n) + O(1)) + t^2O(1) \\
&= \frac{-1}{2}t^2 + o(1) \to -\frac{1}{2}t^2 .
\end{aligned}$$

We see that T_n converges in law to the standard normal distribution. □

Chapter 3

Discrete Stochastic Processes

3.1 Conditional Expectation

Definition. Given a probability space $(\Omega, \mathcal{A}, \mathrm{P})$. A second measure P' on $(\Omega, \mathcal{A})$ is called **absolutely continuous** with respect to P, if $\mathrm{P}[A] = 0$ implies $\mathrm{P}'[A] = 0$ for all $A \in \mathcal{A}$. One writes $\mathrm{P}' \ll \mathrm{P}$.

Example. If $\mathrm{P}[a, b] = b - a$ is the uniform distribution on $\Omega = [0, 1]$ and $\mathcal{A}$ is the Borel σ-algebra, and $Y \in \mathcal{L}^1$ satisfies $Y(x) \geq 0$ for all $x \in \Omega$, then $\mathrm{P}'[a, b] = \int_a^b Y(x)\, dx$ is absolutely continuous with respect to P.

Example. Assume P is again the Lebesgue measure on $[0, 1]$ as in the last example. If $Y(x) = 1_B(x)$, then $\mathrm{P}'[A] = \mathrm{P}[A \cap B]$ for all $A \in \mathcal{A}$. If $\mathrm{P}[B] < 1$, then P is not absolutely continuous with respect to P'. We have $\mathrm{P}'[B^c] = 0$ but $\mathrm{P}[B^c] = 1 - \mathrm{P}[B] > 0$.

Example. If $\mathrm{P}'[A] = \left\{ \begin{array}{cc} 1 & 1/2 \in A \\ 0 & 1/2 \notin A \end{array} \right.$, then P' is not absolutely continuous with respect to P. For $B = \{1/2\}$, we have $\mathrm{P}[B] = 0$ but $\mathrm{P}'[B] = 1 \neq 0$.

The next theorem is a reformulation of a classical theorem of Radon-Nykodym of 1913 and 1930.

Theorem 3.1.1 (Radon-Nykodym equivalent). Given a measure P' which is absolutely continuous with respect to P, then there exists a unique $Y \in \mathcal{L}^1(\mathrm{P})$ with $\mathrm{P}' = Y\mathrm{P}$. The function Y is called the **Radon-Nykodym derivative** of P' with respect to P. It is unique in L^1.

Proof. We can assume without loss of generality that P' is a positive measure (do else the Hahn decomposition $\mathrm{P} = \mathrm{P}^+ - \mathrm{P}^-$), where P^+ and P^-

are positive measures).

(i) Construction: We recall the notation $\mathrm{E}[Y;A] = \mathrm{E}[1_A Y] = \int_A Y \, dP$. The set $\Gamma = \{Y \geq 0 \mid \mathrm{E}[Y;A] \leq \mathrm{P}'[A], \forall A \in \mathcal{A}\}$ is closed under formation of suprema

$$\begin{aligned} \mathrm{E}[Y_1 \vee Y_2; A] &= \mathrm{E}[Y_1; A \cap \{Y_1 > Y_2\}] + \mathrm{E}[Y_2; A \cap \{Y_2 \geq Y_1\}] \\ &\leq \mathrm{P}'[A \cap \{Y_1 > Y_2\}] + \mathrm{P}'[A \cap \{Y_2 \geq Y_1\}] = \mathrm{P}'[A] \end{aligned}$$

and contains a function Y different from 0 since else, P' would be singular with respect to P according to the definition (2.15) of absolute continuity. We claim that the supremum Y of all functions Γ satisfies $Y\mathrm{P} = \mathrm{P}'$: an application of Beppo-Lévi's theorem (2.4.1) shows that the supremum of Γ is in Γ. The measure $\mathrm{P}'' = \mathrm{P}' - Y\mathrm{P}$ is the zero measure since we could do the same argument with a new set Γ for the absolutely continuous part of P''.

(ii) Uniqueness: assume there exist two derivatives Y, Y'. One has then $\mathrm{E}[Y - Y'; \{Y \geq Y'\}] = 0$ and so $Y \geq Y'$ almost everywhere. A similar argument gives $Y' \leq Y$ almost everywhere, so that $Y = Y'$ almost everywhere. In other words, $Y = Y'$ in L^1. □

Theorem 3.1.2 (Existence of conditional expectation, Kolmogorov 1933). Given $X \in \mathcal{L}^1(\mathcal{A})$ and a sub σ-algebra $\mathcal{B} \subset \mathcal{A}$. There exists a random variable $Y \in \mathcal{L}^1(\mathcal{B})$ with $\int_A Y \, dP = \int_A X \, dP$ for all $A \in \mathcal{B}$.

Proof. Define the measures $\tilde{\mathrm{P}}[A] = \mathrm{P}[A]$ and $\mathrm{P}'[A] = \int_A X \, dP = \mathrm{E}[X;A]$ on the probability space $(\Omega, \mathcal{B})$. Given a set $B \in \mathcal{B}$ with $\tilde{P}[B] = 0$, then $\mathrm{P}'[B] = 0$ so that P' is absolutely continuous with respect to $\tilde{P}$. Radon-Nykodym's theorem (3.1.1) provides us with a random variable $Y \in \mathcal{L}^1(\mathcal{B})$ with $\mathrm{P}'[A] = \int_A X \, dP = \int_A Y \, dP$. □

Definition. The random variable Y in this theorem is denoted with $\mathrm{E}[X|\mathcal{B}]$ and called the **conditional expectation of X with respect to $\mathcal{B}$**. The random variable $Y \in \mathcal{L}^1(\mathcal{B})$ is unique in $L^1(\mathcal{B})$. If Z is a random variable, then $\mathrm{E}[X|Z]$ is defined as $\mathrm{E}[X|\sigma(Z)]$. If $\{Z\}_{\mathcal{I}}$ is a family of random variables, then $\mathrm{E}[X|\{Z\}_{\mathcal{I}}]$ is defined as $\mathrm{E}[X|\sigma(\{Z\}_{\mathcal{I}})]$.

Example. If $\mathcal{B}$ is the trivial σ-algebra $\mathcal{B} = \{\emptyset, \Omega\}$, then $\mathrm{E}[X|\mathcal{B}] = X$.

Example. If $\mathcal{B} = \mathcal{A}$, then $\mathrm{E}[X|\mathcal{B}] = \mathrm{E}[X]$.

Example. If $\mathcal{B} = \{\emptyset, Y, Y^c, \Omega\}$ then

$$\mathrm{E}[X|\mathcal{B}](\omega) = \begin{cases} \frac{1}{m(Y)} \int_Y X \, dP & \text{for } \omega \in Y , \\ \frac{\int_{Y^c} X \, dP}{m(Y^c)} & \text{for } \omega \in Y^c . \end{cases}$$

Example. Let $(\Omega, \mathcal{A}, \mathcal{P}) = ([0,1] \times [0,1], \mathcal{A}, dxdy)$, where $\mathcal{A}$ is the Borel σ-algebra defined by the Euclidean distance metric on the square Ω. Let $\mathcal{B}$ be the σ-algebra of sets $A \times [0,1]$, where A is in the Borel σ-algebra of the interval $[0,1]$. If $X(x,y)$ is a random variable on Ω, then $Y = \mathrm{E}[X|\mathcal{B}]$ is the random variable

$$Y(x,y) = \int_0^1 X(x,y)\, dy \; .$$

This **conditional integral** only depends on x.

Remark. This notion of conditional entropy will be important later. Here is a possible interpretation of conditional expectation: for an experiment, the possible outcomes are modeled by a probability space $(\Omega, \mathcal{A})$ which is our "laboratory". Assume that the only information about the experiment are the events in a subalgebra $\mathcal{B}$ of $\mathcal{A}$. It models the "knowledge" obtained from some measurements we can do in the laboratory and $\mathcal{B}$ is generated by a set of random variables $\{Z_i\}_{i \in \mathcal{I}}$ obtained from some measuring devices. With respect to these measurements, our best knowledge of the random variable X is the conditional expectation $\mathrm{E}[X|\mathcal{B}]$. It is a random variable which is a function of the measurements Z_i. For a specific "experiment ω, the conditional expectation $\mathrm{E}[X|\mathcal{B}](\omega)$ is the expected value of $X(\omega)$, conditioned to the σ-algebra $\mathcal{B}$ which contains the events singled out by data from X_i.

Proposition 3.1.3. The conditional expectation $X \mapsto \mathrm{E}[X|\mathcal{B}]$ is the projection from $\mathcal{L}^2(\mathcal{A})$ onto $\mathcal{L}^2(\mathcal{B})$.

Proof. The space $\mathcal{L}^2(\mathcal{B})$ of square integrable $\mathcal{B}$-measurable functions is a linear subspace of $\mathcal{L}^2(\mathcal{A})$. When identifying functions which agree almost everywhere, then $L^2(\mathcal{B})$ is a Hilbert space which is a linear subspace of the Hilbert space $L^2(\mathcal{A})$. For any $X \in \mathcal{L}^2(\mathcal{A})$, there exists a unique projection $p(X) \in \mathcal{L}^2(\mathcal{B})$. The orthogonal complement $\mathcal{L}^2(\mathcal{B})^\perp$ is defined as

$$\mathcal{L}^2(\mathcal{B})^\perp = \{Z \in \mathcal{L}^2(\mathcal{A}) \mid (Z,Y) := \mathrm{E}[Z \cdot Y] = 0 \text{ for all } Y \in \mathcal{L}^2(\mathcal{B})\ \} \; .$$

By the definition of the conditional expectation, we have for $A \in \mathcal{B}$

$$(X - \mathrm{E}[X|\mathcal{B}], 1_A) = \mathrm{E}[X - \mathrm{E}[X|\mathcal{B}]; A] = 0 \; .$$

Therefore $X - \mathrm{E}[X|\mathcal{B}] \in \mathcal{L}^2(\mathcal{B})^\perp$. Because the map $q(X) = \mathrm{E}[X|\mathcal{B}]$ satisfies $q^2 = q$, it is linear and has the property that $(1-q)(X)$ is perpendicular to $\mathcal{L}^2(\mathcal{B})$, the map q is a projection which must agree with p. □

Example. Let $\Omega = \{1,2,3,4\}$ and $\mathcal{A}$ the σ-algebra of all subsets of Ω. Let $\mathcal{B} = \{\emptyset, \{1,3\}, \{2,4\}, \Omega\}$. What is the conditional expectation $Y = \mathrm{E}[X|\mathcal{B}]$

of the random variable $X(k) = k^2$? The Hilbert space $\mathcal{L}^2(\mathcal{A})$ is the four-dimensional space $\mathbb{R}^4$ because a random variable X is now just a vector $X = (X(1), X(2), X(3), X(4)) = (1, 4, 9, 16)$. The Hilbert space $\mathcal{L}^2(\mathcal{B})$ is the set of all vectors $v = (v_1, v_2, v_3, v_4)$ for which $v_1 = v_2$ and $v_3 = v_4$ because functions which would not be constant in (v_1, v_2) would generate a finer algebra. It is the two-dimensional subspace of all vectors $\{v = (a, a, b, b) \mid a, b \in R\}$. The conditional expectation projects onto that plane. The first two components $(X(1), X(2))$ project to $(\frac{X(1)+X(2)}{\sqrt{2}}, \frac{X(1)+X(2)}{\sqrt{2}})$, the second two components project to $(\frac{X(3)+X(4)}{\sqrt{2}}, \frac{X(3)+X(4)}{\sqrt{2}})$. Therefore,

$$\mathrm{E}[X|\mathcal{B}] = (\frac{X(1)+X(2)}{\sqrt{2}}, \frac{X(1)+X(2)}{\sqrt{2}}, \frac{X(3)+X(4)}{\sqrt{2}}, \frac{X(3)+X(4)}{\sqrt{2}}) .$$

Remark. This proposition 3.1.3 means that Y is the least-squares best $\mathcal{B}$-measurable square integrable predictor. This makes conditional expectation important for controlling processes. If $\mathcal{B}$ is the σ-algebra describing the knowledge about a process (like for example the data which a pilot knows about an plane) and X is the random variable (which could be the actual data of the flying plane), we want to know, then $\mathrm{E}[X|\mathcal{B}]$ is the best guess about this random variable, we can make with our knowledge.

Exercice. Given two independent random variables $X, Y \in \mathcal{L}^2$ such that X has the Poisson distribution P_λ and Y has the Poisson distribution P_μ. The random variable $Z = X + Y$ has Poisson distribution $\mathrm{P}_{\lambda+\mu}$ as can be seen with the help of characteristic functions. Let $\mathcal{B}$ be the σ-algebra generated by Z. Show that

$$\mathrm{E}[X|\mathcal{B}] = \frac{\lambda}{\lambda+\mu} Z .$$

Hint: It is enough to show

$$\mathrm{E}[X; \{Z = k\}] = \frac{\lambda}{\lambda+\mu} \mathrm{P}[Z = k] .$$

Even if random variables are only in $\mathcal{L}^1$, the next list of properties of conditional expectation can be remembered better with proposition 3.1.3 in mind which identifies conditional expectation as a projection, if they are in $\mathcal{L}^2$.

Theorem 3.1.4 (Properties of conditional expectation). For given random variables $X, X_n, Y \in \mathcal{L}$, the following properties hold:
(1) Linearity: The map $X \mapsto \mathrm{E}[X|\mathcal{B}]$ is linear.
(2) Positivity: $X \geq 0 \Rightarrow \mathrm{E}[X|\mathcal{B}] \geq 0$.
(3) Tower property: $\mathcal{C} \subset \mathcal{B} \subset \mathcal{A} \Rightarrow \mathrm{E}[E[X|\mathcal{B}]|\mathcal{C}] = \mathrm{E}[X|\mathcal{C}]$.
(4) Conditional Fatou: $|X_n| \leq X$, $\mathrm{E}[\liminf_{n\to\infty} X_n|\mathcal{B}] \leq \liminf_{n\to\infty} \mathrm{E}[X_n|\mathcal{B}]$.
(5) Conditional dominated convergence: $|X_n| \leq X, X_n \to X$ a.e. $\Rightarrow \mathrm{E}[X_n|\mathcal{B}] \to \mathrm{E}[X|\mathcal{B}]$ a.e.
(6) Conditional Jensen: if h is convex, then $\mathrm{E}[h(X)|\mathcal{B}] \geq h(E[X|\mathcal{B}])$. Especially $||\mathrm{E}[X|\mathcal{B}]||_p \leq ||X||_p$.
(7) Extracting knowledge: For $Z \in \mathcal{L}^\infty(\mathcal{B})$, one has $\mathrm{E}[ZX|\mathcal{B}] = Z\mathrm{E}[X|\mathcal{B}]$.
(8) Independence: if X is independent of $\mathcal{C}$, then $\mathrm{E}[X|\mathcal{C}] = \mathrm{E}[X]$.

Proof. (1) For positivity, note that if $Y = \mathrm{E}[X|\mathcal{B}]$ would be negative on a set of positive measure, then $A = Y^{-1}([-1/n, 0]) \in \mathcal{B}$ would have positive probability for some n. This would lead to the contradiction $0 \leq \mathrm{E}[1_A X] = \mathrm{E}[1_A Y] \leq -n^{-1} m(A) < 0$.

(2) Use that $P'' \ll P' \ll P$ implies $P'' = Y'P' = Y'YP$ and $P'' \ll P$ gives $P'' = ZP$ so that $Z = Y'Y$ almost everywhere.

(3) This is especially useful when applied to the algebra $\mathcal{C}_Y = \{\emptyset, Y, Y^c, \Omega\}$. Because $X \leq Y$ almost everywhere if and only if $\mathrm{E}[X|\mathcal{C}_Y] \leq \mathrm{E}[Y|\mathcal{C}_Y]$ for all $Y \in \mathcal{B}$.
(4)-(5) The conditional versions of the Fatou lemma or the dominated convergence theorem are true, if they are true conditioned with $\mathcal{C}_Y$ for each $Y \in \mathcal{B}$. The tower property reduces these statements to versions with $\mathcal{B} = \mathcal{C}_Y$ which are then on each of the sets Y, Y^c the usual theorems.

(6) Chose a sequence $(a_n, b_n) \in \mathbb{R}^2$ such that $h(x) = \sup_n a_n x + b_n$ for all $x \in \mathbb{R}$. We get from $h(X) \geq a_n X + b_n$ that almost surely $\mathrm{E}[h(X)|\mathcal{G}] \geq a_n \mathrm{E}[X|\mathcal{G}] + b_n$. These inequalities hold therefore simultaneously for all n and we obtain almost surely

$$\mathrm{E}[h(X)|\mathcal{G}] \geq \sup_n (a_n \mathrm{E}[X|\mathcal{G}] + b_n) = h(E[X|\mathcal{G}]) .$$

The corollary is obtained with $h(x) = |x|^p$.

(7) It is enough to condition it to each algebra $\mathcal{C}_Y$ for $Y \in \mathcal{B}$. The tower property reduces these statements to linearity.

(8) By linearity, we can assume $X \geq 0$. For $B \in \mathcal{B}$ and $C \in \mathcal{C}$, the random variables $X 1_B$ and 1_C are independent so that

$$\mathrm{E}[X 1_{B \cap C}] = \mathrm{E}[X 1_B 1_C] = \mathrm{E}[X 1_B] \mathrm{P}[C] .$$

The random variable $Y = \mathrm{E}[X|\mathcal{B}]$ is $\mathcal{B}$ measurable and because $Y1_B$ is independent of $\mathcal{C}$ we get

$$\mathrm{E}[(Y1_B)1_C] = \mathrm{E}[Y1_B]\mathrm{P}[C]$$

so that $\mathrm{E}[1_{B\cap C}X] = \mathrm{E}[1_{B\cap C}Y]$. The measures on $\sigma(\mathcal{B},\mathcal{C})$

$$\mu : A \mapsto \mathrm{E}[1_A X], \nu : A \mapsto \mathrm{E}[1_A Y]$$

agree therefore on the π-system of the form $B \cap C$ with $B \in \mathcal{B}$ and $C \in \mathcal{C}$ and consequently everywhere on $\sigma(\mathcal{B},\mathcal{C})$. □

Remark. From the conditional Jensen property in theorem (3.1.4), it follows that the operation of conditional expectation is a positive and continuous operation on $\mathcal{L}^p$ for any $p \geq 1$.

Remark. The properties of Conditional Fatou, Lebesgue and Jensen are statements about functions in $\mathcal{L}^1(\mathcal{B})$ and not about numbers as the usual theorems of Fatou, Lebesgue or Jensen.

Remark. Is there for almost all $\omega \in \Omega$ a probability measure P_ω such that

$$\mathrm{E}[X|\mathcal{B}](\omega) = \int_\Omega X \, dP_\omega \; ?$$

If such a map from Ω to $M_1(\Omega)$ exists and if it is $\mathcal{B}$-measurable, it is called a **regular conditional probability** given $\mathcal{B}$. In general such a map $\omega \mapsto \mathrm{P}_\omega$ does not exist. However, it is known that for a probability space $(\Omega, \mathcal{A}, \mathrm{P})$ for which Ω is a complete separable metric space with Borel σ-algebra $\mathcal{A}$, there exists a regular probability space for any sub σ-algebra $\mathcal{B}$ of $\mathcal{A}$.

Exercice. This exercise deals with conditional expectation.
a) What is $\mathrm{E}[Y|Y]$?
b) Show that if $\mathrm{E}[X|\mathcal{A}] = 0$ and $\mathrm{E}[X|\mathcal{B}] = 0$, then $\mathrm{E}[X|\sigma(\mathcal{A},\mathcal{B})] = 0$.
c) Given $X, Y \in \mathcal{L}^1$ satisfying $\mathrm{E}[X|Y] = Y$ and $\mathrm{E}[Y|X] = X$. Verify that $X = Y$ almost everywhere.

We add a notation which is commonly used.

Definition. The **conditional probability space** $(\Omega, \mathcal{A}, \mathrm{P}[\cdot|\mathcal{B}])$ is defined by

$$\mathrm{P}[B \mid \mathcal{B}] = \mathrm{E}[1_B|\mathcal{B}] \; .$$

For $X \in \mathcal{L}^p$, one has the **conditional moment** $\mathrm{E}[X^p|\mathcal{B}] = \mathrm{E}[X^p|\mathcal{B}]$ if $\mathcal{B}$ be a σ-subalgebra of $\mathcal{A}$. They are $\mathcal{B}$-measurable random variables and generalize the usual moments. Of special interest is the conditional variance:

Definition. For $X \in \mathcal{L}^2$, the **conditional variance** $\mathrm{Var}[X|\mathcal{B}]$ is the random variable $\mathrm{E}[X^2|\mathcal{B}] - \mathrm{E}[X|\mathcal{B}]^2$. Especially, if $\mathcal{B}$ is generated by a random variable Y, one writes $\mathrm{Var}[X|Y] = \mathrm{E}[X^2|Y] - \mathrm{E}[X|Y]^2$.

Remark. Because conditional expectation is a projection, all properties known for the usual variance hold the more general notion of conditional variance. For example, if X, Z are independent random variables in $\mathcal{L}^2$, then $\mathrm{Var}[X + Z|Y] = \mathrm{Var}[X|Y] + \mathrm{Var}[Z|Y]$. One also has the identity $\mathrm{Var}[X|Y] = \mathrm{E}[(X - \mathrm{E}[X|Y])^2|Y]$.

Lemma 3.1.5. (Law of total variance) For $X \in \mathcal{L}^2$ and an arbitrary random variable Y, one has

$$\mathrm{Var}[X] = \mathrm{E}[\mathrm{Var}[X|Y]] + \mathrm{Var}[\mathrm{E}[X|Y]] \, .$$

Proof. By the definition of the conditional variance as well as the properties of conditional expectation:

$$\begin{aligned}
\mathrm{Var}[X] &= \mathrm{E}[X^2] - \mathrm{E}[X]^2 \\
&= \mathrm{E}[\mathrm{E}[X^2|Y]] - \mathrm{E}[\mathrm{E}[X|Y]]^2 \\
&= \mathrm{E}[\mathrm{Var}[X|Y]] + \mathrm{E}[\mathrm{E}[X|Y]^2] - \mathrm{E}[\mathrm{E}[X|Y]]^2 \\
&= \mathrm{E}[\mathrm{Var}[X|Y]] + \mathrm{Var}[\mathrm{E}[X|Y]] \, .
\end{aligned}$$

□

Here is an application which illustrates how one can use of the conditional variance in applications: the Cantor distribution is the singular continuous distribution with the law μ has its support on the standard Cantor set.

Corollary 3.1.6. (Variance of the Cantor distribution) The standard Cantor distribution for the Cantor set on $[0,1]$ has the expectation $1/2$ and the variance $1/8$.

Proof. Let X be a random variable with the Cantor distribution. By symmetry, $\mathrm{E}[X] = \int_0^1 x \, d\mu(x) = 1/2$. Define the σ-algebra

$$\{\emptyset, [0,1/3), [1/3,1], [0,1]\,\}$$

on $\Omega = [0,1]$. It is generated by the random variable $Y = 1_{[0,1/3)}$. Define $Z = \mathrm{E}[X|Y]$. It is a random variable which is constant $1/6$ on $[0,1/3)$ and equal to $5/6$ on $[1/3,1]$. It has the expectation $\mathrm{E}[Z] = (1/6)\mathrm{P}[Y = 1] + (5/6)\mathrm{P}[Y = 0] = 1/12 + 5/12 = 1/2$ and the variance

$$\mathrm{Var}[Z] = \mathrm{E}[Z^2] - \mathrm{E}[Z]^2 = \frac{1}{36}\mathrm{P}[Y = 1] + \frac{25}{36}\mathrm{P}[Y = 0] - 1/4 = 1/9 \, .$$

Define the random variable $W = \mathrm{Var}[X|Y] = \mathrm{E}[X^2|Y] - \mathrm{E}[X|Y]^2 = \mathrm{E}[X^2|Y] - Z$. It is equal to $\int_0^{1/3} (x - 1/6)^2 \, dx$ on $[0, 1/3]$ and equal to $\int_{2/3}^1 (x - 5/6)^2 \, dx$ on $[2/3, 3/3]$. By the self-similarity of the Cantor set, we see that $W = \mathrm{Var}[X|Y]$ is actually constant and equal to $\mathrm{Var}[X]/9$. The identity $\mathrm{E}[\mathrm{Var}[X|Y]] = \mathrm{Var}[X]/9$ implies

$$\mathrm{Var}[X] = \mathrm{E}[\mathrm{Var}[X|Y]] + \mathrm{Var}[\mathrm{E}[X|Y]] = \mathrm{E}[W] + \mathrm{Var}[Z] = \frac{\mathrm{Var}[X]}{9} + \frac{1}{9} \, .$$

Solving for $\mathrm{Var}[X]$ gives $\mathrm{Var}[X] = 1/8$. □

Exercice. Given a probability space $(\Omega, \mathcal{A}, \mathrm{P})$ and a σ-algebra $\mathcal{B} \subset \mathcal{A}$.
a) Show that the map $P : X \in \mathcal{L}^1 \mapsto \mathrm{E}[X|\mathcal{B}]$ is a Markov operator from $\mathcal{L}^1(\mathcal{A}, \mathrm{P})$ to $\mathcal{L}^1(\mathcal{B}, Q)$, where Q is the conditional probability measure on $(\Omega, \mathcal{B})$ defined by $Q[A] = \mathrm{P}[A]$ for $A \in \mathcal{B}$.
b) The map T can also be viewed as a map on the new probability space $(\Omega, \mathcal{B}, Q)$, where Q is the conditional probability. Denote this new map by S. Show that S is again measure preserving and invertible.

Exercice. a) Given a measure preserving invertible map $T : \Omega \to \Omega$ we call $(\Omega, T, \mathcal{A}, \mathrm{P})$ a **dynamical system**. A complex number λ is called an **eigenvalue of** T, if there exists $X \in \mathcal{L}^2$ such that $X(T) = \lambda X$. The map T is said to have **pure point spectrum**, if there exists a countable set of eigenvalues λ_i such that their eigenfuctions X_i span $\mathcal{L}^2$. Show that if T has pure point spectrum, then also S has pure point spectrum.
b) A measure preserving dynamical system $(\Delta, S, \mathcal{B}, \nu)$ is called a **factor** of a measure preserving dynamical system $(\Omega, T, \mathcal{A}, \mu)$ if there exists a measure preserving map $U : \Omega \to \Delta$ such that $S \circ U(x) = U \circ T(x)$ for all $x \in \Omega$. Examples of factors are the system itself or the trivial system $(\Omega, S(x) = x, \mu)$. If S is a factor of T and T is a factor of S, then the two systems are called **isomorphic**. Verify that every factor of a dynamical system $(\Omega, T, \mathcal{A}, \mu)$ can be realized as $(\Omega, T, \mathcal{B}, \mu)$ where $\mathcal{B}$ is a σ-subalgebra of $\mathcal{A}$.
c) It is known that if a measure preserving transformation T on a probability space has pure point spectrum, then the system is isomorphic to a translation on the compact Abelian group $\hat{G}$ which is the dual group of the discrete group G formed by the spectrum $\sigma(T) \subset \mathbb{T}$. Describe the possible **factors** of T and their spectra.

Exercice. Let $\Omega = \mathbb{T}^1$ be the one-dimensional circle. Let $\mathcal{A}$ be the Borel σ-algebra on $\mathbb{T}^1 = \mathbb{R}/(2\pi\mathbb{Z})$ and $P = dx$ the Lebesgue measure. Given $k \in \mathbb{N}$, denote by $\mathcal{B}_k$ the σ-algebra consisting of all $A \in \mathcal{A}$ such that $A + \frac{n2\pi}{k} = A \pmod{2\pi}$ for all $1 \leq n \leq k$. What is the conditional expectation $\mathrm{E}[X|\mathcal{B}_k]$ for a random variable $X \in \mathcal{L}^1$?

3.2 Martingales

It is typical in probability theory is that one considers several σ-algebras on a probability space $(\Omega, \mathcal{A}, \mathrm{P})$. These algebras are often defined by a set of random variables, especially in the case of stochastic processes. Martingales are discrete stochastic processes which generalize the process of summing up IID random variables. It is a powerful tool with many applications.

Definition. A sequence $\{\mathcal{A}_n\}_{n\in\mathbb{N}}$ of sub σ-algebras of $\mathcal{A}$ is called a **filtration**, if $\mathcal{A}_0 \subset \mathcal{A}_1 \subset \cdots \subset \mathcal{A}$. Given a filtration $\{\mathcal{A}_n\}_{n\in\mathbb{N}}$, one calls $(\Omega, \mathcal{A}, \{\mathcal{A}_n\}_{n\in\mathbb{N}}, \mathrm{P})$ a **filtered space**.

Example. If $\Omega = \{0,1\}^{\mathbb{N}}$ is the space of all $0-1$ sequences with the Borel σ-algebra generated by the product topology and $\mathcal{A}_n$ is the finite set of cylinder sets $A = \{x_1 = a_1, \dots, x_n = a_n\}$ with $a_i \in \{0,1\}$, which contains 2^n elements, then $\{\mathcal{A}_n\}_{n\in\mathbb{N}}$ is a filtered space.

Definition. A sequence $X = \{X_n\}_{n\in\mathbb{N}}$ of random variables is called a **discrete stochastic process** or simply **process**. It is a $\mathcal{L}^p$-process, if each X_n is in $\mathcal{L}^p$. A process is called **adapted to the filtration** $\{\mathcal{A}_n\}$ if X_n is $\mathcal{A}_n$-measurable for all $n \in \mathbb{N}$.

Example. For $\Omega = \{0,1\}^{\mathbb{N}}$ as above, the process $X_n(x) = \prod_{i=1}^n x_i$ is a stochastic process adapted to the filtration. Also $S_n(x) = \sum_{i=1}^n x_i$ is adapted to the filtration.

Definition. A $\mathcal{L}^1$-process which is adapted to a filtration $\{\mathcal{A}_n\}$ is called a **martingale** if

$$\mathrm{E}[X_n|\mathcal{A}_{n-1}] = X_{n-1}$$

for all $n \geq 1$. It is called a **supermartingale** if $\mathrm{E}[X_n|\mathcal{A}_{n-1}] \leq X_{n-1}$ and a **submartingale** if $\mathrm{E}[X_n|\mathcal{A}_{n-1}] \geq X_{n-1}$. If we mean either submartingale or supermartingale (or martingale) we speak of a **semimartingale**.

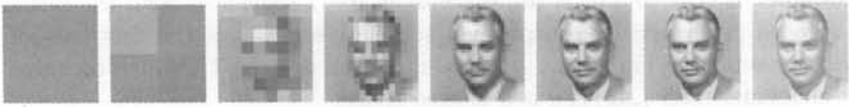
Remark. It immediately follows that for a martingale

$$\mathrm{E}[X_n|\mathcal{A}_m] = X_m$$

if $m < n$ and that $\mathrm{E}[X_n]$ is constant. Allan Gut mentions in [34] that a martingale is an allegory for "life" itself: the expected state of the future given the past history is equal the present state and on average, nothing happens.

Figure. *A random variable X on the unit square defines a gray scale picture if we interpret $X(x, y)$ is the gray value at the point (x, y). It shows Joseph Leo Doob (1910-2004), who developed basic martingale theory and many applications. The partitions $\mathcal{A}_n = \{[k/2^n(k+1)/2^n) \times [j/2^n(j+1)/2^n)\}$ define a filtration of $\Omega = [0, 1] \times [0, 1]$. The sequence of pictures shows the conditional expectations $\mathrm{E}[X_n|\mathcal{A}_n]$. It is a martingale.*

Exercice. Determine from the following sequence of pictures, whether it is a supermartingale or a submartingale. The images get brighter and brighter in average as the resolution becomes better.



Definition. If a martingale X_n is given with respect to a filtered space $\mathcal{A}_n = \sigma(Y_0, \dots, Y_n)$, where Y_n is a given process, X is is called a **martingale with respect** Y.

Remark. The word "martingale" means a gambling system in which losing bets are doubled. It is also the name of a part of a horse's harness or a belt on the back of a man's coat.

Remark. If X is a supermartingale, then $-X$ is a submartingale and vice versa. A supermartingale, which is also a submartingale is a martingale. Since we can change X to $X - X_0$ without destroying any of the martingale properties, we could assume the process is **null at** 0 which means $X_0 = 0$.

Exercice. a) Verify that if X_n, Y_n are two submartingales, then $\sup(X, Y)$ is a submartingale.
b) If X_n is a submartingale, then $\mathrm{E}[X_n] \leq \mathrm{E}[X_{n-1}]$.
c) If X_n is a martingale, then $\mathrm{E}[X_n] = \mathrm{E}[X_{n-1}]$.

Remark. Given a martingale. From the tower property of conditional expectation follows that for $m < n$

$$\mathrm{E}[X_n|\mathcal{A}_m] = \mathrm{E}[E[X_n|\mathcal{A}_{n-1}]|\mathcal{A}_m] = \mathrm{E}[X_{n-1}|\mathcal{A}_m] = \cdots = \mathrm{E}[X_m] .$$

Example. Sum of independent random variables
Let $X_i \in \mathcal{L}^1$ be a sequence of independent random variables with mean $\mathrm{E}[X_i] = 0$. Define $S_0 = 0, S_n = \sum_{k=1}^n X_k$ and $\mathcal{A}_n = \sigma(X_1, \dots, X_n)$ with $\mathcal{A}_0 = \{\emptyset, \Omega\}$. Then S_n is a martingale since S_n is an $\{\mathcal{A}_n\}$-adapted $\mathcal{L}^1$-process and

$$\mathrm{E}[S_n|\mathcal{A}_{n-1}] = \mathrm{E}[S_{n-1}|\mathcal{A}_{n-1}] + \mathrm{E}[X_n|\mathcal{A}_{n-1}] = S_{n-1} + \mathrm{E}[X_n] = S_{n-1} \; .$$

We have used linearity, the independence property of the conditional expectation.

Example. Conditional expectation
Given a random variable $X \in \mathcal{L}^1$ on a filtered space $(\Omega, \mathcal{A}, \{\mathcal{A}_n\}_{n \in \mathbb{N}}, \mathrm{P})$. Then $X_n = \mathrm{E}[X|\mathcal{A}_n]$ is a martingale.
Especially: given a sequence Y_n of random variables. Then $\mathcal{A}_n = \sigma(Y_0, \dots, Y_n)$ is a filtered space and $X_n = \mathrm{E}[X|Y_0, \dots, Y_n]$ is a martingale. Proof: by the tower property

$$\begin{aligned} \mathrm{E}[X_n|Y_0, \dots, Y_{n-1}] &= \mathrm{E}[[X_n|Y_0, \dots, Y_n]|Y_0, \dots, Y_{n-1}] \\ &= \mathrm{E}[X_n|Y_0, \dots, Y_{n-1}] = X_{n-1} \end{aligned}$$

We say X is a **martingale with respect to** Y. Note that because X_n is by definition $\sigma(Y_0, \dots, Y_n)$-measurable, there exist Borel measurable functions $h_n : \mathbb{R}^{n+1} \to \mathbb{R}$ such that $X_n = h_n(Y_0, \dots, Y_{n-1})$.

Example. Product of positive variables
Given a sequence Y_n of independent random variables $Y_n \geq 0$ satisfying with $\mathrm{E}[Y_n] = 1$. Define $X_0 = 1$ and $X_n = \prod_{i=0}^n Y_i$ and $\mathcal{A}_n = \sigma(Y_1, \dots, Y_n)$. Then X_n is a martingale. This is an exercise. Note that the martingale property does not follow directly by taking logarithms.

Example. Product of matrix-valued random variables
Given a sequence of independent random variables Z_n with values in the group $GL(N, \mathbb{R})$ of invertible $N \times N$ matrices and let $\mathcal{A}_n = \sigma(Z_1, \dots, Z_n)$. Assume $\mathrm{E}[\log ||Z_n||] \leq 0$, if $||Z_n||$ denotes the norm of the matrix (the square root of the maximal eigenvalue of $Z_n \cdot Z_n^*$, where Z_n^* is the adjoint). Define the real-valued random variables $X_n = \log ||Z_1 \cdot Z_2 \cdots Z_n||$, where $\cdot$ denotes matrix multiplication. Because $X_n \leq \log ||Z_n|| + X_{n-1}$, we get

$$\begin{aligned} \mathrm{E}[X_n|\mathcal{A}_{n-1}] &\leq \mathrm{E}[\log ||Z_n|| \mid \mathcal{A}_{n-1}] + \mathrm{E}[X_{n-1}|\mathcal{A}_{n-1}] \\ &= \mathrm{E}[\log ||Z_n||] + X_{n-1} \leq X_{n-1} \end{aligned}$$

so that X_n is a supermartingale. In ergodic theory, such a matrix-valued process X_n is called **sub-additive**.

Example. If Z_n is a sequence of matrix valued random variables, we can also look at the sequence of random variables $Y_n = ||Z_1 \cdot Z_2 \cdots Z_n||$. If $\mathrm{E}[||Z_n||] = 1$, then Y_n is a supermartingale.

Example. Polya's urn scheme
An urn contains initially a red and a black ball. At each time $n \geq 1$, a ball is taken randomly, its color noted, and both this ball and another ball of the same color are placed back into the urn. Like this, after n draws, the urn contains $n+2$ balls. Define Y_n as the number of black balls after n moves and $X_n = Y_n/(n+2)$, the fraction of black balls. We claim that X is a martingale with respect to Y: the random variables Y_n take values in $\{1, \dots, n+1\}$. Clearly $\mathrm{P}[Y_{n+1} = k+1 | Y_n = k] = k/(n+2)$ and $\mathrm{P}[Y_{n+1} = k | Y_n = k] = 1 - k/(n+2)$. Therefore

$$\begin{aligned} \mathrm{E}[X_{n+1}|Y_1, \dots, Y_n] &= \frac{1}{n+3}\mathrm{E}[Y_{n+1}|Y_1, \dots, Y_n] \\ &= \frac{1}{n+3}\mathrm{P}[Y_{n+1} = k+1|Y_n = k] \cdot \mathrm{P}[Y_{n+1}] \\ &\quad +\mathrm{P}[Y_{n+1} = k \mid Y_n = k] \cdot \mathrm{P}[Y_n] \\ &= \frac{1}{n+3}[(Y_n+1)\frac{Y_n}{n+2} + Y_n(1 - \frac{Y_n}{n+2})] \\ &= \frac{Y_n}{n+2} = X_n . \end{aligned}$$

Note that X_n is not independent of X_{n-1}. The process "learns" in the sense that if there are more black balls, then the winning chances are better.

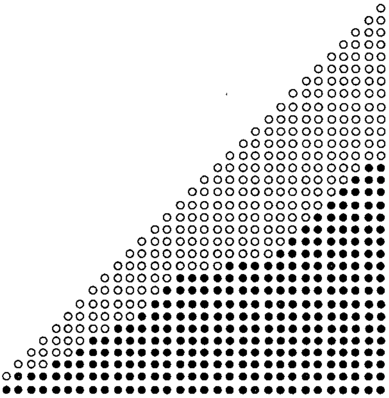


Figure. *A typical run of 30 experiments with Polya's urn scheme.*

Example. Branching processes
Let Z_{ni} be IID, integer-valued random variables with positive finite mean m. Define $Y_0 = 1$ and

$$Y_{n+1} = \sum_{k=1}^{Y_n} Z_{nk}$$

with the convention that for $Y_n = 0$, the sum is zero. We claim that $X_n = Y_n/m^n$ is a martingale with respect to Y. By the independence of Y_n and

$Z_{ni}, i \geq 1$, we have for every n

$$\mathrm{E}[Y_{n+1}|Y_0, \dots, Y_n] = \mathrm{E}[\sum_{k=1}^{Y_n} Z_{nk}|Y_0, \dots, Y_n] = \mathrm{E}[\sum_{k=1}^{Y_n} Z_{nk}] = mY_n$$

so that

$$\mathrm{E}[X_{n+1}|Y_0, \dots, Y_n] = \mathrm{E}[Y_{n+1}|Y_0, \dots Y_n]/m^{n+1} = mY_n/m^{n+1} = X_n \ .$$

The branching process can be used to model population growth, disease epidemic or nuclear reactions. In the first case, think of Y_n as the size of a population at time n and with Z_{ni} the number of progenies of the $i - th$ member of the population, in the n'th generation.

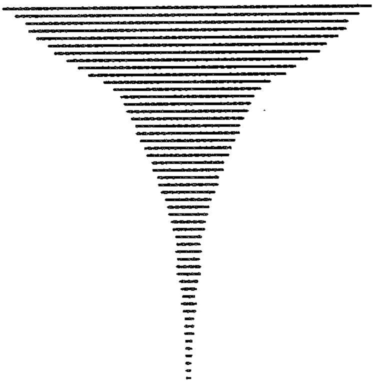


Figure. *A typical growth of Y_n of a branch process. In this example, the random variables Z_{ni} had a Poisson distribution with mean $m = 1.1$. It is possible that the process dies out, but often, it grows exponentially.*

Proposition 3.2.1. Let $\mathcal{A}_n$ be a fixed filtered sequence of σ-algebras. Linear combinations of martingales over $\mathcal{A}_n$ are again martingales over $\mathcal{A}_n$. Submartingales and supermartingales form cones: if for example X, Y are submartingales and $a, b > 0$, then $aX + bY$ is a submartingale.

Proof. Use the linearity and positivity of the conditional expectation. □

Proposition 3.2.2. a) If X is a martingale and u is convex such that $u(X_n) \in \mathcal{L}^1$, then $Y = u(X)$ is a submartingale. Especially, if X is a martingale, then $|X|$ is a submartingale.
b) If u is monotone and convex and X is a submartingale such that $u(X_n) \in \mathcal{L}^1$, then $u(X)$ is a submartingale.

Proof. a) We have by the conditional Jensen property (3.1.4)

$$Y_n = u(X_n) = u(E[X_{n+1}|\mathcal{A}_n]) \leq \mathrm{E}[u(X_{n+1})|\mathcal{A}_n] = \mathrm{E}[Y_{n+1}|\ |\mathcal{A}_n] .$$

b) Use the conditional Jensen property again and the monotonicity of u to get

$$Y_n = u(X_n) \leq u(E[X_{n+1}|\mathcal{A}_n]) \leq \mathrm{E}[u(X_{n+1})|\mathcal{A}_n] = \mathrm{E}[Y_{n+1}|\ |\mathcal{A}_n] .$$

□

Definition. A stochastic process $C = \{C_n\}_{n\geq 1}$ is called **previsible** if C_n is $\mathcal{A}_{n-1}$-measurable. A process X is called **bounded**, if $X_n \in \mathcal{L}^\infty$ and if there exists $K \in \mathbb{R}$ such that $||X_n||_\infty \leq K$ for all $n \in \mathbb{N}$.

Definition. Given a semimartingale X and a previsible process C, the process

$$(\int C\, dX)_n = \sum_{k=1}^{n} C_k(X_k - X_{k-1}) .$$

It is called a **discrete stochastic integral** or a **martingale transform.**

Theorem 3.2.3 (The system can't be beaten). If C is a bounded nonnegative previsible process and X is a supermartingale then $\int C\, dX$ is a supermartingale. The same statement is true for submartingales and martingales.

Proof. Let $Y = \int C\, dX$. From the property of "extracting knowledge" in theorem (3.1.4), we get

$$\mathrm{E}[Y_n - Y_{n-1}|\mathcal{A}_{n-1}] = \mathrm{E}[C_n(X_n - X_{n-1})|\mathcal{A}_{n-1}] = C_n \cdot \mathrm{E}[X_n - X_{n-1}|\mathcal{A}_{n-1}] \leq 0$$

because C_n is nonnegative and X_n is a supermartingale. □

Remark. If one wants to relax the boundedness of C, then one has to strengthen the condition for X. The proposition stays true, if both C and X are $\mathcal{L}^2$-processes.

Remark. Here is an interpretation: if X_n represents your **capital** in a game, then $X_n - X_{n-1}$ are the **net winnings** per unit stake. If C_n is the **stake** on game n, then

$$\int C\, dX = \sum_{k=1}^{n} C_k(X_k - X_{k-1})$$

are the **total winnings** up to time n. A martingale represents a **fair game** since $\mathrm{E}[X_n - X_{n-1}|\mathcal{A}_{n-1}] = 0$, whereas a supermartingale is a game which is **unfavorable** to you. The above proposition tells that you can not find a strategy for putting your stake to make the game fair.

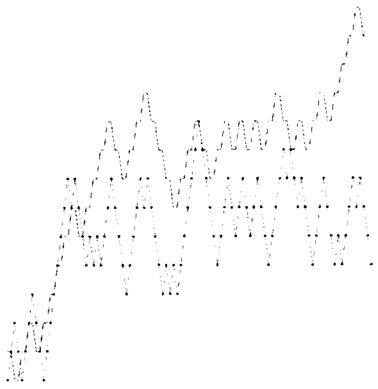


Figure. *In this example, $X_n = \pm 1$ with probability 1/2 and $C_n = 1$ if X_{n-1} is even and $C_n = 0$ if X_{n-1} is odd. The original process X_n is a symmetric random walk and so a martingale. The new process $\int C\, dX$ is again a martingale.*

Exercice. a) Let $Y_1, Y_2, \dots$ be a sequence of independent non-negative random variables satisfying $\mathrm{E}[Y_k] = 1$ for all $k \in \mathbb{N}$. Define $X_0 = 1, X_n = Y_1 \cdots Y_n$ and $\mathcal{A}_n = \sigma(Y_1, Y_2, \dots, Y_n)$. Show that X_n is a martingale.
b) Let Z_n be a sequence of independent random variables taking values in the set of $n \times n$ matrices satisfying $\mathrm{E}[||Z_n||] = 1$. Define $X_0 = 1, X_n = ||Z_1 \cdots Z_n||$. Show that X_n is a supermartingale.

Definition. A random variable T with values in $\overline{\mathbb{N}} = \mathbb{N} \cup \{\infty\}$ is called a **random time**. Define $\mathcal{A}_\infty = \sigma(\bigcup_{n \geq 0} \mathcal{A}_n)$. A random time T is called a **stopping time** with respect to a filtration $\mathcal{A}_n$, if $\{T \leq n\} \in \mathcal{A}_n$ for all $n \in \overline{\mathbb{N}}$.

Remark. A random time T is a stopping time if and only if $\{T = n\} \in \mathcal{A}_n$ for all $n \in \overline{\mathbb{N}}$ since $\{T \leq n\} = \bigcup_{0 \leq k \leq n} \{T = k\} \in \mathcal{A}_n$.

Remark. Here is an interpretation: stopping times are random times, whose occurrence can be determined without pre-knowledge of the future. The term comes from **gambling**. A gambler is forced to stop to play if his capital is zero. Whether or not you stop after the $n-$th game depends only on the history up to and including the time n.

Example. First entry time.
Let X_n be a $\mathcal{A}_n$-adapted process and given a Borel set $B \in \mathcal{B}$ in $\mathbb{R}^d$. Define

$$T(\omega) = \inf\{n \geq 0 \mid X_n(\omega) \in B\}$$

which is the time of first entry of X_n into B. The set $\{T = \infty\}$ is the set which never enters into B. Obviously

$$\{T \leq n\} = \bigcup_{k=0}^{n} \{X_k \in B\} \in \mathcal{A}_n$$

so that T is a stopping time.

Example. "Continuous Black-Jack": let X_i be IID random variables with uniform distribution in $[0,1]$. Define $S_n = \sum_{k=1}^n X_i$ and let $T(\omega)$ be the smallest integer so that $S_n(\omega) > 1$. This is a stopping time. A popular problem asks for the expectation of this random variable T: How many "cards" X_i do we have to draw until we get busted and the sum is larger than 1? We obviously have $\mathrm{P}[T=1]=0$. Now, $\mathrm{P}[T=2]=\mathrm{P}[X_2 > 1-X_1]$ is the area of region $\{(x,y) \in [0,1] \times [0,1] \mid y > 1-x \}$ which is 1/2. Similarly $\mathrm{P}[T=3] = \mathrm{P}[X_3 > 1 - X_1 - X_2]$ is the volume of the solid $\{(x,y,z) \in [0,1]^3 \mid z > 1-x-y \}$ which is $1/6 = 1/3!$. Inductively we see $\mathrm{P}[T=k] = 1/k!$ and the expectation of T is $\mathrm{E}[T] = \sum_{k=1}^\infty k/k! = \sum_{k=0}^\infty 1/k! = e$. This means that if we play Black-Jack with uniformly distributed random variables and threshold 1, we expect to get busted in more than 2, but less than 3 "cards".

Example. Last exit time.
Assume the same setup as in 1). But this time

$$T(\omega) = \sup\{n \geq 0 \mid X_n(\omega) \in B\}$$

is **not** a stopping time since it is impossible to know that X will return to B after some time k without knowing the whole future.

Proposition 3.2.4. Let T_1, T_2 be two stopping times. The infimum $T_1 \wedge T_2$, the maximum $T_1 \vee T_2$ as well as the sum $T_1 + T_2$ are stopping times.

Proof. This is obvious from the definition because $\mathcal{A}_n$-measurable functions are closed by taking minima, maxima and sums. □

Definition. Given a stochastic process X_n which is adapted to a filtration $\mathcal{A}_n$ and let T be a stopping time with respect to $\mathcal{A}_n$, define the random variable

$$X_T(\omega) = \begin{cases} X_{T(\omega)}(\omega) & , T(\omega) < \infty \\ 0 & , \text{else} \end{cases}$$

or equivalently $X_T = \sum_{n=0}^\infty X_n 1_{\{T=n\}}$. The process $X_n^T = X_{T \wedge n}$ is called the **stopped process**. It is equal to X_T for times $T \leq n$ and equal to X_n if $T \geq n$.

Proposition 3.2.5. If X is a supermartingale and T is a stopping time, then the stopped process X^T is a supermartingale. In particular $\mathrm{E}[X^T] \leq \mathrm{E}[X_0]$. The same statement is true if supermartingale is replaced by martingale in which case $\mathrm{E}[X^T] = \mathrm{E}[X_0]$.

Proof. Define the "stake process" $C^{(T)}$ by $C_n^{(T)} = 1_{n \leq T}$. You can think of it as betting 1 unit and quit playing immediately after time T. Define then the "winning process"

$$(\int C^{(T)}\, dX)_n = \sum_{k=1}^{n} C_k(X_k - X_{k-1}) = X_{T \wedge n} - X_0 \; .$$

or shortly $\int C^{(T)}\, dX = X_T - X_0$. The process C is previsible, since it can only take values 0 and 1 and $\{C_n^{(T)} = 0\} = \{T \leq n-1\} \in \mathcal{A}_{n-1}$. The claim follows from the "system can't be beaten" theorem. □

Remark. It is important that we take the stopped process X^T and not the random variable X_T:
for the random walk X on $\mathbb{Z}$ starting at 0, let T be the stopping time $T = \inf\{n \mid X_n = 1\}$. This is the martingale strategy in casino which gave the name of these processes. As we will see later on, the random walk is recurrent $\mathrm{P}[T < \infty] = 1$ in one dimensions. However

$$1 = \mathrm{E}[X_T] \neq \mathrm{E}[X_0] = 0 \; .$$

The above theorem gives $\mathrm{E}[X^T] = \mathrm{E}[X_0]$.

When can we say $\mathrm{E}[X_T] = \mathrm{E}[X_0]$? The answer gives Doob's optimal stopping time theorem:

Theorem 3.2.6 (Doob's optimal stopping time theorem). Let X be a supermartingale and T be a stopping time. If one of the five following conditions are true

(i) T is bounded.
(ii) X is bounded and T is almost everywhere finite.
(iii) $T \in \mathcal{L}^1$ and $|X_n - X_{n-1}|$ is bounded.
(iv) $X_T \in \mathcal{L}^1$ and $\lim_{k \to \infty} \mathrm{E}[X_k; \{T > k\}] = 0$.
(v) X is uniformly integrable and T is almost everywhere finite.

then $\mathrm{E}[X_T] \leq \mathrm{E}[X_0]$.
If X is a martingale and any of the five conditions is true, then $\mathrm{E}[X_T] = \mathrm{E}[X_0]$.

Proof. We know that $\mathrm{E}[X_{T \wedge n} - X_0] \leq 0$ because X is a supermartingale.
(i) Because T is bounded, we can take $n = \sup T(\omega) < \infty$ and get

$$\mathrm{E}[X_T - X_0] = \mathrm{E}[X_{T \wedge n} - X_0] \leq 0 \; .$$

(ii) Use the dominated convergence theorem to get

$$\lim_{n \to \infty} \mathrm{E}[X_{T \wedge n} - X_0] \leq 0 \; .$$

(iii) We have a bound $|X_n - X_{n-1}| \leq K$ and so

$$|X_{T\wedge n} - X_0| = |\sum_{k=1}^{T\wedge n} X_k - X_{k-1}| \leq KT .$$

Because $KT \in \mathcal{L}^1$, the result follows from the dominated convergence theorem.

(iv) By (i), we get $\mathrm{E}[X_0] = \mathrm{E}[X_{T\wedge k}] = \mathrm{E}[X_T; \{T \leq k\}] + \mathrm{E}[X_k; \{T > k\}]$ and taking the limit gives $\mathrm{E}[X_0] = \lim_{n\to\infty} \mathrm{E}[X_k; \{T \leq k\}] \to \mathrm{E}[X_T]$ by the dominated convergence theorem and the assumption.

(v) The uniformly integrability $\mathrm{E}[|X_n|; |X_n| > R] \to 0$ for $R \to \infty$ assures that $X_T \in \mathcal{L}^1$ since $\mathrm{E}[|X_T|] \leq k \cdot \max_{1\leq i\leq n} \mathrm{E}[|X_k|] + \sup_n \mathrm{E}[|X_n|; \{T > k\}] < \infty$. Since $|E[X_k; \{T > k\}]| \leq \sup_n \mathrm{E}[|X_n|; \{T > k\}] \to 0$, we can apply (iv).

If X is a martingale, we use the supermartingale case for both X and $-X$. □

Remark. The interpretation of this result is that a fair game cannot be made unfair by sampling it with **bounded** stopping times.

Theorem 3.2.7 (No winning strategy). Assume X is a martingale and suppose $|X_n - X_{n-1}|$ is bounded. Given a previsible process C which is bounded and let $T \in \mathcal{L}^1$ be a stopping time, then $\mathrm{E}[(\int C dX)_T] = 0$.

Proof. We know that $\int C\, dX$ is a martingale and since $(\int C\, dX)_0 = 0$, the claim follows from the optimal stopping time theorem part (iii). □

Remark. The **martingale strategy** mentioned in the introduction shows that for unbounded stopping times, there is a winning strategy. With the martingale strategy one has $T = n$ with probability $1/2^n$. The player always wins, she just has to double the bet until the coin changes sign. But it assumes an "infinitely thick wallet". With a finite but large initial capital, there is a very small risk to lose, but then the loss is large. You see that in the real world: players with large capital in the stock market mostly win, but if they lose, their loss can be huge.

Martingales can be characterized involving stopping times:

Theorem 3.2.8 (Komatsu's lemma). Let X be an $\mathcal{A}_n$-adapted sequence of random variables in $\mathcal{L}^1$ such that for every bounded stopping time T

$$\mathrm{E}[X_T] = \mathrm{E}[X_0] ,$$

then X is a martingale with respect to $\mathcal{A}_n$.

Proof. Fix $n \in \mathbb{N}$ and $A \in \mathcal{A}_n$. The map

$$T = n + 1 - 1_A = \begin{cases} n & \omega \in A \\ n+1 & \omega \notin A \end{cases}$$

is a stopping time because $\sigma(T) = \{\emptyset, A, A^c, \Omega\ \} \subset \mathcal{A}_n$. Apply $\mathrm{E}[X_T] = \mathrm{E}[X_0]$ and $\mathrm{E}[X_{T'}] = \mathrm{E}[X_0]$ for the bounded constant stopping time $T' = n + 1$ to get

$$\begin{aligned} \mathrm{E}[X_n; A] + \mathrm{E}[X_{n+1}; A^c] &= \mathrm{E}[X_T] = \mathrm{E}[X_0] = \mathrm{E}[X_{T'}] = \mathrm{E}[X_{n+1}] \\ &= \mathrm{E}[X_{n+1}; A] + \mathrm{E}[X_{n+1}; A^c] \end{aligned}$$

so that $\mathrm{E}[X_{n+1}; A] = \mathrm{E}[X_n; A]$. Since this is true, for any $A \in \mathcal{A}_n$, we know that $\mathrm{E}[X_{n+1}|\mathcal{A}_n] = \mathrm{E}[X_n|\mathcal{A}_n] = X_n$ and X is a martingale. □

Example. The gambler's ruin problem is the following question: Let Y_i be IID with $\mathrm{P}[Y_i = \pm 1] = 1/2$ and let $X_n = \sum_{k=1}^{n} Y_i$ be the random walk with $X_0 = 0$. We know that X is a martingale with respect to Y. Given $a, b > 0$, we define the stopping time

$$T = \min\{n \geq 0 \mid X_n = b, \text{ or } X_n = -a\ \} .$$

We want to compute $\mathrm{P}[X_T = -a]$ and $\mathrm{P}[X_T = b]$ in dependence of a, b.

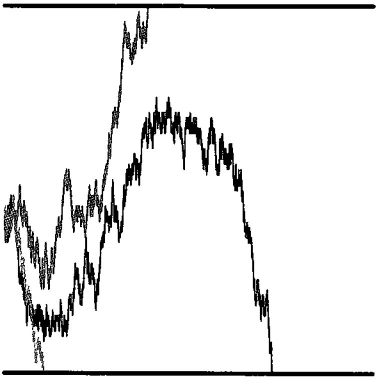


Figure. *Three samples of a process X_n starting at $X_0 = 0$. The process is stopped with the stopping time T, when X_n hits the lower bound $-a$ or the upper bound b. If X_n is the winning of a first gambler, which is the loss of a second gambler, then T is the time, for which one of the gamblers is broke. The initial capital of the first gambler is a, the initial capital of the second gambler is b.*

Remark. If Y_i are the outcomes of a series of fair gambles between two players A and B and the random variables X_n are the net change in the fortune of the gamblers after n independent games. If at the beginning, A has fortune a and B has fortune b, then $\mathrm{P}[X_T = -a]$ is the **ruin probability** of A and $\mathrm{P}[X_T = b]$ is the **ruin probability** of B.

Proposition 3.2.9.

$$\mathrm{P}[X_T = -a] = 1 - \mathrm{P}[X_T = b] = \frac{b}{(a+b)} .$$

Proof. T is finite almost everywhere. One can see this by the law of the iterated logarithm,

$$\limsup_n \frac{X_n}{\Lambda_n} = 1, \ \liminf_n \frac{X_n}{\Lambda_n} = -1 .$$

(We will give later a direct proof the finiteness of T, when we treat the random walk in more detail.) It follows that $\mathrm{P}[X_T = -a] = 1 - \mathrm{P}[X_T = b]$. We check that X_k satisfies condition (iv) in Doob's stopping time theorem: since X_T takes values in $\{a, b\ \}$, it is in $\mathcal{L}^1$ and because on the set $\{T > k\ \}$, the value of X_k is in $(-a, b)$, we have $|E[X_k; \{T > k\ \}]| \leq \max\{a, b\}\mathrm{P}[T > k] \to 0$. □

Remark. The boundedness of T is necessary in Doob's stopping time theorem. Let $T = \inf\{n \mid X_n = 1\ \}$. Then $\mathrm{E}[X_T] = 1$ but $\mathrm{E}[X_0] = 0]$ which shows that some condition on T or X has to be imposed. This fact leads to the "martingale" gambling strategy defined by doubling the bet when loosing. If the casinos would not impose a bound on the possible inputs, this gambling strategy would lead to wins. But you have to go there with enough money. One can see it also like this, If you are A and the casino is B and $b = 1$, $a = \infty$ then $\mathrm{P}[X_T = b] = 1$, which means that the casino is ruined with probability 1.

Theorem 3.2.10 (Wald's identity). Assume T is a stopping time of a $\mathcal{L}^1$-process Y for which Y_i are IID random variables with expectation $\mathrm{E}[Y_i] = m$ and $T \in \mathcal{L}^1$. The process $S_n = \sum_{k=1}^n Y_k$ satisfies

$$\mathrm{E}[S_T] = m\mathrm{E}[T] .$$

Proof. The process $X_n = S_n - n\mathrm{E}[Y_1]$ is a martingale satisfying condition (iii) in Doob's stopping time theorem. Therefore

$$0 = \mathrm{E}[X_0] = \mathrm{E}[X_T] = \mathrm{E}[S_T - T\mathrm{E}[Y_1]] .$$

Now solve for $\mathrm{E}[S_T]$. □

3.3 Doob's convergence theorem

Definition. Given a stochastic process X and two real numbers $a < b$, we define the random variable

$$\begin{aligned} U_n[a,b](\omega) &= \max\{k \in \mathbb{N} \mid \exists \\ & \quad 0 \le s_1 < t_1 < \cdots < s_k < t_k \le n, \\ & \quad X_{s_i}(\omega) < a, X_{t_i}(\omega) > b, 1 \le i \le k\} \end{aligned}$$

called the **number of up-crossings** of $[a,b]$. Denote with $U_\infty[a,b]$ the limit

$$U_\infty[a,b] = \lim_{n\to\infty} U_n[a,b] .$$

Because $n \mapsto U_n[a,b]$ is monotone, this limit exists in $\mathbb{N} \cup \{\infty\}$.

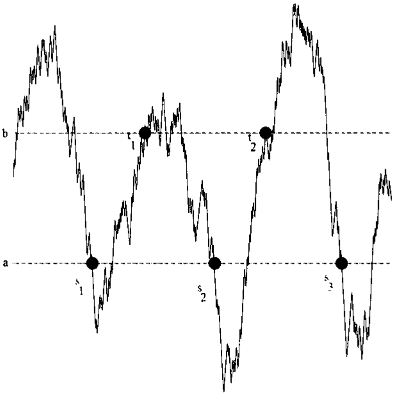


Figure. *A random walk crossing two values $a < b$. An up-crossing is a time s, where $X_s < a$ until the time, when the first time $X_t > b$. The random variable $U_n[a,b]$ with values in $\mathbb{N}$ measures the number of up-crossings in the time interval $[0,n]$.*

Theorem 3.3.1 (Doob's up-crossing inequality). If X is a supermartingale. Then

$$(b-a)\mathrm{E}[U_n[a,b]] \le \mathrm{E}[(X_n - a)^-] .$$

Proof. Define $C_1 = 1_{\{X_0 < a\}}$ and inductively for $n \ge 2$ the process

$$C_n := 1_{\{C_{n-1}=1\}} 1_{\{X_{n-1} \le b\}} + 1_{\{C_{n-1}=0\}} 1_{\{X_{n-1} < a\}} .$$

It is previsible. Define the winning process $Y = \int C \, dX$ which satisfies by definition $Y_0 = 0$. We have the **winning inequality**

$$Y_n(\omega) \geq (b-a)U_n[a,b](\omega) - (X_n(\omega) - a)^- .$$

Every up-crossing of $[a,b]$ increases the Y-value (the winning) by at least $(b-a)$, while $(X_n - a)^-$ is essentially the loss during the last interval of play.

Since C is previsible, bounded and nonnegative, we know that Y_n is also a supermartingale (see "the system can't be beaten") and we have therefore $\mathrm{E}[Y_n] \leq 0$. Taking expectation of the winning inequality, gives the claim. □

Remark. The proof uses the following strategy for putting your stakes C: wait until X gets below a. Play then unit stakes until X gets above b and stop playing. Wait again until X gets below a, etc.

Definition. We say, a stochastic process X_n is **bounded in $\mathcal{L}^p$**, if there exists $M \in \mathbb{R}$ such that $||X_n||_p \leq M$ for all $n \in \mathbb{N}$.

Corollary 3.3.2. If X is a supermartingale which is bounded in $\mathcal{L}^1$. Then

$$\mathrm{P}[U_\infty[a,b] = \infty] = 0 .$$

Proof. By the up-crossing lemma, we have for each $n \in \mathbb{N}$

$$(b-a)\mathrm{E}[U_n[a,b]] \leq |a| + \mathrm{E}[|X_n|] \leq |a| + \sup_n ||X_n||_1 < \infty .$$

By the dominated convergence theorem

$$(b-a)\mathrm{E}[U_\infty[a,b]] < \infty ,$$

which gives the claim. □

Remark. If $S_n = \sum_{k=1}^n X_k$ is the one dimensional random walk, then it is a martingale which is unbounded in $\mathcal{L}^1$. In this case, $\mathrm{E}[U_\infty[a,b]] = \infty$.

Theorem 3.3.3 (Doob's convergence theorem). Let X_n be a supermartingale which is bounded in $\mathcal{L}^1$. Then

$$X_\infty = \lim_{n \to \infty} X_n$$

exists almost everywhere.

Proof.

$$\begin{aligned}\Lambda &= \{\omega \in \Omega \mid X_n \text{ has no limit in } [-\infty,\infty]\,\} \\ &= \{\omega \in \Omega \mid \liminf X_n < \limsup X_n\,\} \\ &= \bigcup_{a<b,a,b\in\mathbb{Q}} \{\omega \in \Omega \mid \liminf X_n < a < b < \limsup X_n\,\} \\ &= \bigcup_{a<b,a,b\in\mathbb{Q}} \Lambda_{a,b}\,.\end{aligned}$$

Since $\Lambda_{a,b} \subset \{U_\infty[a,b] = \infty\,\}$ we have $\mathrm{P}[\Lambda_{a,b}] = 0$ and therefore also $\mathrm{P}[\Lambda] = 0$. Therefore $X_\infty = \lim_{n\to\infty} X_n$ exists almost surely. By Fatou's lemma

$$\mathrm{E}[|X_\infty|] = \mathrm{E}[\liminf_{n\to\infty} |X_n|) \le \liminf_{n\to\infty} \mathrm{E}[|X_n|] \le \sup_n \mathrm{E}[|X_n|] < \infty$$

so that $\mathrm{P}[X_\infty < \infty] = 1$. □

Example. Let X be a random variable on $([0,1),\mathcal{A},\mathrm{P})$, where P is the Lebesgue measure. The finite σ-algebra $\mathcal{A}_n$ generated by the intervals

$$A_k = [\frac{k}{2^n}, \frac{k+1}{2^n})$$

defines a filtration and $X_n = \mathrm{E}[X|\mathcal{A}_n]$ is a martingale which converges. We will see below with Lévys upward theorem (3.4.2 that the limit actually is the random variable X.

Example. Let X_k be IID random variables in $\mathcal{L}^1$. For $0 < \lambda < 1$, the branching random walk $S_n = \sum_{k=0}^n \lambda^k X_k$ is a martingale which is bounded in $\mathcal{L}^1$ because

$$||S_n||_1 \le \frac{1}{1-\lambda}||X_0||_1\,.$$

The martingale converges by Doob's convergence theorem almost surely. One can also deduce this from Kolmogorov's theorem (2.11.3) if $X_k \in \mathcal{L}^2$. Doobs convergence theorem (3.3.3) assures convergence for $X_k \in \mathcal{L}^1$.

Remark. Of course, we can replace supermartingale by submartingale or martingale in the theorem.

Example. We look again at Polya's urn scheme, which was defined earlier. Since the process Y giving the fraction of black balls is a martingale and bounded $0 \le Y \le 1$, we can apply the convergence theorem: there exists Y_∞ with $Y_n \to Y_\infty$.

Corollary 3.3.4. If X is a non-negative supermartingale, then $X_\infty = \lim_{n\to\infty} X_n$ exists almost everywhere and is finite.

Proof. Since the supermartingale property gives $\mathrm{E}[|X_n|] = \mathrm{E}[X_n] \leq \mathrm{E}[X_0]$, the process X_n is bounded in $\mathcal{L}^1$. Apply Doob's convergence theorem. □

Remark. This corollary is also true for non-positive submartingales or martingales, which are either nonnegative or non-positive.

Example. For the Branching process, we had IID random variables Z_{ni} with positive finite mean m and defined $Y_0 = 0$, $Y_{n+1} = \sum_{k=1}^{Y_n} Z_{nk}$. We saw that the process $X_n = Y_n/m^n$ is non-negative and a martingale. According to the above corollary, the limit X_∞ exists almost everywhere. It is an interesting problem to find the distribution of X_∞: Assume Z_{ni} have the generating function $f(\theta) = \mathrm{E}[\theta^{Z_{ni}}]$.

(i) Y_n has the generating function $f^n(\theta) = f(f^{n-1})(\theta)$.
We prove this by induction. For $n = 1$ this is trivial. Using the independence of Z_{nk} we have

$$\mathrm{E}[\theta^{Y_{n+1}}|Y_n = k] = f(\theta)^k$$

and so

$$\mathrm{E}[\theta^{Y_{n+1}}|Y_n] = f(\theta)^{Z_n} .$$

By the tower property, this leads to

$$\mathrm{E}[\theta^{Y_{n+1}}] = \mathrm{E}[f(\theta)^{Z_n}] .$$

Write $\alpha = f(\theta)$ and use induction to simplify the right hand side to

$$\mathrm{E}[f(\theta)^{Y_n}] = \mathrm{E}[\alpha^{Y_n}] = f^n(\alpha) = f^n(f(\theta)) = f^{n+1}(\theta) .$$

(ii) In order to find the distribution of X_∞ we calculate instead the characteristic function

$$L(\lambda) = L(X_\infty)(\lambda) = \mathrm{E}[\exp(i\lambda X_\infty)] .$$

Since $X_n \to X_\infty$ almost everywhere, we have $L(X_n)(\lambda) \to L(X_\infty)(\lambda)$. Since $X_n = Y_n/m^n$ and $\mathrm{E}[\theta^{Y_n}] = f^n(\theta)$, we have

$$L(X_n)(\lambda) = f^n(e^{i\lambda/m^n})$$

so that L satisfies the **functional equation**

$$L(\lambda m) = f(L(\lambda)) .$$

Theorem 3.3.5 (Limit distribution of the branching process). For the branching process defined by IID random variables Z_{ni} having the generating function f, the Fourier transform $L(\lambda) = \mathrm{E}[e^{i\lambda X_\infty}]$ of the distribution of the limit martingale X_∞ can be computed by solving the functional equation

$$L(\lambda \cdot m) = f(L(\lambda)) .$$

Remark. If f has no analytic extension to the complex plane, we have to replace the Fourier transform with the Laplace transform

$$L(\lambda) = \mathrm{E}[e^{-\lambda X_\infty}] .$$

Remark. Related to Doob's convergence theorem for supermartingales is Kingman's subadditive ergodic theorem, which generalizes Birkhoff's ergodic theorem and which we state without proof. Neither of the two theorems are however corollaries of each other.

Definition. A sequence of random variables X_n is called **subadditive** with respect to a measure preserving transformation T, if $X_{m+n} \leq X_m + X_n(T^m)$ almost everywhere.

Theorem 3.3.6 (The subadditive ergodic theorem of Kingmann). Given a sequence of random variables, which $X_n : X \to \mathbb{R} \cup \{-\infty\}$ with $X_n^+ := \max(0, X_n) \in L^1(X)$ and which is subadditive with respect to a measure preserving transformation T. Then there exists a T-invariant integrable measurable function $X : \Omega \to \mathbb{R} \cup \{-\infty\}$ such that $\frac{1}{n}X_n(x) \to X(x)$ for almost all $x \in X$. Furthermore $\frac{1}{n}\mathrm{E}[X_n] \to \mathrm{E}[X]$.

If the condition of boundedness of the process in Doob's convergence theorem is strengthened a bit by assuming that X_n is uniformly integrable, then one can reverse in some sense the convergence theorem:

Theorem 3.3.7 (Doob's convergence theorem for uniformly integrable supermartingales). A supermartingale X_n is uniformly integrable if and only if there exists X such that $X_n \to X$ in $\mathcal{L}^1$.

Proof. If X_n is uniformly integrable, then X_n is bounded in $\mathcal{L}^1$ and Doob's convergence theorem gives $X_n \to X$ almost everywhere. But a uniformly integrable family X_n which converges almost everywhere converges in $\mathcal{L}^1$. On the other hand, a sequence $X_n \in \mathcal{L}^1$ converging to $X \in \mathcal{L}^1$ is uniformly integrable. □

Theorem 3.3.8 (Characterization of uniformly integrable martingales). An $\mathcal{A}_n$-adapted process is an uniformly integrable martingale if and only if $X_n \to X$ in $\mathcal{L}^1$ and $X_n = \mathrm{E}[X|\mathcal{A}_n]$.

Proof. By Doob's convergence theorem for uniformly integrable supermartingales (3.3.7), we know the "if"-part. We already know that $X_n = \mathrm{E}[X|\mathcal{A}_n]$ is a martingale. What we have to show is that it is uniformly integrable. Given $\epsilon > 0$. Choose $\delta > 0$ such that for all $A \in \mathcal{A}$, the condition $\mathrm{P}[A] < \delta$ implies $\mathrm{E}[|X|; A] < \epsilon$. Choose further $K \in \mathbb{R}$ such that $K^{-1} \cdot \mathrm{E}[|X|] < \delta$. By Jensen's inequality

$$|X_n| = |E[X|\mathcal{A}_n]| \leq \mathrm{E}[|X||\mathcal{A}_n] \leq \mathrm{E}[|X|] \, .$$

Therefore

$$K \cdot \mathrm{P}[|X_n| > K] \leq \mathrm{E}[|X_n|] \leq \mathrm{E}[|X|] \leq \delta \cdot K$$

so that $\mathrm{P}[|X_n| > K] < \delta$. By definition of conditional expectation , $|X_n| \leq \mathrm{E}[|X||\mathcal{A}_n]$ and $\{|X_n| > K\} \in \mathcal{A}_n$

$$\mathrm{E}[|X_n|; |X_n| > K] \leq \mathrm{E}[|X|; |X_n| > K] < \epsilon \, .$$

□

Remark. As a summary we can say that supermartingale X_n which is either bounded in $\mathcal{L}^1$ or nonnegative or uniformly integrable converges almost everywhere.

Exercice. Let S and T be stopping times satisfying $S \leq T$.
a) Show that the process

$$C_n(\omega) = 1_{\{S(\omega) < n \leq T(\omega)\}}$$

is previsible.
b) Show that for every supermartingale X and stopping times $S \leq T$ the inequality

$$\mathrm{E}[X_T] \leq \mathrm{E}[X_S]$$

holds.

Exercice. In Polya's urn process, let Y_n be the number of black balls after n steps. Let $X_n = Y_n/(n+2)$ be the fraction of black balls. We have seen that X is a martingale.
a) Prove that $\mathrm{P}[Y_n = k] = 1/(n+1)$ for every $1 \leq k \leq n+1$.
b) Compute the distribution of the limit X_∞.

Exercice. a) Which polynomials f can you realize as generating functions of a probability distribution? Denote this class of polynomials with $\mathcal{P}$.
b) Design a martingale X_n, where the iteration of polynomials $P \in \mathcal{P}$ plays a role.

c) Use one of the consequences of Doob's convergence theorem to show that the dynamics of every polynomial $P \in \mathcal{P}$ on the positive axis can be conjugated to a linear map $T : z \mapsto mz$: there exists a map L such that

$$L \circ T(z) = P \circ L(z)$$

for every $z \in \mathbb{R}^+$.

Example. The branching process $Y_{n+1} = \sum_{k=1}^{Y_n} Z_{nk}$ defined by random variables Z_{nk} having generating function f and mean m defines a martingale $X_n = Y_n/m^n$. We have seen that the Laplace transform $L(\lambda) = \mathrm{E}[e^{-\lambda X_\infty}]$ of the limit X_∞ satisfies the functional equation

$$L(m\lambda) = f(L(\lambda)) .$$

We assume that the IID random variables Z_{nk} have the geometric distribution $\mathrm{P}[Z = k] = p(1-p)^k = pq^k$ with parameter $0 < p < 1$. The probability generating function of this distribution is

$$f(\theta) = \mathrm{E}[\theta^Z] = \sum_{k=1}^{\infty} pq^k\theta^k = \frac{p}{1-q\theta} .$$

As we have seen in proposition (2.12.5),

$$\mathrm{E}[Z] = \sum_{k=1}^{\infty} pq^k k = \frac{q}{p} .$$

The function $f^n(\theta)$ can be computed as

$$f^n(\theta) = \frac{pm^n(1-\theta) + q\theta - p}{qm^n(1-\theta) + q\theta - p} .$$

This is because f is a **Möbius transformation** and iterating f corresponds to look at the power $A^n = \begin{bmatrix} 0 & p \\ -q & 1 \end{bmatrix}^n$. This power can be computed by diagonalisating A:

$$A^n = (q-p)^{-1} \begin{bmatrix} 1 & p \\ 1 & q \end{bmatrix} \begin{bmatrix} p^n & 0 \\ 0 & q^n \end{bmatrix} \begin{bmatrix} q & -p \\ -1 & 1 \end{bmatrix} .$$

We get therefore

$$L(\lambda) = \mathrm{E}[e^{-\lambda X_\infty}] = \lim_{n\to\infty} \mathrm{E}[e^{-\lambda Y_n/m^n}] = \lim_{n\to\infty} f_n(e^{\lambda/m^n}) = \frac{p\lambda + q - p}{q\lambda + q - p} .$$

If $m \leq 1$, then the law of X_∞ is a Dirac mass at 0. This means that the process dies out. We see that in this case directly that $\lim_{n\to\infty} f_n(\theta) = 1$. In the case $m > 1$, the law of X_∞ has a point mass at 0 of weight $p/q = 1/m$ and an absolutely continuous part $(1/m - 1)^2 e^{(1/m-1)x}\, dx$. This can be seen by performing a "look up" in a table of Laplace transforms

$$L(\lambda) = \frac{p}{q} e^{-\lambda 0} + \int_0^\infty (1-p/q)^2 e^{(p/q-1)x} \cdot e^{-\lambda x}\, dx .$$

Definition. Define $p_n = \mathrm{P}[Y_n = 0]$, the probability that the process dies out until time n. Since $p_n = f^n(0)$ we have $p_{n+1} = f(p_n)$. If $f(p) = p$, p is called the **extinction probability**.

Proposition 3.3.9. For a branching process with $\mathrm{E}[Z] \geq 1$, the extinction probability is the unique solution of $f(x) = x$ in $(0, 1)$. For $\mathrm{E}[Z] \leq 1$, the extinction probability is 1.

Proof. The generating function $f(\theta) = \mathrm{E}[\theta^Z] = \sum_{n=0}^{\infty} \mathrm{P}[Z = n]\theta^n = \sum_n p_n \theta^n$ is analytic in $[0, 1]$. It is nondecreasing and satisfies $f(1) = 1$. If we assume that $\mathrm{P}[Z = 0] > 0$, then $f(0) > 0$ and there exists a unique solution of $f(x) = x$ satisfying $f'(x) < 1$. The orbit $f^n(u)$ converges to this fixed point for every $u \in (0, 1)$ and this fixed point is the extinction probability of the process. The value of $f'(0) = \mathrm{E}[Z]$ decides whether there exists an attracting fixed point in the interval $(0, 1)$ or not. □

3.4 Lévy's upward and downward theorems

Lemma 3.4.1. Given $X \in \mathcal{L}^1$. Then the class of random variables

$$\{Y = \mathrm{E}[X|\mathcal{B}] \mid \mathcal{B} \subset \mathcal{A}, \mathcal{B} \text{ is } \sigma - \text{algebra}\}$$

is uniformly integrable.

Proof. Given $\epsilon > 0$. Choose $\delta > 0$ such that for all $A \in \mathcal{A}$, $\mathrm{P}[A] < \delta$ implies $\mathrm{E}[|X|; A] < \epsilon$. Choose further $K \in \mathbb{R}$ such that $K^{-1} \cdot \mathrm{E}[|X|] < \delta$. By Jensen's inequality, $Y = \mathrm{E}[X|\mathcal{B}]$ satisfies

$$|Y| = |E[X|\mathcal{B}]| \leq \mathrm{E}[|X||\mathcal{B}] \leq \mathrm{E}[|X|] \,.$$

Therefore

$$K \cdot \mathrm{P}[|X_n| > K] \leq \mathrm{E}[|Y|] \leq \mathrm{E}[|X|] \leq \delta \cdot K$$

so that $\mathrm{P}[|Y| > K] < \delta$. Now, by definition of conditional expectation , $|Y| \leq \mathrm{E}[|X||\mathcal{B}]$ and $\{|Y| > K\} \in \mathcal{B}$

$$\mathrm{E}[|X_{\mathcal{B}}|; |X_{\mathcal{B}}| > K] \leq \mathrm{E}[|X|; |X_{\mathcal{B}}| > K] < \epsilon \,.$$

□

Definition. Denote by $\mathcal{A}_\infty$ the σ-algebra generated by $\bigcup_n \mathcal{A}_n$.

Theorem 3.4.2 (Lévy's upward theorem). Given $X \in \mathcal{L}^1$. Then $X_n = \mathrm{E}[X|\mathcal{A}_n]$ is a uniformly integrable martingale and X_n converges in $\mathcal{L}^1$ to $X_\infty = \mathrm{E}[X|\mathcal{A}_\infty]$.

Proof. The process X is a martingale. The sequence X_n is uniformly integrable by the above lemma. Therefore X_∞ exists almost everywhere by Doob's convergence theorem for uniformly integrable martingales, and since the family X_n is uniformly integrable, the convergence is in $\mathcal{L}^1$. We have to show that $X_\infty = Y := \mathrm{E}[X|\mathcal{A}_\infty]$.
By proving the claim for the positive and negative part, we can assume that $X \geq 0$ (and so $Y \geq 0$). Consider the two measures

$$Q_1(A) = \mathrm{E}[X; A],\ Q_2(A) = \mathrm{E}[X_\infty; A] .$$

Since $\mathrm{E}[X_\infty|\mathcal{A}_n] = \mathrm{E}[X|\mathcal{A}_n]$, we know that Q_1 and Q_2 agree on the π-system $\bigcup_n \mathcal{A}_n$. They agree therefore everywhere on $\mathcal{A}_\infty$. Define the event $A = \{\mathrm{E}[X|\mathcal{A}_\infty] > X_\infty \} \in \mathcal{A}_\infty$. Since $Q_1(A) - Q_2(A) = \mathrm{E}[E[X|\mathcal{A}_\infty] - X_\infty]; A] = 0$ we have $\mathrm{E}[X|\mathcal{A}_\infty] \leq X_\infty$ almost everywhere. Similarly also $X_\infty \leq X|\mathcal{A}_\infty]$ almost everywhere. □

As an application, we see a martingale proof of Kolmogorov's $0 - 1$ law:

Corollary 3.4.3. For any sequence $\mathcal{A}_n$ of independent σ-algebras, the tail σ-algebra $\mathcal{T} = \bigcap_n \mathcal{B}_n$ with $\mathcal{B}_n = \bigcup_{m>n} \mathcal{A}_m$ is trivial.

Proof. Given $A \in \mathcal{T}$, define $X = 1_A \in \mathcal{L}^\infty(\mathcal{T})$ and the σ-algebras $\mathcal{C}_n = \sigma(\mathcal{A}_1, \dots, \mathcal{A}_n)$. By Lévy's upward theorem (3.4.2),

$$X = \mathrm{E}[X|\mathcal{C}_\infty] = \lim_{n\to\infty} \mathrm{E}[X|\mathcal{C}_n] .$$

But since $\mathcal{C}_n$ is independent of $\mathcal{A}_n$ and X is $\mathcal{C}_n$ measurable, we have

$$\mathrm{P}[A] = \mathrm{E}[X] = \mathrm{E}[X|\mathcal{C}_n] \to X$$

and because X takes only the value 0 or 1 and $X = \mathrm{P}[A]$ shows that it must be constant, we get $\mathrm{P}[A] = 1$ or $\mathrm{P}[A] = 0$. □

Definition. A sequence $\mathcal{A}_{-n}$ of σ-algebras $\mathcal{A}_{-n}$ satisfying

$$\cdots \subset \mathcal{A}_{-n} \subset \mathcal{A}_{-(n-1)} \subset \cdots \subset \mathcal{A}_{-1}$$

is called a **downward filtration**. Define $\mathcal{A}_{-\infty} = \bigcap_n \mathcal{A}_{-n}$.

Theorem 3.4.4 (Lévy's downward theorem). Given a downward filtration $\mathcal{A}_{-n}$ and $X \in \mathcal{L}^1$. Define $X_{-n} = \mathrm{E}[X|\mathcal{A}_{-n}]$. Then $X_{-\infty} = \lim_{n\to\infty} X_{-n}$ converges in $\mathcal{L}^1$ and $X_{-\infty} = \mathrm{E}[X|\mathcal{A}_{-\infty}]$.

Proof. Apply Doob's up-crossing lemma to the uniformly integrable martingale

$$X_k, -n \leq k \leq -1 \ :$$

for all $a < b$, the number of up-crossings is bounded

$$U_k[a,b] \leq (|a| + ||X||_1)/(b-a) \, .$$

This implies in the same way as in the proof of Doob's convergence theorem that $\lim_{n\to\infty} Y_{-n}$ converges almost everywhere.
We show now that $X_{-\infty} = \mathrm{E}[X|\mathcal{A}_{-\infty}]$: given $A \in \mathcal{A}_{-\infty}$. We have $\mathrm{E}[X;A] = \mathrm{E}[X_{-n};A] \to \mathrm{E}[X_{-\infty};A]$. The same argument as before shows that $X_{-\infty} = \mathrm{E}[X;\mathcal{A}_{-\infty}]$. □

Lets also look at a martingale proof of the strong law of large numbers.

Corollary 3.4.5. Given $X_n \in \mathcal{L}^1$ which are IID and have mean m. Then $S_n/n \to m$ in $\mathcal{L}^1$.

Proof. Define the downward filtration $\mathcal{A}_{-n} = \sigma(S_n, S_{n+1}, \dots)$.
Since $\mathrm{E}[X_1|\mathcal{A}_{-n}] = \mathrm{E}[X_i|\mathcal{A}_{-n}] = \mathrm{E}[X_i|S_n, S_{n+1}, \dots] = X_i$, and $\mathrm{E}[X_1|\mathcal{A}_n] = S_n/n$. We can apply Lévy's downward theorem to see that S_n/n converges in $\mathcal{L}^1$. Since the limit X is in $\mathcal{T}$, it is by Kolmogorov's 0-1 law a constant c and $c = \mathrm{E}[X] = \lim_{n\to\infty} \mathrm{E}[S_n/n] = m$. □

3.5 Doob's decomposition of a stochastic process

Definition. A process X_n is **increasing**, if $\mathrm{P}[X_n \leq X_{n+1}] = 1$.

Theorem 3.5.1 (Doob's decomposition). Let X_n be an $\mathcal{A}_n$-adapted $\mathcal{L}^1$-process. Then

$$X = X_0 + N + A$$

where N is a martingale null at 0 and A is a previsible process null at 0. This decomposition is unique in L^1. X is a submartingale if and only if A is increasing.

Proof. If X has a Doob decomposition $X = X_0 + N + A$, then

$$\mathrm{E}[X_n - X_{n-1}|\mathcal{A}_{n-1}] = \mathrm{E}[N_n - N_{n-1}|\mathcal{A}_n] + \mathrm{E}[A_n - A_{n-1}|\mathcal{A}_{n-1}] = A_n - A_{n-1}$$

which means that

$$A_n = \sum_{k=1}^{n} \mathrm{E}[X_k - X_{k-1} | \mathcal{A}_{n-1}] \, .$$

If we **define** A like this, we get the required decomposition and the submartingale characterization is also obvious. □

Remark. The corresponding result for continuous time processes is deeper and called **Doob-Meyer decomposition theorem**. See theorem (4.17.2).

Lemma 3.5.2. Given $s, t, u, v \in \mathbb{N}$ with $s \leq t \leq u \leq v$. If X_n is a $\mathcal{L}^2$-martingale, then

$$\mathrm{E}[(X_t - X_s)(X_v - X_u)] = 0$$

and

$$\mathrm{E}[X_n^2] = \mathrm{E}[X_0^2] + \sum_{k=1}^{n} \mathrm{E}[(X_k - X_{k-1})^2] \, .$$

Proof. Because $\mathrm{E}[X_v - X_u | \mathcal{A}_u] = X_u - X_u = 0$, we know that $X_v - X_u$ is orthogonal to $\mathcal{L}^2(\mathcal{A}_u)$. The first claim follows since $X_t - X_x \in \mathcal{L}^2(\mathcal{A}_u)$. The formula

$$X_n = X_0 + \sum_{k=1}^{n} (X_k - X_{k-1})$$

expresses X_n as a sum of orthogonal terms and Pythagoras theorem gives the second claim. □

Corollary 3.5.3. A $\mathcal{L}^2$-martingale X is bounded in $\mathcal{L}^2$ if and only if $\sum_{k=1}^{\infty} \mathrm{E}[(X_k - X_{k-1})^2] < \infty$.

Proof.

$$\mathrm{E}[X_n^2] = \mathrm{E}[X_0^2] + \sum_{k=1}^{n} \mathrm{E}[(X_k - X_{k-1})^2] \leq \mathrm{E}[X_0^2] + \sum_{k=1}^{\infty} \mathrm{E}[(X_k - X_{k-1})^2] < \infty \, .$$

If on the other hand, X_n is bounded in $\mathcal{L}^2$, then $||X_n||_2 \leq K < \infty$ and $\sum_k \mathrm{E}[(X_k - X_{k-1})^2] \leq K + \mathrm{E}[X_0^2]$. □

Theorem 3.5.4 (Doob's convergence theorem for L^2-martingales). Let X_n be a $\mathcal{L}^2$-martingale which is bounded in $\mathcal{L}^2$, then there exists $X \in \mathcal{L}^2$ such that $X_n \to X$ in $\mathcal{L}^2$.

Proof. If X is bounded in $\mathcal{L}^2$, then, by monotonicity of the norm $||X||_1 \leq ||X||_2$, it is bounded in $\mathcal{L}^1$ so that by Doob's convergence theorem, $X_n \to X$ almost everywhere for some X. By Pythagoras and the previous corollary (3.5.3), we have

$$\mathrm{E}[(X - X_n)^2] \leq \sum_{k \geq n+1} \mathrm{E}[(X_k - X_{k-1})^2] \to 0$$

so that $X_n \to X$ in $\mathcal{L}^2$. □

Definition. Let X_n be a martingale in $\mathcal{L}^2$ which is null at 0. The conditional Jensen's inequality (3.1.4) shows that X_n^2 is a submartingale. Doob's decomposition theorem allows to write $X^2 = N + A$, where N is a martingale and A is a previsible increasing process. Define $A_\infty = \lim_{n\to\infty} A_n$ point wise, where the limit can take the value ∞ also. One writes also $\langle X \rangle$ for A so that

$$X^2 = N + \langle X \rangle \; .$$

Lemma 3.5.5. Assume X is a $\mathcal{L}^2$-martingale. X is bounded in $\mathcal{L}^2$ if and only if $\mathrm{E}[\langle X \rangle_\infty] < \infty$.

Proof. From $X^2 = N + A$, we get $\mathrm{E}[X_n^2] = \mathrm{E}[A_n]$ since for a martingale N, the equality $\mathrm{E}[N_n] = \mathrm{E}[N_0]$ holds and N is null at 0. Therefore, X is in $\mathcal{L}^2$ if and only if $\mathrm{E}[A_\infty] < \infty$ since $= \mathrm{E}[X_n^2] = \mathrm{E}[A_n]$ and A_n is increasing. □

We can now relate the convergence of X_n to the finiteness of $A_\infty = \langle X \rangle_\infty$.

Proposition 3.5.6. a) If $\lim_{n\to\infty} X_n(\omega)$ converges, then $A_\infty(\omega) < \infty$.
b) On the other hand, if $||X_n - X_{n-1}||_\infty \leq K$, then $A_\infty(\omega) < \infty$ implies the convergence of $\lim_{n\to\infty} X_n(\omega)$.

Proof. a) Because A is previsible, we can define for every k a stopping time $S(k) = \inf\{n \in \mathbb{N} \mid A_{n+1} > k\ \}$. The stopped process $A^{S(k)}$ is previsible because for $B \in \mathcal{B}_{\mathbb{R}}$ and $n \in \mathbb{N}$,

$$\{A_{n \wedge S(k)} \in B\} = A_1 \cup A_2$$

with

$$\begin{aligned} A_1 &= \bigcup_{i=0}^{n-1} \{S(k) = i; A_i \in B\} \in \mathcal{A}_{n-1} \\ A_2 &= \{A_n \in B\} \cap \{S(k) \leq n-1\}^c \in \mathcal{A}_{n-1} . \end{aligned}$$

Since

$$(X^{S(k)})^2 - A^{S_k} = (X^2 - A)^{S(k)}$$

is a martingale, we see that $\langle X^{S(k} \rangle = A^{S(k)}$. The later process $A^{S(k)}$ is bounded by k so that by the above lemma $X^{S(k)}$ is bounded in $\mathcal{L}^2$. Therefore $\lim_{n\to\infty} X_{n \wedge S(k)}$ exists almost surely. Combining this with

$$\{A_\infty < \infty\} = \bigcup_k \{S_k = \infty\}$$

proves the claim.
b) Suppose the claim is wrong and that

$$\mathrm{P}[A_\infty = \infty, \sup_n |X_n| < \infty] > 0 .$$

Then,

$$\mathrm{P}[T(c) = \infty; A_\infty = \infty] > 0$$

where $T(c)$ is the stopping time

$$T(c) = \inf\{n \mid |X_n| > c\} .$$

Now

$$\mathrm{E}[X^2_{T(c)\wedge n} - A_{T(c)\wedge n}] = 0$$

and $X^{T(c)}$ is bounded by $c + K$. Thus

$$\mathrm{E}[A_{T(c)\wedge n}] \leq (c+K)^2$$

for all n. This is a contradiction to $\mathrm{P}[A_\infty = \infty, \sup_n |X_n| < \infty] > 0$. □

Theorem 3.5.7 (A strong law for martingales). Let X be a $\mathcal{L}^2$-martingale zero at 0 and let $A = \langle X \rangle$. Then

$$\frac{X_n}{A_n} \to 0$$

almost surely on $\{A_\infty = \infty\ \}$.

Proof. (i) Césaro's lemma: Given $0 = b_0 < b_1 \leq \dots, b_n \leq b_{n+1} \to \infty$ and a sequence $v_n \in \mathbb{R}$ which converges $v_n \to v_\infty$, then $\frac{1}{b_n} \sum_{k=1}^n (b_k - b_{k-1}) v_k \to v_\infty$.

Proof. Let $\epsilon > 0$. Choose m such that $v_k > v_\infty - \epsilon$ if $k \geq m$. Then

$$\begin{aligned} \liminf_{n\to\infty} \frac{1}{b_n} \sum_{k=1}^n (b_k - b_{k-1}) v_k &\geq \liminf_{n\to\infty} \frac{1}{b_n} \sum_{k=1}^m (b_k - b_{k-1}) v_k \\ &\quad + \frac{b_n - b_m}{b_n} (v_\infty - \epsilon) \\ &\geq 0 + v_\infty - \epsilon \end{aligned}$$

Since this is true for every $\epsilon > 0$, we have $\liminf \geq v_\infty$. By a similar argument $\limsup \geq v_\infty$. □

(ii) Kronecker's lemma: Given $0 = b_0 < b_1 \leq \dots, b_n \leq b_{n+1} \to \infty$ and a sequence x_n of real numbers. Define $s_n = x_1 + \dots + x_n$. Then the convergence of $u_n = \sum_{k=1}^n x_k/b_k$ implies that $s_n/b_n \to 0$.

Proof. We have $u_n - u_{n-1} = x_n/b_n$ and

$$s_n = \sum_{k=1}^n b_k(u_k - u_{k-1}) = b_n u_n - \sum_{k=1}^n (b_k - b_{k-1}) u_{k-1} .$$

Césaro's lemma (i) implies that s_n/b_n converges to $u_\infty - u_\infty = 0$. □

(iii) Proof of the claim: since A is increasing and null at 0, we have $A_n > 0$ and $1/(1+A_n)$ is bounded. Since A is previsible, also $1/(1+A_n)$ is previsible, we can define the martingale

$$W_n = (\int (1+A)^{-1} dX)_n = \sum_{k=1}^n \frac{X_k - X_{k-1}}{1 + A_k} .$$

Moreover, since $(1 + A_n)$ is $\mathcal{A}_{n-1}$-measurable, we have

$$\mathrm{E}[(W_n - W_{n-1})^2 | \mathcal{A}_{n-1}] = (1+A_n)^{-2}(A_n - A_{n-1}) \leq (1+A_{n-1})^{-1} - (1+A_n)^{-1}$$

almost surely. This implies that $\langle W \rangle_\infty \leq 1$ so that $\lim_{n\to\infty} W_n$ exists almost surely. Kronecker's lemma (ii) applied point wise implies that on $\{A_\infty = \infty\}$

$$\lim_{n\to\infty} X_n/(1 + A_n) = \lim_{n\to\infty} X_n/A_n \to 0 .$$

□

3.6 Doob's submartingale inequality

Theorem 3.6.1 (Doob's submartingale inequality). Let X be a non-negative submartingale. For any $\epsilon > 0$

$$\epsilon \cdot \mathrm{P}[\sup_{1 \leq k \leq n} X_k \geq \epsilon] \leq \mathrm{E}[X_n; \{\sup_{1 \leq k \leq n} X_k \geq \epsilon\}] \leq \mathrm{E}[X_n] .$$

Proof. The set $A = \{\sup_{1 \leq k \leq n} X_k \geq \epsilon\}$ is a disjoint union of the sets

$$\begin{aligned} A_0 &= \{X_0 \geq \epsilon\} \in \mathcal{A}_0 \\ A_k &= \{X_k \geq \epsilon\} \cap (\bigcup_{i=0}^{n} A_i^c) \in \mathcal{A}_k . \end{aligned}$$

Since X is a submartingale, and $X_k \geq \epsilon$ on A_k we have for $k \leq n$

$$\mathrm{E}[X_n; A_k] \geq \mathrm{E}[X_k; A_k] \geq \epsilon \mathrm{P}[A_k] .$$

Summing up from $k = 0$ to n gives the result. □

We have seen the following result already as part of theorem (2.11.1). Here it appears as a special case of the submartingale inequality:

Theorem 3.6.2 (Kolmogorov's inequality). Given $X_n \in \mathcal{L}^2$ IID with $\mathrm{E}[X_i] = 0$ and $S_n = \sum_{k=1}^n X_k$. Then for $\epsilon > 0$,

$$\mathrm{P}[\sup_{1 \leq k \leq n} |S_k| \geq \epsilon] \leq \frac{\mathrm{Var}[S_n]}{\epsilon^2} .$$

Proof. S_n is a martingale with respect to $\mathcal{A}_n = \sigma(X_1, X_2, \dots, X_n)$. Because $u(x) = x^2$ is convex, S_n^2 is a submartingale. Now apply the submartingale inequality (3.6.1). □

Here is an other proof of the law of iterated logarithm for independent $N(0,1)$ random variables.

Theorem 3.6.3 (Special case of law of iterated logarithm). Given X_n IID with standard normal distribution $N(0,1)$. Then $\limsup_{n \to \infty} S_n/\Lambda(n) = 1$.

Proof. We will use for

$$1 - \Phi(x) = \int_x^\infty \phi(y)\, dy = \int_x^\infty (2\pi)^{-1/2} \exp(-y^2/2)\, dy$$

the elementary estimates

$$(x + x^{-1})^{-1}\phi(x) \leq 1 - \Phi(x) \leq x^{-1}\phi(x) .$$

(i) S_n is a martingale relative to $\mathcal{A}_n = \sigma(X_1, \dots, X_n)$. The function $x \mapsto e^{\theta x}$ is convex on $\mathbb{R}$ so that $e^{\theta S_n}$ is a submartingale. The submartingale inequality (3.6.1) gives

$$\mathrm{P}[\sup_{1 \leq k \leq n} S_k > \epsilon] = \mathrm{P}[\sup_{1 \leq k \leq n} e^{\theta S_k} > e^{\theta\epsilon}] \leq e^{-\theta\epsilon}\mathrm{E}[e^{\theta S_n}] = e^{-\theta\epsilon}e^{\theta^2 \cdot n/2} .$$

For given $\epsilon > 0$, we get the best estimate for $\theta = \epsilon/n$ and obtain

$$\mathrm{P}[\sup_{1 \leq k \leq n} S_k > \epsilon] \leq e^{-\epsilon^2/(2n)} .$$

(ii) Given $K > 1$ (close to 1). Choose $\epsilon_n = K\Lambda(K^{n-1})$. The last inequality in (i) gives

$$\mathrm{P}[\sup_{1 \leq k \leq K^n} S_k \geq \epsilon_n] \leq \exp(-\epsilon_n^2/(2K^n)) = (n-1)^{-K}(\log K)^{-K} .$$

The Borel-Cantelli lemma assures that for large enough n and $K^{n-1} \leq k \leq K^n$

$$S_k \leq \sup_{1 \leq k \leq K^n} S_k \leq \epsilon_n = K\Lambda(K^{n-1}) \leq K\Lambda(k)$$

which means for $K > 1$ almost surely

$$\limsup_{k \to \infty} \frac{S_k}{\Lambda(k)} \leq K .$$

By taking a sequence of K's converging down to 1, we obtain almost surely

$$\limsup_{k \to \infty} \frac{S_k}{\Lambda(k)} \leq 1 .$$

(iii) Given $N > 1$ (large) and $\delta > 0$ (small). Define the independent sets

$$A_n = \{S(N^{n+1}) - S(N^n) > (1-\delta)\Lambda(N^{n+1} - N^n)\} .$$

Then

$$\mathrm{P}[A_n] = 1 - \Phi(y) = (2\pi)^{-1/2}(y + y^{-1})^{-1}e^{-y^2/2}$$

with $y = (1-\delta)(2\log\log(N^{n-1} - N^n))^{1/2}$. Since $\mathrm{P}[A_n]$ is up to logarithmic terms equal to $(n \log N)^{-(1-\delta)^2}$, we have $\sum_n \mathrm{P}[A_n] = \infty$. Borel-Cantelli shows that $\mathrm{P}[\limsup_n A_n] = 1$ so that

$$S(N^{n+1}) > (1-\delta)\Lambda(N^{n+1} - N^n) + S(N^n) .$$

By (ii), $S(N^n) > -2\Lambda(N^n)$ for large n so that for infinitely many n, we have

$$S(N^{n+1}) > (1-\delta)\Lambda(N^{n+1}-N^n) - 2\Lambda(N^n) .$$

It follows that

$$\limsup_n \frac{S_n}{\Lambda_n} \geq \limsup_n \frac{S(N^{n+1})}{\Lambda(N^{n+1})} \geq (1-\delta)(1-\frac{1}{N})^{1/2} - 2N^{-1/2} .$$

□

3.7 Doob's $\mathcal{L}^p$ inequality

Lemma 3.7.1. (Corollary of Hölder inequality) Fix $p > 1$ and q satisfying $p^{-1}+q^{-1} = 1$. Given $X, Y \in \mathcal{L}^p$ satisfying

$$\epsilon \mathrm{P}[|X| \geq \epsilon] \leq \mathrm{E}[|Y|; |X| \geq \epsilon]$$

$\forall \epsilon > 0$, then $||X||_p \leq q \cdot ||Y||_p$.

Proof. Integrating the assumption multiplied with $p\epsilon^{p-2}$ gives

$$L = \int_0^\infty p\epsilon^{p-1}\mathrm{P}[|X| \geq \epsilon] \; d\epsilon \leq \int_0^\infty p\epsilon^{p-2}\mathrm{E}[|Y|; |X| \geq \epsilon] \; d\epsilon =: R .$$

By Fubini's theorem, the the left hand side is

$$L = \int_0^\infty \mathrm{E}[p\epsilon^{p-1}1_{\{|X| \geq \epsilon\}}] \; d\epsilon = \mathrm{E}[\int_0^\infty p\epsilon^{p-1}1_{\{|X| \geq \epsilon\}} d\epsilon] = \mathrm{E}[|X|^p] .$$

Similarly, the right hand side is $R = \mathrm{E}[q \cdot |X|^{p-1}|Y|]$. With Hölder's inequality, we get

$$\mathrm{E}[|X|^p] \leq \mathrm{E}[q|X|^{p-1}|Y|] \leq q||Y||_p \cdot |||X|^{p-1}||_q .$$

Since $(p-1)q = p$, we can substitute $|||X|^{p-1}||_q = \mathrm{E}[|X|^p]^{1/q}$ on the right hand side, which gives the claim. □

Theorem 3.7.2 (Doob's L^p inequality). Given a non-negative submartingale X which is bounded in $\mathcal{L}^p$. Then $X^* = \sup_n X_n$ is in $\mathcal{L}^p$ and satisfies

$$||X^*|| \leq q \cdot \sup_n ||X_n||_p .$$

Proof. Define $X_n^* = \sup_{1 \leq k \leq n} X_k$ for $n \in \mathbb{N}$. From Doob's submartingale inequality (3.6.1) and the above lemma (3.7.1), we see that

$$||X_n^*||_p \leq q||X_n||_p \leq q \sup_n ||X_n||_p \, .$$

□

Corollary 3.7.3. Given a non-negative submartingale X which is bounded in $\mathcal{L}^p$. Then $X_\infty = \lim_{n\to\infty} X_n$ exists in $\mathcal{L}_p$ and $||X_\infty||_p = \sup_n ||X_n||_p$.

Proof. The submartingale X is dominated by the element X^* in the $\mathcal{L}^p$-inequality. The supermartingale $-X$ is bounded in $\mathcal{L}^p$ and so bounded in $\mathcal{L}^1$. We know therefore that $X_\infty = \lim_{n\to\infty} X_n$ exists almost everywhere. From $|X_n - X_\infty|^p \leq (2X^*)^p \in \mathcal{L}^p$ and the dominated convergence theorem we deduce $X_n \to X_\infty$ in $\mathcal{L}^p$. □

Corollary 3.7.4. Given a martingale Y bounded in $\mathcal{L}^p$ and $X = |Y|$. Then

$$X_\infty = \lim_{n\to\infty} X_n$$

exists in $\mathcal{L}_p$ and $||X_\infty||_p = \sup_n ||X_n||_p$.

Proof. Use the above corollary for the submartingale $X = |Y|$. □

Theorem 3.7.5 (Kakutani's theorem). Let X_n be a non-negative independent $\mathcal{L}^1$ process with $\mathrm{E}[X_n] = 1$ for all n. Define $S_0 = 1$ and $S_n = \prod_{k=1}^n X_k$. Then $S_\infty = \lim_n S_n$ exists, because S_n is a nonnegative $\mathcal{L}^1$ martingale. Then S_n is uniformly integrable if and only if $\prod_{n=1}^\infty \mathrm{E}[X_n^{1/2}] > 0$.

Proof. Define $a_n = \mathrm{E}[X_n^{1/2}]$. The process

$$T_n = \frac{X_1^{1/2}}{a_1} \frac{X_2^{1/2}}{a_2} \cdots \frac{X_n^{1/2}}{a_n}$$

is a martingale. We have $\mathrm{E}[T_n^2] = (a_1 a_2 \cdots a_n)^{-2} \leq (\prod_n a_n)^{-1} < \infty$ so that T is bounded in $\mathcal{L}^2$, By Doob's $\mathcal{L}^2$-inequality

$$\mathrm{E}[\sup_n |S_n|] \leq \mathrm{E}[\sup_n |T_n|^2] \leq 4 \sup_n \mathrm{E}[|T_n|^2] < \infty$$

so that S is dominated by $S^* = \sup_n |S_n| \in \mathcal{L}^1$. This implies that S is uniformly integrable.

If S_n is uniformly integrable, then $S_n \to S_\infty$ in $\mathcal{L}^1$. We have to show that $\prod_{n=1}^\infty a_n > 0$. Aiming to a contradiction, we assume that $\prod_n a_n = 0$. The martingale T defined above is a nonnegative martingale which has a limit T_∞. But since $\prod_n a_n = 0$ we must then have that $S_\infty = 0$ and so $S_n \to 0$ in $\mathcal{L}^1$. This is not possible because $\mathrm{E}[S_n] = 1$ by the independence of the X_n. □

Here are examples, where martingales occur in applications:

Example. This example is a primitive model for the Stock and Bond market. Given $a < r < b < \infty$ real numbers. Define $p = (r - a)/(b - a)$. Let ϵ_n be IID random variables taking values $1, -1$ with probability p respectively $1 - p$. Define a process B_n (**bonds** with fixed interest rate f) and S_n (**stocks** with fluctuating interest rates) by

$$B_n = (1 + r)^n B_{n-1}, B_0 = 1$$

$$S_n = (1 + R_n) S_{n-1}, S_0 = 1$$

with $R_n = (a + b)/2 + \epsilon_n (a - b)/2$. Given a sequence A_n (the **portfolio**), your fortune is X_n and satisfies

$$X_n = (1 + r) X_{n-1} + A_n S_{n-1} (R_n - r) .$$

We can write $R_n - r = \frac{1}{2}(b - a)(Z_n - Z_{n-1})$ with the martingale

$$Z_n = \sum_{k=1}^n (\epsilon_k - 2p + 1) .$$

The process $Y_n = (1 + r)^{-n} X_n$ satisfies then

$$\begin{aligned} Y_n - Y_{n-1} &= (1 + r)^{-n} A_n S_{n-1} (R_n - r) \\ &= \frac{1}{2}(b - a)(1 + r)^{-n} A_n S_{n-1} (Z_n - Z_{n-1}) \\ &= C_n (Z_n - Z_{n-1}) \end{aligned}$$

showing us that Y is the stochastic integral $\int C \, dZ$. So, if your portfolio A_n is previsible ($\mathcal{A}_{n-1}$ measurable), then Y is a martingale.

Example. Let X, $X_1, X_2 \dots$ be independent random variables satisfying that the law of X is $N(0, \sigma^2)$ and the law of X_k is $N(0, \sigma_k^2)$. We define the random variables

$$Y_k = X + X_k$$

which we consider as a **noisy observation of the random variable** X. Define $\mathcal{A}_n = \sigma(X_1, \dots, X_n)$ and the martingale

$$M_n = \mathrm{E}[X|\mathcal{A}_n] .$$

By Doob's martingale convergence theorem (3.5.4), we know that M_n converges in $\mathcal{L}^2$ to a random variable M_∞. One can show that

$$\mathrm{E}[(X - M_n)^2] = (\sigma^{-2} + \sum_{k=1}^{n} \sigma_k^{-2})^{-1} .$$

This implies that $X = M_\infty$ if and only if $\sum_n \sigma_n^{-2} = \infty$. If the noise grows too much, for example for $\sigma_n = n$, then we can not recover X from the observations Y_k.

3.8 Random walks

Consider the d-dimensional lattice $\mathbb{Z}^d$, where each point has $2d$ neighbors. A particle starts at the origin $0 \in \mathbb{Z}^d$ and makes in each time step a random step into one of the $2d$ directions. What is the probability that the particle returns back to the origin?

Definition. Define a sequence of IID random variables X_n which take values in

$$I = \{e \in \mathbb{Z}^d \mid |e| = \sum_{i=1}^{d} |e_i| = 1 \}$$

and which have the uniform distribution defined by $\mathrm{P}[X_n = e] = (2d)^{-1}$ for all $e \in I$. The random variable $S_n = \sum_{i=1}^{n} X_i$ with $S_0 = 0$ describes the position of the particle at time n. The discrete stochastic process S_n is called the **random walk** on the lattice $\mathbb{Z}^d$.

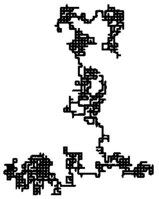


Figure. *A random walk sample path $S_1(\omega), \dots, S_n(\omega)$ in the lattice $\mathbb{Z}^2$ after 2000 steps. $B_n(\omega)$ is the number of revisits of the starting points* 0.

As a probability space, we can take $\Omega = I^{\mathbb{N}}$ with product measure $\nu^{\mathbb{N}}$, where ν is the measure on E, which assigns to each point e the probability

$\nu(\{e\}) = (2d)^{-1}$. The random variables X_n are then defined by $X_n(\omega) = \omega_n$. Define the sets $A_n = \{S_n = 0\}$ and the random variables

$$Y_n = 1_{A_n} \cdot$$

if the particle has returned to position $0 \in \mathbb{Z}^d$ at time n, then $Y_n = 1$, otherwise $Y_n = 0$. The sum $B_n = \sum_{k=0}^{n} Y_k$ counts the number of visits of the origin 0 of the particle up to time n and $B = \sum_{k=0}^{\infty} Y_k$ counts the total number of visits at the origin. The expectation

$$\mathrm{E}[B] = \sum_{n=0}^{\infty} \mathrm{P}[S_n = 0]$$

tells us how many times a particle is expected to return to the origin. We write $\mathrm{E}[B] = \infty$, if the sum diverges. In this case, the particle returns back to the origin infinitely many times.

Theorem 3.8.1 (Polya). $\mathrm{E}[B] = \infty$ for $d = 1, 2$ and $\mathrm{E}[B] < \infty$ for $d > 2$.

Proof. Fix $n \in \mathbb{N}$ and define $a^{(n)}(k) = \mathrm{P}[S_n = k]$ for $k \in \mathbb{Z}^d$. Because the particle can reach in time n only a bounded region, the function $a^{(n)} : \mathbb{Z}^d \to \mathbb{R}$ is zero outside a bounded set. We can therefore define its Fourier transform

$$\phi_{S_n}(x) = \sum_{k \in \mathbb{Z}^d} a^{(n)}(k) e^{2\pi i k \cdot x}$$

which is smooth function on $\mathbb{T}^d = \mathbb{R}^d/\mathbb{Z}^d$. It is the characteristic function of S_n because

$$\mathrm{E}[e^{ixS_n}] = \sum_{k \in \mathbb{Z}^d} \mathrm{P}[S_n = k] e^{ik \cdot x} .$$

The characteristic function ϕ_X of X_k is

$$\phi_X(x) = \frac{1}{2d} \sum_{|j|=1} e^{2\pi i x_j} = \frac{1}{d} \sum_{i=1}^{d} \cos(2\pi x_i) .$$

Because the S_n is a sum of n independent random variables X_j

$$\phi_{S_n} = \phi_{X_1}(x)\phi_{X_2}(x)\dots\phi_{X_n}(x) = \frac{1}{d^n}(\sum_{i=1}^{d} \cos(2\pi x_i))^n .$$

Note that $\phi_{S_n}(0) = \mathrm{P}[S_n = 0]$.

We now show that $\mathrm{E}[B] = \sum_{n \geq 0} \phi_{S_n}(0)$ is finite if and only if $d < 3$. The Fourier inversion formula gives

$$\sum_n \mathrm{P}[S_n = 0] = \int_{\mathbb{T}^d} \sum_{n=0}^{\infty} \phi_X^n(x)\, dx = \int_{\mathbb{T}^d} \frac{1}{1 - \phi_X(x)}\, dx .$$

A Taylor expansion $\phi_X(x) = 1 - \sum_j \frac{x_j^2}{2}(2\pi)^2 + \dots$ shows

$$\frac{1}{2} \cdot \frac{(2\pi)^2}{2d}|x|^2 \leq 1 - \phi_X(x) \leq 2 \cdot \frac{(2\pi)^2}{2d}|x|^2 .$$

The claim of the theorem follows because the integral

$$\int_{\{|x|<\epsilon\}} \frac{1}{|x|^2} \, dx$$

over the ball of radius ϵ in $\mathbb{R}^d$ is finite if and only if $d \geq 3$. □

Corollary 3.8.2. The particle returns to the origin infinitely often almost surely if $d \leq 2$. For $d \geq 3$, almost surely, the particle returns only finitely many times to zero and $\mathrm{P}[\lim_{n\to\infty} |S_n| = \infty] = 1$.

Proof. If $d > 2$, then $A_\infty = \limsup_n A_n$ is the subset of Ω, for which the particles returns to 0 infinitely many times. Since $\mathrm{E}[B] = \sum_{n=0}^\infty \mathrm{P}[A_n]$, the Borel-Cantelli lemma gives $\mathrm{P}[A_\infty] = 0$ for $d > 2$. The particle returns therefore back to 0 only finitely many times and in the same way it visits each lattice point only finitely many times. This means that the particle eventually leaves every bounded set and converges to infinity.
If $d \leq 2$, let p be the probability that the random walk returns to 0:

$$p = \mathrm{P}[\bigcup_n A_n] .$$

Then p^{m-1} is the probability that there are at least m visits in 0 and the probability is $p^{m-1} - p^m = p^{m-1}(1-p)$ that there are exactly m visits. We can write

$$\mathrm{E}[B] = \sum_{m\geq 1} m p^{m-1}(1-p) = \frac{1}{1-p} .$$

Because $\mathrm{E}[B] = \infty$, we know that $p = 1$. □

The use of characteristic functions allows also to solve combinatorial problems like to count the number of closed paths starting at zero in the graph:

Proposition 3.8.3. There are

$$\int_{\mathbb{T}^d} (\sum_{k=1}^d \cos(2\pi i x_k))^n \, dx_1 \cdots dx_d$$

closed paths of length n in the lattice $\mathbb{Z}^d$.

Proof. If we know the probability $\mathrm{P}[S_n = 0]$ that a path returns to 0 in n step, then $(2d)^n \mathrm{P}[S_n = 0]$ is the number of closed paths in $\mathbb{Z}^d$ of length n. But $\mathrm{P}[S_n = 0]$ is the zeroth Fourier coefficient

$$\int_{\mathbb{T}^d} \phi_{S_n}(x) \, dx = \int_{\mathbb{T}^d} \frac{1}{d^n} (\sum_{k=1}^{d} \cos(2\pi i x_k))^n \, dx$$

of ϕ_{S_n}. □

Example. In the case $d = 1$, we have

$$\int_0^1 2^{2n} \cos^{2n}(2\pi x) \, dx = \left(\begin{array}{c} 2n \\ n \end{array} \right)$$

closed paths of length $2n$ starting at 0. We know that also because

$$\mathrm{P}[S_{2n} = 0] = \left(\begin{array}{c} 2n \\ n \end{array} \right) \frac{1}{2^n} \frac{1}{2^n} .$$

The lattice $\mathbb{Z}^d$ can be generalized to an arbitrary graph G which is a **regular graph** that is a graph, where each vertex has the same number of neighbors. A convenient way is to take as the graph the Cayley graph of a discrete group G with generators $a_1, \dots, a_d$.

Corollary 3.8.4. If G is the Cayley graph of an Abelian group $\mathcal{G}$ then the random walk on G is recurrent if and only at most two of the generators have infinite order.

Proof. An Abelian group $\mathcal{G}$ is isomorphic to $\mathbb{Z}^k \times \mathbb{Z}_{n_1} \times \dots \mathbb{Z}_{n_d}$. The characteristic function of X_n is a function on the dual group $\hat{\mathcal{G}}$

$$\sum_{n=0}^{\infty} P[S_n = 0] = \sum_{n=0}^{\infty} \int_{\hat{\mathcal{G}}} \phi_{S_n}(x) \, dx = \sum_{n=0}^{\infty} \int_{\hat{\mathcal{G}}} \phi_X^n(x) \, dx = \int_{\hat{\mathcal{G}}} \frac{1}{1 - \phi_X(x)} \, dx$$

is finite if and only if $\hat{\mathcal{G}}$ contains a three dimensional torus which means $k > 2$. □

The recurrence properties on non-Abelian groups is more subtle, because characteristic functions loose then some of their good properties.

Example. An other generalization is to add a **drift** by changing the probability distribution ν on I. Given $p_j \in (0,1)$ with $\sum_{|j|=1} p_j = 1$. In this case

$$\phi_X(x) = \sum_{|j|=1} p_j e^{2\pi i x_j} .$$

We have recurrence if and only if

$$\int_{\mathbb{T}^d} \frac{1}{1-\phi_X(x)} \, dx = \infty .$$

Take for example the case $d = 1$ with drift parameterized by $p \in (0,1)$. Then

$$\phi_X(x) = pe^{2\pi i x} + (1-p)e^{-2\pi i x} = \cos(2\pi x) + i(2p-1)\sin(2\pi x) .$$

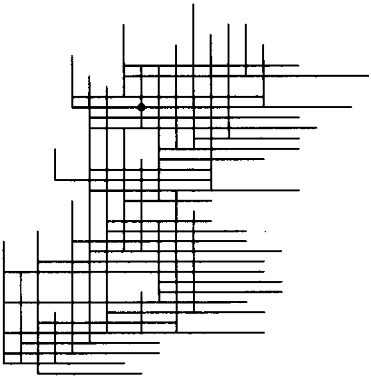
which shows that

$$\int_{\mathbb{T}^d} \frac{1}{1-\phi_X(x)} \, dx < \infty$$

if $p \neq 1/2$. A random walk with drift will almost certainly not return to 0 infinitely often.

Example. An other generalization of the random walk is to take identically distributed random variables X_n with values in I, which need not to be independent. An example which appears in number theory in the case $d = 1$ is to take the probability space $\Omega = \mathbb{T}^1 = \mathbb{R}/\mathbb{Z}$, an irrational number α and a function f which takes each value in I on an interval $[\frac{k}{2d}, \frac{k+1}{2d})$. The random variables $X_n(\omega) = f(\omega + n\alpha)$ define an ergodic discrete stochastic process but the random variables are not independent. A random walk $S_n = \sum_{k=1}^n X_k$ with random variables X_k which are dependent is called a **dependent random walk**.

Figure. *If Y_k are IID random variables with uniform distribution in $[0, a]$, then $Z_n = \sum_{k=1}^n Y_k \bmod 1$ are dependent. Define $X_k = (1,0)$ if $Z_k \in [0,1/4)$, $X_k = (-1,0)$ if $Z_k \in [1/4,1/2)$, $X_k = (0,1)$ if $Z_k \in [1/2,3/4)$ and $X_k = (0,-1)$ if $Z_k \in [3/4,1)$. Also X_k are no more independent. For small a, there can belong intervals, where X_k is the same because Z_k stays in the same quarter interval. The picture shows a typical path of the process $S_n = \sum_{k=1}^n X_k$.*



Example. An example of a one-dimensional dependent random walk is the problem of "**almost alternating sums**" [52]. Define on the probability space $\Omega = ([0,1], \mathcal{A}, dx)$ the random variables $X_n(x) = 21_{[0,1/2]}(x + n\alpha) - 1$, where α is an irrational number. This produces a symmetric random walk, but unlike for the usual random walk, where $S_n(x)$ grows like $\sqrt{n}$, one sees a much slower growth $S_n(0) \leq \log(n)^2$ for almost all α and for special numbers like the **golden ratio** $(\sqrt{5}+1)/2$ or the **silver ratio** $\sqrt{2}+1$ one has for infinitely many n the relation

$$a \cdot \log(n) + 0.78 \leq S_n(0) \leq a \cdot \log(n) + 1$$

with $a = 1/(2\log(1+\sqrt{2}))$. It is not known whether $S_n(0)$ grows like $\log(n)$ for almost all α.

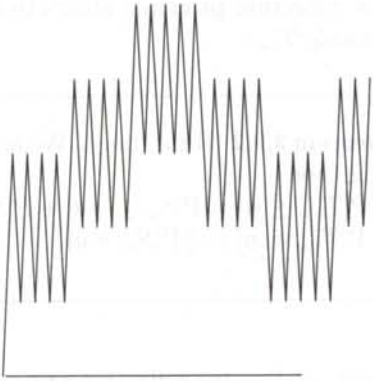


Figure. *An almost periodic random walk in one dimensions. Instead of flipping coins to decide whether to go up or down, one turns a wheel by an angle α after each step and goes up if the wheel position is in the right half and goes down if the wheel position is in the left half. While for periodic α the growth of S_n is either linear (like for $\alpha = 0$), or zero (like for $\alpha = 1/2$), the growth for most irrational α seems to be logarithmic.*

3.9 The arc-sin law for the 1D random walk

Definition. Let X_n denote the $\{-1,1\}$-valued random variable with $\mathrm{P}[S_n = \pm 1] = 1/2$ and let $S_n = \sum_{k=1}^n X_k$ be the random walk. We have seen that it is a martingale with respect to X_n. Given $a \in \mathbb{Z}$, we define the stopping time

$$T_a = \min\{n \in \mathbb{N} \mid S_n = a\} .$$

Theorem 3.9.1 (Reflection principle). For integers $a, b > 0$, one has

$$\mathrm{P}[a + S_n = b \, , \, T_{-a} \leq n] = \mathrm{P}[S_n = a + b] .$$

Proof. The number of paths from a to b passing zero is equal to the number of paths from $-a$ to b which in turn is the number of paths from zero to $a + b$. □

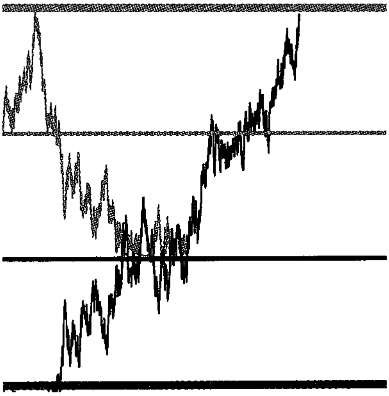


Figure. *The proof of the reflection principle: reflect the part of the path above 0 at the line 0. To every path which goes from a to b and touches 0 there corresponds a path from $-a$ to b.*

The reflection principle allows to compute the distribution of the random variable T_{-a}:

Theorem 3.9.2 (Ruin time). We have the following distribution of the stopping time:
a) $\mathrm{P}[T_{-a} \leq n] = \mathrm{P}[S_n \leq -a] + P[S_n > a]$.
b) $\mathrm{P}[T_{-a} = n] = \frac{a}{n}\mathrm{P}[S_n = a]$.

Proof. a) Use the reflection principle in the third equality:

$$\begin{aligned}
\mathrm{P}[T_{-a} \leq n] &= \sum_{b \in \mathbb{Z}} \mathrm{P}[T_{-a} \leq n, a + S_n = b] \\
&= \sum_{b \leq 0} \mathrm{P}[a + S_n = b] + \sum_{b > 0} \mathrm{P}[T_{-a} \leq n, a + S_n = b] \\
&= \sum_{b \leq 0} \mathrm{P}[a + S_n = b] + \sum_{b > 0} \mathrm{P}[S_n = a + b] \\
&= \mathrm{P}[S_n \leq -a] + \mathrm{P}[S_n > a]
\end{aligned}$$

b) From

$$\mathrm{P}[S_n = a] = \left(\begin{array}{c} n \\ \frac{a+n}{2} \end{array} \right)$$

we get

$$\frac{a}{n}\mathrm{P}[S_n = a] = \frac{1}{2}(\mathrm{P}[S_{n-1} = a - 1] - \mathrm{P}[S_{n-1} = a + 1]) \; .$$

Also

$$\begin{aligned}
\mathrm{P}[S_n > a] - \mathrm{P}[S_{n-1} > a] &= \mathrm{P}[S_n > a \; , \; S_{n-1} \leq a] \\
&\quad + \mathrm{P}[S_n > a \; , \; S_{n-1} > a] - \mathrm{P}[S_{n-1} > a] \\
&= \frac{1}{2}(\mathrm{P}[S_{n-1} = a] - \mathrm{P}[S_{n-1} = a + 1])
\end{aligned}$$

and analogously

$$P[S_n \leq -a] - P[S_{n-1} \leq -a] = \frac{1}{2}(P[S_{n-1} = a-1] - P[S_{n-1} = a]) .$$

Therefore, using a)

$$\begin{array}{rcl} P[T_{-a} = n] & = & P[T_{-a} \leq n] - P[T_{-a} \leq n-1] \\ & = & P[S_n \leq -a] - P[S_{n-1} \leq -a] \\ & & P[S_n > a] - P[S_{n-1} > a] \\ & = & \frac{1}{2}(P[S_{n-1} = a] - P[S_{n-1} = a+1]) \\ & & \frac{1}{2}(P[S_{n-1} = a-1] - P[S_{n-1} = a]) \\ & = & \frac{1}{2}(P[S_{n-1} = a-1] - P[S_{n-1} = a+1]) = \frac{a}{n}P[S_n = a] \end{array}$$

□

Theorem 3.9.3 (Ballot theorem).

$$P[S_n = a \, , \, S_1 > 0, \dots, S_{n-1} > 0] = \frac{a}{n} \cdot P[S_n = a] .$$

Proof. When reversing time, the number of paths from 0 to a of length n which do no more hit 0 is the number of paths of length n which start in a and for which $T_{-a} = n$. Now use the previous theorem

$$P[T_{-a} = n] = \frac{a}{n}P[S_n = a] .$$

□

Corollary 3.9.4. The distribution of the first return time is

$$P[T_0 > 2n] = P[S_{2n} = 0] .$$

Proof.

$$\begin{array}{rcll} P[T_0 > 2n] & = & \frac{1}{2}P[T_{-1} > 2n-1] + \frac{1}{2}P[T_1 > 2n-1] & \\ & = & P[T_{-1} > 2n-1] & \text{(by symmetry)} \\ & = & P[S_{2n-1} > -1 \text{ and } S_{2n-1} \leq 1] & \\ & = & P[S_{2n-1} \in \{0,1\}] & \\ & = & P[S_{2n-1} = 1] = P[S_{2n} = 0] . & \end{array}$$

□

Remark. We see that $\lim_{n\to\infty} \mathrm{P}[T_0 > 2n] = 0$. This restates that the random walk is recurrent. However, the expected return time is very long:

$$\mathrm{E}[T_0] = \sum_{n=0}^{\infty} n\mathrm{P}[T_0 = n] = \sum_{n=0}^{\infty} \mathrm{P}[T_0 > n] = \sum_{n=0}^{\infty} \mathrm{P}[S_n = 0] = \infty$$

because by the **Stirling formula** $n! \sim n^n e^{-n}\sqrt{2\pi n}$, one has $\left(\begin{array}{c} 2n \\ n \end{array}\right) \sim 2^{2n}/\sqrt{\pi n}$ and so

$$\mathrm{P}[S_{2n} = 0] = \left(\begin{array}{c} 2n \\ n \end{array}\right) \frac{1}{2^{2n}} \sim (\pi n)^{-1/2} .$$

Definition. We are interested now in the random variable

$$L(\omega) = \max\{0 \leq n \leq 2N \mid S_n(\omega) = 0\}$$

which describes the **last visit of the random walk in 0 before time** $2N$. If the random walk describes a game between two players, who play over a time $2N$, then L is the time when one of the two players does no more give up his leadership.

Theorem 3.9.5 (Arc Sin law). L has the discrete arc-sin distribution:

$$\mathrm{P}[L = 2n] = \frac{1}{2^{2N}} \left(\begin{array}{c} 2n \\ n \end{array}\right) \left(\begin{array}{c} 2N-2n \\ N-n \end{array}\right)$$

and for $N \to \infty$, we have

$$\mathrm{P}[\frac{L}{2N} \leq z] \to \frac{2}{\pi} \arcsin(\sqrt{z}) .$$

Proof.

$$\mathrm{P}[L = 2n] = \mathrm{P}[S_{2n} = 0] \cdot \mathrm{P}[T_0 > 2N - 2n] = \mathrm{P}[S_{2n} = 0] \cdot \mathrm{P}[S_{2N-2n} = 0]$$

which gives the first formula. The Stirling formula gives $\mathrm{P}[S_{2k} = 0] \sim \frac{1}{\sqrt{\pi k}}$ so that

$$\mathrm{P}[L = 2k] = \frac{1}{\pi} \frac{1}{\sqrt{k(N-k)}} = \frac{1}{N} f(\frac{k}{N})$$

with

$$f(x) = \frac{1}{\pi\sqrt{x(1-x)}} .$$

It follows that

$$\mathrm{P}[\frac{L}{2N} \leq z] \to \int_0^z f(x)\, dx = \frac{2}{\pi} \arcsin(\sqrt{z}) \, .$$

□

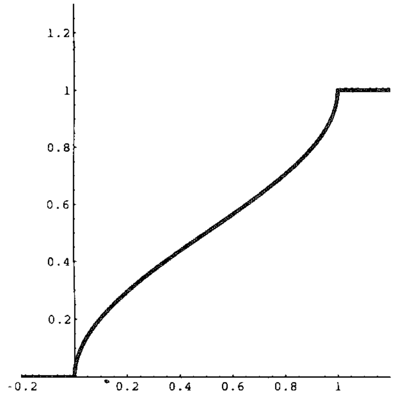


Figure. *The distribution function* $\mathrm{P}[L/2N \leq z]$ *converges in the limit* $N \to \infty$ *to the function* $2\arcsin(\sqrt{z})/\pi$.

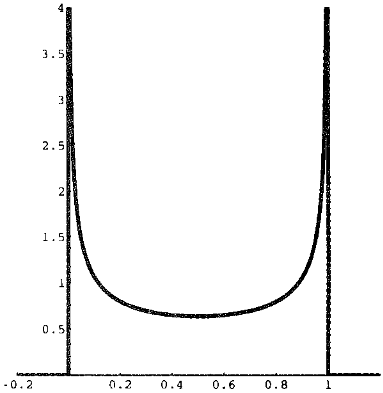


Figure. *The density function of this distribution in the limit* $N \to \infty$ *is called the arc-sin distribution.*

Remark. From the shape of the arc-sin distribution, one has to expect that the winner takes the final leading position either early or late.

Remark. The arc-sin distribution is a natural distribution on the interval $[0, 1]$ from the different points of view. It belongs to a measure which is the **Gibbs measure** of the quadratic map $x \mapsto 4 \cdot x(1-x)$ on the unit interval maximizing the Boltzmann-Gibbs entropy. It is a **thermodynamic equilibrium measure** for this quadratic map. It is the measure μ on the interval $[0, 1]$ which minimizes the energy

$$I(\mu) = -\int_0^1 \int_0^1 \log|E - E'| \, d\mu(E) \, d\mu(E') \, .$$

One calls such measures also **potential theoretical equilibrium measures**.

3.10 The random walk on the free group

Definition. The **free group** F_d with d generators is the set of finite words w written in the $2d$ letters

$$A = \{a_1, a_2, \dots, a_d, a_1^{-1}, a_2^{-1}, \dots, a_d^{-1}\}$$

modulo the identifications $a_i a_i^{-1} = a_i^{-1} a_i = 1$. The group operation is concatenating words $v \circ w = vw$. The inverse of $w = w_1 w_2 \cdots w_n$ is $w^{-1} = w_n^{-1} \cdots w_2^{-1} w_1^{-1}$. Elements w in the group F_d can be uniquely represented by reduced words obtained by deleting all words vv^{-1} in w. The identity e in the group F_d is the empty word. We denote by $l(w)$ the length of the reduced word of w.

Definition. Given a free group G with generators A and let X_k be uniformly distributed random variables with values in A. The stochastic process $S_n = X_1 \cdots X_n$ is called the **random walk on the group** G. Note that the group operation $X_k \cdot$ needs not to be commutative. The random walk on the free group can be interpreted as a walk on a **tree**, because the Cayley graph of the group F_d with generators A contains no non-contractible closed circles.

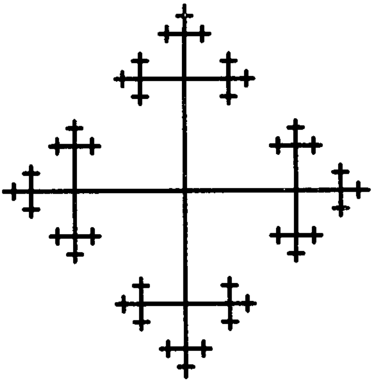


Figure. *Part of the Cayley graph of the free group F_2 with two generators a, b. It is a tree. At every point, one can go into 4 different directions. Going into one of these directions corresponds to multiplying with a, a^{-1}, b or b^{-1}.*

Definition. Define for $n \in \mathbb{N}$

$$r_n = \mathrm{P}[S_n = e\ ,\ S_1 \neq e, S_2 \neq e, \dots S_{n-1} \neq e]$$

which is the probability of returning for the first time to e if one starts at e. Define also for $n \in \mathbb{N}$

$$m_n = \mathrm{P}[S_n = e]$$

with the convention $m^{(0)} = 1$. Let r and m be the probability generating functions of the sequences r_n and m_n:

$$m(x) = \sum_{n=0}^{\infty} m_n x^n,\ r(x) = \sum_{n=0}^{\infty} r_n x^n\ .$$

These sums converge for $|x| < 1$.

Lemma 3.10.1. (Feller)

$$m(x) = \frac{1}{1 - r(x)}\ .$$

Proof. Let T be the stopping time

$$T = \min\{n \in \mathbb{N} \mid S_n = e\} .$$

With $\mathrm{P}[T = n] = r_n$, the function $r(x) = \sum_n r_n x^n$ is the probability generating function of T. The probability generating function of a sum independent random variables is the product of the probability generating functions. Therefore, if T_i are independent random variables with distribution T, then $\sum_{i=1}^{n} T_i$ has the probability generating function $x \mapsto r^n(x)$. We have

$$\begin{aligned}
\sum_{n=0}^{\infty} m_n x^n &= \sum_{n=0}^{\infty} \mathrm{P}[S_n = e] x^n \\
&= \sum_{n=0}^{\infty} \sum_{0 \leq n_1 < n_2 < \cdots < n_k} \mathrm{P}[S_{n_1} = e, S_{n_2} = e, \dots, S_{n_k} = e, \\
&\quad S_n \neq e \text{ for } n \notin \{n_1, \dots, n_k\}] x^n \\
&= \sum_{n=0}^{\infty} \mathrm{P}[\sum_{k=1}^{n} T_k = n] x^n = \sum_{k=0}^{\infty} r^n(x) = \frac{1}{1 - r(x)} .
\end{aligned}$$

□

Remark. This lemma is true for the random walk on a Cayley graph of any **finitely presented group.**

The numbers r_{2n+1} are zero for odd $2n+1$ because an even number of steps are needed to come back. The values of r_{2n} can be computed by using basic combinatorics:

Lemma 3.10.2. (Kesten)

$$r_{2n} = \frac{1}{(2d)^n} \frac{1}{n} \binom{2n-2}{n-1} 2d(2d-1)^{n-1} .$$

Proof. We have

$$r_{2n} = \frac{1}{2d} |\{w_1 w_2 \dots w_n \in G, w^k = w_1 w_2 \dots w_k \neq e\}| .$$

To count the number of such words, map every word with $2n$ letters into a path in $\mathbb{Z}^2$ going from $(0,0)$ to (n,n) which is away from the diagonal except at the beginning or the end. The map is constructed in the following way: for every letter, we record a horizontal or vertical step of length 1. If $l(w^k) = l(w^{k-1}) + 1$, we record a horizontal step. In the other case, if $l(w^k) = l(w^{k-1}) - 1$, we record a vertical step. The first step is horizontal independent of the word. There are

$$\frac{1}{n} \binom{2n-2}{n-1}$$

such paths since by the distribution of the stopping time in the one dimensional random walk

$$\begin{aligned} \mathrm{P}[T_{-1} = 2n-1] &= \frac{1}{2n-1} \cdot \mathrm{P}[S_{2n-1} = 1] \\ &= \frac{1}{2n-1} \binom{2n-1}{n} \\ &= \frac{1}{n} \binom{2n-2}{n-1} . \end{aligned}$$

Counting the number of words mapped into the same path, we see that we have in the first step $2d$ possibilities and later $(2d-1)$ possibilities in each of the $n-1$ horizontal step and only 1 possibility in a vertical step. We have therefore to multiply the number of paths by $2d(2d-1)^{n-1}$. □

Theorem 3.10.3 (Kesten). For the free group F_d, we have

$$m(x) = \frac{2d-1}{(d-1) + \sqrt{d^2 - (2d-1)x^2}} .$$

Proof. Since we know r^{2n} we can compute

$$r(x) = \frac{d - \sqrt{d^2 - (2d-1)x^2}}{2d-1}$$

and get the claim with Feller's lemma $m(x) = 1/(1 - r(x))$. □

Remark. The Cayley graph of the free group is also called the **Bethe lattice.** One can read of from this formula that the spectrum of the free Laplacian $L : l^2(F_d) \to l^2(F_d)$ on the **Bethe lattice** given by

$$Lu(g) = \sum_{a \in A} u(g+a)$$

is the whole interval $[-a, a]$ with $a = 2\sqrt{2d-1}$.

Corollary 3.10.4. The random walk on the free group F_d with d generators is recurrent if and only if $d = 1$.

Proof. Denote as in the case of the random walk on $\mathbb{Z}^d$ with B the random variable counting the total number of visits of the origin. We have then again $\mathrm{E}[B] = \sum_n \mathrm{P}[S_n = e] = \sum_n m_n = m(1)$. We see that for $d = 1$ we

have $m(1) = \infty$ and that $m(d) < \infty$ for $d > 1$. This establishes the analog of Polya's result on $\mathbb{Z}^d$ and leads in the same way to the recurrence:
(i) $d = 1$: We know that $\mathbb{Z}_1 = F_1$, and that the walk in $\mathbb{Z}^1$ is recurrent.
(ii) $d \geq 2$: define the event $A_n = \{S_n = e\}$. Then $A_\infty = \limsup_n A_n$ is the subset of Ω, for which the walk returns to e infinitely many times. Since for $d \geq 2$,

$$\mathrm{E}[B] = \sum_{n=0}^{\infty} \mathrm{P}[A_n] m(d) < \infty ,$$

The Borel-Cantelli lemma gives $\mathrm{P}[A_\infty] = 0$ for $d > 2$. The particle returns therefore to 0 only finitely many times and similarly it visits each vertex in F_d only finitely many times. This means that the particle eventually leaves every bounded set and escapes to infinity. □

Remark. We could say that the problem of the random walk on a discrete group G is **solvable** if one can give an algebraic formula for the function $m(x)$. We have seen that the classes of Abelian finitely generated and free groups are solvable. Trying to extend the class of solvable random walks seems to be an interesting problem. It would also be interesting to know, whether there exists a group such that the function $m(x)$ is transcendental.

3.11 The free Laplacian on a discrete group

Definition. Let G be a countable discrete group and $A \subset G$ a finite set which generates G. The **Cayley graph** Γ of (G, A) is the graph with edges G and sites (i, j) satisfying $i - j \in A$ or $j - i \in A$.

Remark. We write the composition in G additively even so we do not assume that G is Abelian. We allow A to contain also the identity $e \in G$. In this case, the Cayley graph contains two closed loops of length 1 at each site.

Definition. The **symmetric random walk** on $\Gamma(G, A)$ is the process obtained by summing up independent uniformly distributed $(A \cup A^{-1})$-valued random variables X_n. More generally, we can allow the random variables X_n to be independent but have any distribution on $A \cup A^{-1}$. This distribution is given by numbers $p_a = p_a^{-1} \in [0, 1]$ satisfying $\sum_{a \in A \cup A^{-1}} p_a = 1$.

Definition. The **free Laplacian** for the random walk given by (G, A, p) is the linear operator on $l^2(G)$ defined by

$$L_{gh} = p_{g-h} \, .$$

Since we assumed $p_a = p_{a^{-1}}$, the matrix L is symmetric: $L_{gh} = L_{hg}$ and the spectrum

$$\sigma(L) = \{E \in \mathbb{C} \mid (L - E) \text{ is invertible } \}$$

is a compact subset of the real line.

Remark. One can interpret L as the transition probability matrix of the random walk which is a ”Markov chain”. We will come back to this interpretation later.

Example. $G = \mathbb{Z}$, $A = \{1\}$. $p = p_a = 1/2$ for $a = 1, -1$ and $p_a = 0$ for $a \notin \{1, -1\}$. The matrix

$$L = \begin{bmatrix}
\cdot & \cdot & & & & & & \\
\cdot & 0 & p & & & & & \\
& p & 0 & p & & & & \\
& & p & 0 & p & & & \\
& & & p & 0 & p & & \\
& & & & p & 0 & p & \\
& & & & & p & 0 & \cdot \\
& & & & & & \cdot & \cdot
\end{bmatrix}$$

is also called a **Jacobi matrix**. It acts on the Hilbert space $l^2(\mathbb{Z})$ by $(Lu)_n = p(u_{n+1} + u_{n-1})$.

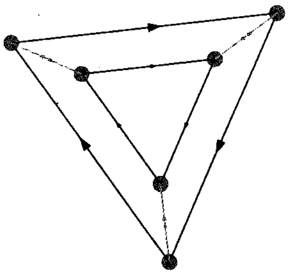
Example. Let $G = D_3$ be the **dihedral group** which has the presentation $G = \langle a, b | a^3 = b^2 = (ab)^2 = 1 \rangle$. The group is the symmetry group of the equilateral triangle. It has 6 elements and it is the smallest non-Abelian group. Let us number the group elements with integers $\{1, 2 = a, 3 = a^2, 4 = b, 5 = ab, 6 = a^2b\}$. We have for example $3 \star 4 = a^2b = 6$ or $3 \star 5 = a^2ab = a^3b = b = 4$. In this case $A = \{a, b\}, A^{-1} = \{a^{-1}, b\}$ so that $A \cup A^{-1} = \{a, a^{-1}, b\}$. The Cayley graph of the group is a graph with 6 vertices. We could take the uniform distribution $p_a = p_b = p_{a^{-1}} = 1/3$ on $A \cup A^{-1}$, but lets instead chose the distribution $p_a = p_{a^{-1}} = 1/4, p_b = 1/2$, which is natural if we consider multiplication by b and multiplication by b^{-1} as different.

Example. The free Laplacian on D_3 with the random walk transition probabilities $p_a = p_{a^{-1}} = 1/4, p_b = 1/2$ is the matrix

$$L = \begin{bmatrix}
0 & 1/4 & 1/4 & 1/2 & 0 & 0 \\
1/4 & 0 & 1/4 & 0 & 1/2 & 0 \\
1/4 & 0 & 0 & 0 & 0 & 1/2 \\
1/2 & 0 & 0 & 0 & 1/4 & 1/4 \\
0 & 1/2 & 0 & 1/4 & 0 & 1/4 \\
0 & 0 & 1/2 & 1/4 & 0 & 0
\end{bmatrix}$$

which has the eigenvalues $(-3 \pm \sqrt{5})/8, (5 \pm \sqrt{5})/8, 1/4, -3/4$.

Figure. *The Cayley graph of the dihedral group $G = D_3$ is a regular graph with 6 vertices and 9 edges.*



A basic question is: what is the relation between the spectrum of L, the structure of the group G and the properties of the random walk on G..

Definition. As before, let m_n be the probability that the random walk starting in e returns in n steps to e and let

$$m(x) = \sum_{n \in G} m_n x^n$$

be the generating function of the sequence m_n.

Proposition 3.11.1. The norm of L is equal to $\limsup_{n\to\infty}(m_n)^{1/n}$, the inverse of the radius of convergence of $m(x)$.

Proof. Because L is symmetric and real, it is self-adjoint and the spectrum of L is a subset of the real line $\mathbb{R}$ and the spectral radius of L is equal to its norm $||L||$.
We have $[L^n]_{ee} = m_n$ since $[L^n]_{ee}$ is the sum of products $\prod_{j=1}^n p_{a_j}$ each of which is the probability that a specific path of length n starting and landing at e occurs.
It remains therefore to verify that

$$\limsup_{n\to\infty} ||L^n||^{1/n} = \limsup_{n\to\infty}[L^n]_{ee}^{1/n}$$

and since the $\geq$ direction is trivial we have only to show that $\leq$ direction. Denote by $\mathrm{E}(\lambda)$ the spectral projection matrix of L, so that $dE(\lambda)$ is a projection-valued measure on the spectrum and the spectral theorem says that L can be written as $L = \int \lambda \, dE(\lambda)$. The measure $\mu_e = dE_{ee}$ is called a **spectral measure** of L. The real number $E(\lambda) - E(\mu)$ is nonzero if and only if there exists some spectrum of L in the interval $[\lambda, \mu)$. Since

$$\frac{(-1)}{\lambda}\sum_n \frac{[L^n]_{ee}}{\lambda^n} = \int_{\mathbb{R}} (E - \lambda)^{-1} \, dk(E)$$

can't be analytic in λ in a point λ_0 of the support of dk which is the spectrum of L, the claim follows. □

Remark. We have seen that the matrix L defines a spectral measure μ_e on the real line. It can be defined for any group element g, not only $g = e$ and is the same measure. It is therefore also the so called **density of states** of L. If we think of μ as playing the role of the law for random variables, then the **integrated density of states** $E(\lambda) = F_L(\lambda) = \int_{-\infty} d\mu(\lambda)$ plays the role of the distribution function for real-valued random variables.

Example. The Fourier transform $U : l^2(\mathbb{Z}^1) \to L^2(\mathbb{T}^1)$:

$$\hat{u}(x) = (Uu)(x) = \sum_{n \in \mathbb{Z}} u_n e^{inx}$$

diagonalises the matrix L for the random walk on $\mathbb{Z}^1$

$$\begin{aligned}
(ULU^*)\hat{u}(x) &= ((UL)(u_n)(x) = pU(u_{n+1} + u_{n-1})(x) \\
&= p\sum_{n \in \mathbb{Z}} (u_{n+1} + u_{n-1})e^{inx} \\
&= p\sum_{n \in \mathbb{Z}} u_n(e^{i(n-1)x} + e^{i(n+1)x}) \\
&= p\sum_{n \in \mathbb{Z}} u_n(e^{ix} + e^{-ix})e^{inx} \\
&= p\sum_{n \in \mathbb{Z}} u_n 2\cos(x)e^{inx} \\
&= 2p\cos(x) \cdot \hat{u}(x) .
\end{aligned}$$

This shows that the spectrum of ULU^* is $[-1, 1]$ and because U is an unitary transformation, also the spectrum of L is in $[-1, 1]$.

Example. Let $G = \mathbb{Z}^d$ and $A = \{e_i\}_{i=1}^d$, where $\{e_i\}$ is the standard bases. Assume $p = p_a = 1/(2d)$. The analogous Fourier transform $F : l^2(\mathbb{Z}^d) \to L^2(\mathbb{T}^d)$ shows that FLF^* is the multiplication with $\frac{1}{d}\sum_{j=1}^d \cos(x_j)$. The spectrum is again the interval $[-1, 1]$.

Example. The Fourier diagonalisation works for any discrete Abelian group with finitely many generators.

Example. $G = F_d$ the free group with the natural d generators. The spectrum of L is

$$[-\frac{\sqrt{2d-1}}{d}, \frac{\sqrt{2d-1}}{d}]$$

which is strictly contained in $[-1, 1]$ if $d > 1$.

Remark. Kesten has shown that the spectral radius of L is equal to 1 if and only if the group G has an invariant mean. For example, for a finite graph, where L is a stochastic matrix, for which each column is a probability vector, the spectral radius is 1 because L^T has the eigenvector $(1, \dots, 1)$ with eigenvalue 1.

Random walks and Laplacian can be defined on any graph. The spectrum of the Laplacian on a finite graph is an invariant of the graph but there are non-isomorphic graphs with the same spectrum. There are known infinite self-similar graphs, for which the Laplacian has pure point spectrum [63]. There are also known infinite graphs, such that the Laplacian has purely singular continuous spectrum [95]. For more on spectral theory on graphs, start with [6].

3.12 A discrete Feynman-Kac formula

Definition. A **discrete Schrödinger operator** is a bounded linear operator L on the Hilbert space $l^2(\mathbb{Z}^d)$ of the form

$$(Lu)(n) = \sum_{i=1}^{d} u(n+e_i) - 2u(n) + u(n-e_i) + V(n)u(n) \; ,$$

where V is a bounded function on $\mathbb{Z}^d$. They are discrete versions of operators $L = -\Delta + V(x)$ on $L^2(\mathbb{R}^d)$, where Δ is the free Laplacian. Such operators are also called **Jacobi matrices.**

Definition. The **Schrödinger equation**

$$i\hbar\dot{u} = Lu, \; u(0) = u_0$$

is a differential equation in $l^2(\mathbb{Z}^d, \mathbb{C})$ which describes the motion of a complex valued wave function u of a classical quantum mechanical system. The constant $\hbar$ is called the **Planck constant** and $i = \sqrt{-1}$ is the imaginary unit. Lets assume to have units where $\hbar = 1$ for simplicity.

Remark. The solution of the Schrödinger equation is

$$u_t = e^{\frac{t}{i}L} u_0 \; .$$

The solution exists for all times because the von Neumann series

$$e^{tL} = 1 + tL + \frac{t^2L^2}{2!} + \frac{t^3L^3}{3!} + \cdots$$

is in the space of bounded operators.

Remark. It is an achievement of the physicist Richard Feynman to see that the evolution as a **path integral**. In the case of differential operators L, where this idea can be made rigorous by going to imaginary time and one can write for $L = -\Delta + V$

$$e^{-t:}u(x) = \mathrm{E}_x[e^{\int_0^t V(\gamma(s)) \, ds} u_0(\gamma(t))] \; ,$$

where E_x is the expectation value with respect to the measure P_x on the Wiener space of Brownian motion starting at x.

Here is a discrete version of the **Feynman-Kac formula**:

Definition. The Schrödinger equation with **discrete time** is defined as

$$i(u_{t+\epsilon} - u_t) = \epsilon L u_t ,$$

where $\epsilon > 0$ is fixed. We get the evolution

$$u_{t+n\epsilon} = (1 - i\epsilon L)^n u_t$$

and we denote the right hand side with $\tilde{L}^n u_t$.

Definition. Denote by $\Gamma_n(i,j)$ the set of paths of length n in the graph G having as edges $\mathbb{Z}^d$ and sites pairs $[i,j]$ with $|i-j| \leq 1$. The graph G is the **Cayley graph** of the group $\mathbb{Z}^d$ with the generators $A \cup A^{-1} \cup \{e\}$, where $A = \{e_1, \dots, e_d, \}$ is the set of natural generators and where e is the identity.

Definition. Given a path γ of finite length n, we use the notation

$$\exp(\int_\gamma L) = \prod_{i=1}^{n} L_{\gamma(i),\gamma(i+1)} \cdot$$

Let Ω is the set of all paths on G and E denotes the expectation with respect to a measure P of the random walk on G starting at 0.

Theorem 3.12.1 (Discrete Feynman-Kac formula). Given a discrete Schrödinger operator L. Then

$$(L^n u)(0) = \mathrm{E}_0[\exp(\int_0^n L)\, u(\gamma(n))] .$$

Proof.

$$\begin{aligned} (L^n u)(0) &= \sum_j (L^n)_{0j} u(j) \\ &= \sum_j \sum_{\gamma \in \Gamma_n(0,j)} \exp(\int_0^n L)\, u(j) \\ &= \sum_{\gamma \in \Gamma_n} \exp(\int_0^n L) u(\gamma(n)) . \end{aligned}$$

□

Remark. This discrete random walk expansion corresponds to the Feynman-Kac formula in the continuum. If we extend the potential to all the sites of the Cayley graph by putting $V([k,k]) = V(k)$ and $V([k,l]) = 0$ for $k \neq l$, we can define $\exp(\int_\gamma V)$ as the product $\prod_{i=1}^n V([\gamma(i), \gamma(i+1)])$. Then

$$(L^n u)(0) = \mathrm{E}[\exp(\int_0^n V) u(\gamma(n))]$$

which is formally the Feynman-Kac formula.

In order to compute $(\tilde{L}^n u)(k)$ with $\tilde{L} = (1 - k\epsilon L)$, we have to take the potential $\tilde{v}$ defined by

$$\tilde{v}([k,k]) = 1 - i\epsilon v(\gamma(k)) .$$

Remark. The Schrödinger equation with discrete time has the disadvantage that the time evolution of the quantum mechanical system is no more unitary. This draw-back could be overcome by considering also $i\hbar(u_t - u_{t-\epsilon}) = \epsilon L u_t$ so that the propagator from $u_{t-\epsilon}$ to $u_{t+\epsilon}$ is given by the unitary operator

$$U = (1 - \frac{i\epsilon}{\hbar} L)(1 + \frac{i\epsilon}{\hbar} L)^{-1}$$

which is a **Cayley transform** of L. See also [50], where the idea is disussed to use $\tilde{L} = \arccos(aL)$, where L has been rescaled such that aL has norm smaller or equal to 1. The time evolution can then be computed by iterating the map $A : (\psi, \phi) \mapsto (2aL\psi - \phi, \psi)$ on $H \oplus H$.

3.13 Discrete Dirichlet problem

Also for other partial differential equations, solutions can be described probabilistically. We look here at the Dirichlet problem in a bounded discrete region. The formula which we derive in this situation holds also in the continuum limit, where the random walk is replaced by Brownian motion.

Definition. The **discrete Laplacian** on $\mathbb{Z}^2$ is defined as

$$\Delta f(n,m) = f(n+1,m) + f(n-1,m) + f(n,m+1) + f(n,m-1) - 4f(n,m) .$$

With the discrete partial derivatives

$$\delta_x^+ f(n,m) = \frac{1}{2}(f(n+1,m) - f(n,m)),\ \delta_x^- f(n,m) = \frac{1}{2}(f(n,m) - f(n-1,m)) ,$$

$$\delta_y^+ f(n,m) = \frac{1}{2}(f(n,m+1) - f(n,m)),\ \delta_y^- f(n,m) = \frac{1}{2}(f(n,m) - f(n,m-1)) ,$$

the Laplacian is the sum of the second derivatives as in the continuous case, where $\Delta = f_{xx} + f_{yy}$:

$$\Delta = \delta_x^+ \delta_x^- + \delta_y^+ \delta_y^- .$$

The discrete Laplacian in $\mathbb{Z}^3$ is defined in the same way as a discretisation of $\Delta = f_{xx} + f_{yy} + f_{zz}$. The setup is analogue in higher dimensions

$$(\Delta u)(n) = \frac{1}{2d} \sum_{i=1}^d (u(n + e_i) + u(n - e_i) - 2u(n)) ,$$

where $e_1, \dots, e_d$ is the standard basis in $\mathbb{Z}^d$.

Definition. A **bounded region** D **in** $\mathbb{Z}^d$ is a finite subset of $\mathbb{Z}^d$. Two points are **connected** in D if they are connected in $\mathbb{Z}^3$. The boundary δD of D consists of all lattice points in D which have a neighboring lattice point which is outside D. Given a function f on the boundary δD, the **discrete Dirichlet problem** asks for a function u on D which satisfies the discrete Laplace equation $\Delta u = 0$ in the interior $\text{int}(D)$ and for which $u = f$ on the boundary δD.

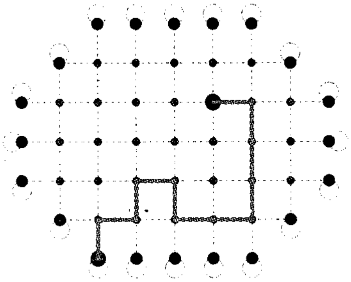


Figure. *The discrete Dirichlet problem is a problem in linear algebra. One algorithm to solve the problem can be restated as a probabilistic "path integral method". To find the value of u at a point x, look at the "discrete Wiener space" of all paths γ starting at x and ending at some boundary point $S_T(\omega) \in \delta D$ of D. The solution is $u(x) = \mathrm{E}_x[f(S_T)]$.*

Definition. Let $\Omega_{x,n}$ denote the set of all paths of length n in D which start at a point $x \in D$ and end up at a point in the boundary δD. It is a subset of $\Gamma_{x,n}$, the set of all paths of length n in $\mathbb{Z}^d$ starting at x. Lets call it the **discrete Wiener space of order** n defined by x and D. It is a subset of the set $\Gamma_{x,n}$ which has 2^{dn} elements. We take the uniform distribution on this finite set so that $\mathrm{P}_{x,n}[\{\gamma\}] = 1/2^{dn}$.

Definition. Let L be the matrix for which $L_{x,y} = 1/(2d)$ if $x, y \in \mathbb{Z}^d$ are connected by a path and x is in the interior of D. The matrix L is a bounded linear operator on $l^2(D)$ and satisfies $L_{x,z} = L_{z,x}$ for $x, z \in \text{int}(D) = D \backslash \delta D$. Given $f : \delta D \to \mathbb{R}$, we extend f to a function $F(x) = 0$ on $\int D = D \setminus \delta D$ and $F(x) = f(x)$ for $x \in \delta D$. The discrete Dirichlet problem can be restated as the problem to find the solution u to the system of linear equations

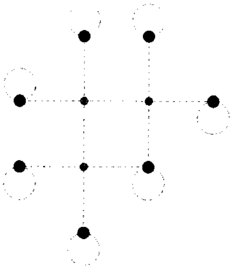
$$(1 - L)u = f \,.$$

Lemma 3.13.1. The number of paths in $\Omega_{x,n}$ starting at $x \in D$ and ending at a different point $y \in D$ is equal to $(2d)^n L^n_{xy}$.

Proof. Use induction. By definition, L_{xz} is $1/(2d)$ if there is a path from x to z. The integer $L^n_{x,y}$ is the number of paths of length n from x to y. □

Figure. *Here is an example of a problem where $D \subset \mathbb{Z}^2$ has 10 points:*

$$4L = \begin{bmatrix} 0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0 \\ 1&0&1&0&1&0&0&1&0&0 \\ 0&1&0&1&0&1&0&0&1&0 \\ 0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&1&0&1&0&0&1&1 \\ 0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0 \end{bmatrix} .$$



Only the rows corresponding to interior points are nonzero.

Definition. For a function f on the boundary δD, define

$$\mathrm{E}_{x,n}[f] = \sum_{y \in \delta D} f(y) L^n_{x,y}$$

and

$$\mathrm{E}_x[f] = \sum_{n=0}^{\infty} \mathrm{E}_{x,n}[f] .$$

This functional defines for every point $x \in D$ a probability measure μ_x on the boundary δD. It is the discrete analog of the **harmonic measure** in the continuum. The measure P_x on the set of paths satisfies $\mathrm{E}_x[1] = 1$ as we will just see.

Proposition 3.13.2. Let S_n be the random walk on $\mathbb{Z}^d$ and let T be the stopping time which is the first exit time of S from D. The solution to the discrete Dirichlet problem is

$$u(x) = \mathrm{E}_x[f(S_T)] .$$

Proof. Because $(1 - L)u = f$ and

$$\mathrm{E}_{x,n}[f] = (L^n f)_x ,$$

we have from the geometric series formula

$$(1-A)^{-1} = \sum_{k=0}^{n} A^k$$

the result

$$u(x) = (1-L)^{-1} f(x) = \sum_{n=0}^{\infty} [L^n f]_x = \sum_{n=0}^{\infty} \mathrm{E}_{x,n}[f] = \mathrm{E}_x[S_T] \; .$$

Define the matrix K by $K_{jj} = 1$ for $j \in \delta D$ and $K_{ij} = L_{ji}/4$ else. The matrix K is a **stochastic matrix**: its column vectors are probability vectors. The matrix K has a maximal eigenvalue 1 and so norm 1 (K^T has the maximal eigenvector $(1, 1, \dots, 1)$ with eigenvalue 1 and since eigenvalues of K agree with eigenvalues of K^T). Because $||L|| < 1$, the spectral radius of L is smaller than 1 and the series converges. If $f = 1$ on the boundary, then $u = 1$ everywhere. From $\mathrm{E}_x[1] = 1$ follows that the discrete Wiener measure is a probability measure on the set of all paths starting at x. □

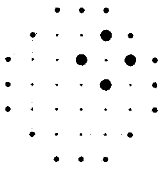


Figure. *The random walk defines a diffusion process.*

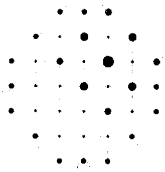


Figure. *The diffusion process after time $t =$* 2.

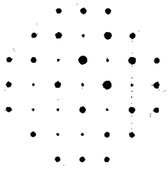


Figure. *The diffusion process after time $t =$* 3.

The path integral result can be generalized and the increased generality makes it even simpler to describe:

Definition. Let (D, E) be an arbitrary finite directed graph, where D is a finite set of n **vertices** and $E \subset D \times D$ is the set of **edges**. Denote an edge connecting i with j with e_{ij}. Let K be a **stochastic matrix** on $l^2(D)$: the entries satisfy $K_{ij} \geq 0$ and its column vectors are probability vectors $\sum_{i \in D} K_{ij} = 1$ for all $j \in D$. The stochastic matrix encodes the graph and additionally defines a random walk on D if K_{ij} is interpreted as the transition probability to hop from j to i. Lets call a point $j \in \delta D$ a **boundary point**, if $K_{jj} = 1$. The complement $\mathrm{int}D = D \backslash \delta D$ consists of **interior points**. Define the matrix L as $L_{jj} = 0$ if j is a boundary point and $L_{ij} = K_{ji}$ otherwise.

The **discrete Wiener space** $\Omega_x \subset D$ on D is the set of all finite paths $\gamma = (x = x_0, x_1, x_2, \dots, x_n)$ starting at a point $x \in D$ for which $K_{x_i x_{i+1}} > 0$. The **discrete Wiener measure** on this countable set is defined as $\mathrm{P}_x[\{\gamma\}] = \prod_{j=0}^{n-1} K_{j,j+1}$. A function u on D is called **harmonic** if $(Lu)_x = 0$ for all $x \in D$. The **discrete Dirichlet problem on the graph** is to find a function u on D which is harmonic and which satisfies $u = f$ on the boundary δD of D.

Theorem 3.13.3 (The Dirichlet problem on graphs). Assume D is a directed graph. If S_n is the random walk starting at x and T is the stopping time to reach the boundary of D, then the solution

$$u = \mathrm{E}_x[f(S_T)]$$

is the expected value of S_T on the discrete Wiener space of all paths starting at x and ending at the boundary of D.

Proof. Let F be the function on D which agrees with f on the boundary of D and which is 0 in the interior of D. The Dirichlet problem on the graph is the system of linear equations $(1 - L)u = f$. Because the matrix L has spectral radius smaller than 1, the problem is given by the geometric series

$$u = \sum_{n=0}^{\infty} L^n f \, .$$

But this is the sum $\mathrm{E}_x[f(S_T)]$ over all paths γ starting at x and ending at the boundary of f. □

Example. Lets look at a directed graph (D, E) with 5 vertices and 2 boundary points. The Laplacian on D is defined by the stochastic matrix

$$K = \begin{bmatrix} 0 & 1/3 & 0 & 0 & 0 \\ 1/2 & 0 & 1 & 0 & 0 \\ 1/4 & 1/2 & 0 & 0 & 0 \\ 1/8 & 1/6 & 0 & 1 & 0 \\ 1/8 & 0 & 0 & 0 & 1 \end{bmatrix}$$

or the Laplacian

$$L = \begin{bmatrix} 0 & 1/2 & 1/4 & 1/8 & 1/8 \\ 1/3 & 0 & 1/2 & 1/6 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} .$$

Given a function f on the boundary of D, the solution u of the discrete Dirichlet problem $(1 - L)u = f$ on this graph can be written as a path integral $\sum_{n=0}^{\infty} L^n f = \mathrm{E}_x[f(S_T)]$ for the random walk S_n on D stopped at the boundary δD.

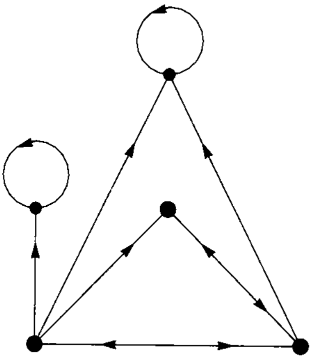


Figure. *The directed graph (D,E) with 5 vertices and 2 boundary points.*

Remark. The interplay of random walks on graphs and discrete partial differential equations is relevant in electric networks. For mathematical treatments, see [19, 99].

3.14 Markov processes

Definition. Given a measurable space $(S,\mathcal{B})$ called **state space**, where S is a set and $\mathcal{B}$ is a σ-algebra on S. A function $P : S \times \mathcal{B} \to \mathbb{R}$ is called a **transition probability function** if $P(x,\cdot)$ is a probability measure on $(S,\mathcal{B})$ for all $x \in S$ and if for every $B \in \mathcal{B}$, the map $s \to P(s,B)$ is $\mathcal{B}$-measurable. Define $P^1(x,B) = P(x,B)$ and inductively the measures $P^{n+1}(x,B) = \int_S P^n(y,B)P(x,dy)$, where we write $\int P(x,dy)$ for the integration on S with respect to the measure $P(x,\cdot)$.

Example. If S is a finite set and $\mathcal{B}$ is the set of all subsets of S. Given a stochastic matrix K and a point $s \in S$, the measures $P(s,\cdot)$ are the probability vectors, which are the columns of K.

Remark. The transition probability functions are elements in $\mathcal{L}(S, M_1(S))$, where $M_1(S)$ is the set of Borel probability measures on S. With the multiplication

$$(P \circ Q)(x,B) = \int_S P(y,B)\, dQ(x)$$

we get a commutative semi-group. The relation $P^{n+m} = P^n \circ P^m$ is also called the **Chapmann-Kolmogorov equation**.

Definition. Given a probability space $(\Omega,\mathcal{A},\mathrm{P})$ with a filtration $\mathcal{A}_n$ of σ-algebras. An $\mathcal{A}_n$-adapted process X_n with values in S is called a **discrete**

time Markov process if there exists a transition probability function P such that

$$\mathrm{P}[X_n \in B \mid \mathcal{A}_k](\omega) = P^{n-k}(X_k(\omega), B) .$$

Definition. If the state space S is a discrete space, a finite or countable set, then the Markov process is called a **Markov chain**, A Markov chain is called a **denumerable Markov chain**, if the state space S is countable, a **finite Markov chain**, if the state space is finite.

Remark. It follows from the definition of a Markov process that X_n satisfies the **elementary Markov property**: for $n > k$,

$$\mathrm{P}[X_n \in B \mid X_1, \dots, X_k] = \mathrm{P}[X_n \in B \mid X_k] .$$

This means that the probability distribution of X_n is determined by knowing the probability distribution of X_{n-1}. The future depends only on the present and not on the past.

Theorem 3.14.1 (Markov processes exist). For any state space $(S, \mathcal{B})$ and any transition probability function P, there exists a corresponding Markov process X.

Proof. Choose a probability measure μ on $(S, \mathcal{B})$ and define on the product space $(\Omega, \mathcal{A}) = (S^{\mathbb{N}}, \mathcal{B}^{\mathbb{N}})$ the π-system $\mathcal{C}$ consisting of of cylinder-sets $\prod_{n \in \mathbb{N}} B_n$ given by a sequence $B_n \in \mathcal{B}$ such that $B_n = S$ except for finitely many n. Define a measure $P = P_\mu$ on $(\Omega, \mathcal{C})$ by requiring

$$\mathrm{P}[\omega_k \in B_k, k = 1, \dots n] = \int_{B_0} \mu(dx_0) \int_{B_1} P(x_0, dx_1) \dots \int_{B_n} P(x_{n-1}, dx_n) .$$

This measure has a unique extension to the σ-algebra $\mathcal{A}$.
Define the increasing sequence of σ-algebras $\mathcal{A}_n = \mathcal{B}^n \times \prod_{i=1}^n \{\emptyset, \Omega\}$ containing cylinder sets. The random variables $X_n(\omega) = x_n$ are $\mathcal{A}^n$-adapted. In order to see that it is a Markov process, we have to check that

$$\mathrm{P}[X_n \in B_n \mid \mathcal{A}_{n-1}](\omega) = P(X_{n-1}(\omega), B_n)$$

which is a special case of the above requirement by taking $B_k = S$ for $k \neq n$. □

Example. Independent S-valued random variables
Assume the measures $P(x, \cdot)$ are independent of x. Call this measure P. In this case

$$\mathrm{P}[X_n \in B_n \mid \mathcal{A}_{n-1}](\omega) = \mathrm{P}[B_n]$$

which means that $\mathrm{P}[X_n \in B_n \mid \mathcal{A}_{n-1}] = \mathrm{P}[X_n \in B_n]$. The $S-$ valued random variables X_n are independent and have identical distribution and P is the law of X_n. Every sequence of IID random variables is a Markov process.

Example. Countable and finite state Markov chains.
Given a Markov process with finite or countable state space S. We define the transition matrix P_{ij} on the Hilbert space $l^2(S)$ by

$$P_{ij} = P(i, \{j\}) \, .$$

The matrix P transports the law of X_n into the law of X_{n+1}.
The transition matrix P_{ij} is a **stochastic matrix**: each column is a probability vector: $\sum_j P_{ij} = 1$ with $P_{ij} \geq 0$. Every measure on S can be given by a vector $\pi \in l^2(S)$ and $P\pi$ is again a measure. If X_0 is constant and equal to i and X_n is a Markov process with transition probability P, then $P^n_{ij} = \mathrm{P}[X_n = j]$.

Example. Sum of independent S-valued random variables Let S be a countable Abelian group and let π be a probability distribution on S assigning to each $j \in S$ the weight π_j. Define $P_{ij} = \pi_{j-i}$. Now X_n is the sum of n independent random variables with law π. The sum changes from i to j with probability $P_{ij} = p_{i-j}$.

Example. Branching processes Given $S = \{0, 1, 2 \dots \} = \mathbb{N}$ with fixed probability distribution π. If X is a S-valued random variable with distribution π then $\sum_{k=1}^n X_k$ has a distribution which we denote by $\pi^{(n)}$. Define the matrix $P_{ij} = \pi^{(i)}_j$. The Markov chain with this transition probability matrix on S is called a **branching process**.

Definition. The transition probability function P acts also on measures π of S by

$$\mathcal{P}(\pi)(B) = \int_S P(x, B) \; d\pi(x) \, .$$

A probability measure π is called **invariant** if $\mathcal{P}\pi = \pi$. An invariant measure π on S is called **stationary measure** of the Markov process.

This operator on measures leaves a subclass of measures with densities with respect to some measure ν invariant. We can so assign a **Markov operator** to a transition probability function:

Lemma 3.14.2. For any $x \in S$ define the measure

$$\nu(B) = \sum_{n=0}^{\infty} \frac{1}{2^n} \mathcal{P}^n(x, B)$$

on $(S, \mathcal{B})$ has the property that if μ is absolutely continuous with respect to ν, then also $\mathcal{P}\mu$ is absolutely continuous with respect to ν.

Proof. Given $\mu = f \cdot \nu$ with $f \in L^1(S)$. Then

$$\mathcal{P}\mu = \int_S P(x,B) f(x) \; d\nu(x)$$

is absolutely continuous with respect to ν because $\mathcal{P}\mu(B) = 0$ implies $P(x,B) = 0$ for almost all x with $f(x) > 0$ and so $f\nu(B) = 0$. □

Corollary 3.14.3. To each transition probability function can be assigned a Markov operator $\mathcal{P} : L^1(S,\nu) \to L^1(S,\nu)$.

Proof. Choose ν as above and define

$$\mathcal{P}f_1 = f_2$$

if $\mathcal{P}\mu_1 = \mu_2$ with $\mu_i = f_i\nu_i$. To check that $\mathcal{P}$ is a Markov operator, we have to check $\mathcal{P}f \geq 0$ if $f \geq 0$, which follows from

$$\mathcal{P}f\nu(B) = \int_S P(x,B) f(x) \; d\nu(x) \geq 0 \, .$$

We also have to show that $||\mathcal{P}f||_1 = 1$ if $||f||_1$. It is enough to show this for elementary functions $f = \sum_j a_j 1_{B_j}$ with $a_j > 0$ with $B_j \in \mathcal{B}$ satisfying $\sum_j a_j \nu(B_j) = 1$ satisfies $||\mathcal{P}1_B\nu|| = \nu(B)$. But this is obvious $||\mathcal{P}1_B\nu|| = \int_B P(x,\cdot) \; d\nu(x) = \nu(B)$. □

We see that the abstract approach to study Markov operators on $L^1(S)$ is more general, than looking at transition probability measures. This point of view can reduce some of the complexity, when dealing with discrete time Markov processes.

Chapter 4

Continuous Stochastic Processes

4.1 Brownian motion

Definition. Let $(\Omega, \mathcal{A}, \mathrm{P})$ be a probability space and let $T \subset \mathbb{R}$ be time. A collection of random variables X_t, $t \in T$ with values in $\mathbb{R}$ is called a **stochastic process**. If X_t takes values in $S = \mathbb{R}^d$, it is called a **vector-valued stochastic process** but one often abbreviates this by the name stochastic process too. If the time T can be a discrete subset of $\mathbb{R}$, then X_t is called a **discrete time stochastic process**. If time is an interval, $\mathbb{R}^+$ or $\mathbb{R}$, it is called a **stochastic process with continuous time**. For any fixed $\omega \in \Omega$, one can regard $X_t(\omega)$ as a function of t. It is called a **sample function** of the stochastic process. In the case of a vector-valued process, it is a **sample path**, a curve in $\mathbb{R}^d$.

Definition. A stochastic process is called **measurable**, if $X : T \times \Omega \to S$ is measurable with respect to the product σ-algebra $\mathcal{B}(T) \times \mathcal{A}$. In the case of a real-valued process ($S = \mathbb{R}$), one says X is **continuous in probability** if for any $t \in \mathbb{R}$ the limit $X_{t+h} \to X_t$ takes place in probability for $h \to 0$. If the sample function $X_t(\omega)$ is a continuous function of t for almost all ω, then X_t is called a **continuous stochastic process**. If the sample function is **a right continuous** function in t for almost all $\omega \in \Omega$, X_t is called a **right continuous stochastic process**. Two stochastic process X_t and Y_t satisfying $\mathrm{P}[X_t - Y_t = 0] = 1$ for all $t \in T$ are called **modifications of each other** or **indistinguishable**. This means that for almost all $\omega \in \Omega$, the sample functions coincide $X_t(\omega) = Y_t(\omega)$.

Definition. A $\mathbb{R}^n$-valued random vector X is called **Gaussian**, if it has the multidimensional characteristic function

$$\phi_X(s) = \mathrm{E}[e^{is \cdot X}] = e^{-(s,Vs)/2 + i(m,s)}$$

for some nonsingular symmetric $n \times n$ matrix V and vector $m = \mathrm{E}[X]$. The matrix V is called **covariance matrix** and the vector m is called the **mean vector**.

Example. A normal distributed random variable X is a Gaussian random variable. The covariance matrix is in this case the scalar $\mathrm{Var}[X]$.

Example. If V is a symmetric matrix with determinant $\det(V) \neq 0$, then the random variable

$$X(x) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(V)}} e^{-(x-m,V^{-1}(x-m))/2}$$

on $\Omega = \mathbb{R}^n$ is a Gaussian random variable with covariance matrix V. To see that it has the required multidimensional characteristic function $\phi_X(u)$. Note that because V is symmetric, one can diagonalize it. Therefore, the computation can be done in a bases, where V is diagonal. This reduces the situation to characteristic functions for normal random variables.

Example. A set of random variables $X_1, \dots, X_n$ are called **jointly Gaussian** if any linear combination $\sum_{i=1}^n a_i X_i$ is a Gaussian random variable too. For a jointly Gaussian set of of random variables X_j, the vector $X = (X_1, \dots, X_n)$ is a Gaussian random vector.

Example. A **Gaussian process** is a $\mathbb{R}^d$-valued stochastic process with continuous time such that $(X_{t_0}, X_{t_1}, \dots, X_{t_n})$ is jointly Gaussian for any $t_0 \leq t_1 < \cdots < t_n$. It is called **centered** if $m_t = \mathrm{E}[X_t] = 0$ for all t.

Definition. An $\mathbb{R}^d$-valued continuous Gaussian process X_t with **mean vector** $m_t = \mathrm{E}[X_t]$ and the covariance matrix $V(s,t) = \mathrm{Cov}[X_s, X_t] = \mathrm{E}[(X_s - m_s)\cdot(X_t - m_t)^*]$ is called **Brownian motion** if for any $0 \leq t_0 < t_1 < \cdots < t_n$, the random vectors $X_{t_0}, X_{t_{i+1}} - X_{t_i}$ are independent and the covariance matrix V satisfies $V(s,t) = V(r,r)$, where $r = \min(s,t)$ and $s \mapsto V(s,s)$. It is called the **standard Brownian motion** if $m_t = 0$ for all t and $V(s,t) = \min\{s,t\}$.

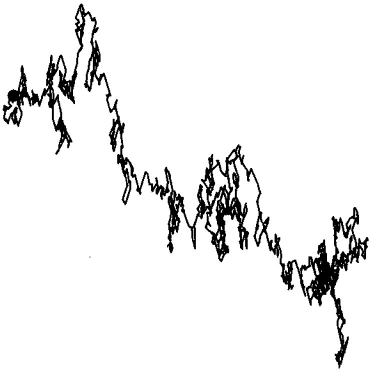


Figure. *A path $X_t(\omega_1)$ of Brownian motion in the plane $S = \mathbb{R}^2$ with a drift $m_t = \mathrm{E}[X_t] = (t,0)$. This is not standard Brownian motion. The process $Y_t = X_t - (t,0)$ is standard Brownian motion.*

Recall that for two random vectors X, Y with mean vectors m, n, the covariance matrix is $\text{Cov}[X,Y]_{ij} = \text{E}[(X_i - m_i)(Y_j - n_j)]$. We say $\text{Cov}[X,Y] = 0$ if this matrix is the zero matrix.

Lemma 4.1.1. A Gaussian random vector (X, Y) with random vectors X, Y satisfying $\text{Cov}[X,Y] = 0$ has the property that X and Y are independent.

Proof. We can assume without loss of generality that the random variables X, Y are centered. Two $\mathbb{R}^n$-valued Gaussian random vectors X and Y are independent if and only if

$$\phi_{(X,Y)}(s,t) = \phi_X(s) \cdot \phi_Y(t), \forall s, t \in \mathbb{R}^n$$

Indeed, if V is the covariance matrix of the random vector X and W is the covariance matrix of the random vector Y, then

$$U = \begin{bmatrix} U & \text{Cov}[X,Y] \\ \text{Cov}[Y,X] & V \end{bmatrix} = \begin{bmatrix} U & 0 \\ 0 & V \end{bmatrix}$$

is the covariance matrix of the random vector (X, Y). With $r = (t, s)$, we have therefore

$$\begin{aligned} \phi_{(X,Y)}(r) &= \text{E}[e^{ir\cdot(X,Y)}] = e^{-\frac{1}{2}(r\cdot Ur)} \\ &= e^{-\frac{1}{2}(s\cdot Vs) - \frac{1}{2}(t\cdot Wt)} \\ &= e^{-\frac{1}{2}(s\cdot Vs)}e^{-\frac{1}{2}(t\cdot Wt)} \\ &= \phi_X(s)\phi_Y(t) \ . \end{aligned}$$

□

Example. In the context of this lemma, one should mention that there exist uncorrelated normal distributed random variables X, Y which are **not** independent [109]: Proof. Let X be Gaussian on $\mathbb{R}$ and define for $\alpha > 0$ the variable $Y(\omega) = -X(\omega)$, if $\omega > \alpha$ and $Y = X$ else. Also Y is Gaussian and there exists α such that $\text{E}[XY] = 0$. But X and Y are not independent and $X+Y = 0$ on $[-\alpha, \alpha]$ shows that $X+Y$ is not Gaussian. This example shows why Gaussian vectors (X, Y) are defined directly as $\mathbb{R}^2$ valued random variables with some properties and not as a vector (X, Y) where each of the two component is a one-dimensional random Gaussian variable.

Proposition 4.1.2. If X_t is a Gaussian process with covariance $V(s,t) = V(r,r)$ with $r = \min(s,t)$, then it is Brownian motion.

Proof. By the above lemma (4.1.1), we only have to check that for all $i < j$

$$\text{Cov}[X_{t_0}, X_{t_{j+1}} - X_{t_j}] = 0, \ \text{Cov}[X_{t_{i+1}} - X_{t_i}, X_{t_{j+1}} - X_{t_j}] = 0 \ .$$

But by assumption

$$\text{Cov}[X_{t_0}, X_{t_{j+1}} - X_{t_j}] = V(t_0, t_{j+1}) - V(t_0, t_j) = V(t_0, t_0) - V(t_0, t_0) = 0$$

and

$$\begin{aligned} \text{Cov}[X_{t_{i+1}} - X_{t_i}, X_{t_{j+1}} - X_{t_j}] &= V(t_{i+1}, t_{j+1}) - V(t_{i+1}, t_j) \\ &\quad -V(t_i, t_{j+1}) + V(t_i, t_j) \\ &= V(t_{i+1}, t_{i+1}) - V(t_{i+1}, t_{i+1}) \\ &\quad -V(t_i, t_i) + V(t_i, t_i) = 0 \ . \end{aligned}$$

□

Remark. Botanist **Robert Brown** was studying the fertilization process in a species of flowers in 1828. While watching pollen particles in water through a microscope, he observed small particles in "rapid oscillatory motion". While previous studies concluded that these particles were alive, Brown's explanation was that matter is composed of small "active molecules", which exhibit a rapid, irregular motion having its origin in the particles themselves and not in the surrounding fluid. Brown's contribution was to establish Brownian motion as an important phenomenon, to demonstrate its presence in inorganic as well as organic matter and to refute by experiment incorrect mechanical or biological explanations of the phenomenon. The book [73] includes more on the history of Brownian motion.

The construction of Brownian motion happens in two steps: one first constructs a Gaussian process which has the desired properties and then shows that it has a modification which is continuous.

Proposition 4.1.3. Given a separable real Hilbert space $(H, ||\cdot||)$. There exists a probability space $(\Omega, \mathcal{A}, \text{P})$ and a family $X(h), h \in H$ of real-valued random variables on Ω such that $h \mapsto X(h)$ is linear, and $X(h)$ is Gaussian, centered and $\text{E}[X(h)^2] = ||h||^2$.

Proof. Pick an orthonormal basis $\{e_n\}$ in H and attach to each e_n a centered Gaussian IID random variable $X_n \in \mathcal{L}^2$ satisfying $||X_n||_2 = 1$. Given a general $h = \sum h_n e_n \in H$, define

$$X(h) = \sum_n h_n X_n$$

which converges in $\mathcal{L}^2$. Because X_n are independent, they are orthonormal in $\mathcal{L}^2$ so that

$$||X(h)||_2^2 = \sum_n h_n^2 ||X_n||_2 = \sum_n h_n^2 = ||h||_2^2 \ .$$

□

Definition. If we choose $H = L^2(\mathbb{R}^+, dx)$, the map $X : H \mapsto \mathcal{L}^2$ is also called a **Gaussian measure**. For a Borel set $A \subset \mathbb{R}^+$ we define then $X(A) = X(1_A)$. The term "measure" is warranted by the fact that $X(A) = \sum_n X(A_n)$ if A is a countable disjoint union of Borel sets A_n. One also has $X(\emptyset) = 0$.

Remark. The space $X(H) \subset \mathcal{L}^2$ is a Hilbert space isomorphic to H and in particular

$$\mathrm{E}[X(h)X(h')] = (h, h') .$$

We know from the above lemma that h and h' are orthogonal if and only if $X(h)$ and $X(h')$ are independent and that

$$\mathrm{E}[X(A)X(B)] = \mathrm{Cov}[X(A), X(B)] = (1_A, 1_B) = |A \cap B| .$$

Especially $X(A)$ and $X(B)$ are independent if and only if A and B are disjoint.

Definition. Define the process $B_t = X([0,t])$. For any sequence $t_1, t_2, \dots \in T$, this process has independent increments $B_{t_i} - B_{t_{i-1}}$ and is a Gaussian process. For each t, we have $\mathrm{E}[B_t^2] = t$ and for $s < t$, the increment $B_t - B_s$ has variance $t - s$ so that

$$\mathrm{E}[B_s B_t] = \mathrm{E}[B_s^2] + \mathrm{E}[B_s(B_t - B_s)] = \mathrm{E}[B_s^2] = s .$$

This model of Brownian motion has everything except continuity.

Theorem 4.1.4 (Kolmogorov's lemma). Given a stochastic process X_t with $t \in [a,b]$ for which there exist three constants $p > r, K$ such that

$$\mathrm{E}[|X_{t+h} - X_t|^p] \leq K \cdot h^{1+r}$$

for every $t, t+h \in [a,b]$, then X_t has a modification Y_t which is almost everywhere continuous: for all $s,t \in [a,b]$

$$|Y_t(\omega) - Y_s(\omega)| \leq C(\omega)\ |t-s|^\alpha, 0 < \alpha < \frac{r}{p} .$$

Proof. We can assume without loss of generality that $a = 0, b = 1$ because we can translate and rescale the time variable to be in this situation. Define $\epsilon = r - \alpha p$. By the Chebychev-Markov inequality (2.5.4)

$$\mathrm{P}[|X_{t+h} - X_t|] \geq |h|^\alpha] \leq |h|^{-\alpha p}\mathrm{E}[|X_{t+h} - X_t|^p] \leq K|h|^{1+\epsilon}$$

so that

$$\mathrm{P}[|X_{(k+1)/2^n} - X_{k/2^n}| \geq 2^{-n\alpha}] \leq K 2^{-n(1+\epsilon)} .$$

Therefore

$$\sum_{n=1}^{\infty} \sum_{k=0}^{2^n-1} \mathrm{P}[|X_{(k+1)/2^n} - X_{k/2^n}| \geq 2^{-n\alpha}] < \infty .$$

By the first Borel-Cantelli's lemma (2.2.2), there exists $n(\omega) < \infty$ almost everywhere such that for all $n \geq n(\omega)$ and $k = 0, \dots, 2^n - 1$

$$|X_{(k+1)/2^n}(\omega) - X_{k/2^n}(\omega)| < 2^{-n\alpha} .$$

Let $n \geq n(\omega)$ and $t \in [k/2^n, (k+1)/2^n]$ of the form $t = k/2^n + \sum_{i=1}^m \gamma_i/2^{n+i}$ with $\gamma_i \in \{0,1\}$. Then

$$|X_t(\omega) - X_{k2^{-n}}(\omega)| \leq \sum_{i=1}^m \gamma_i 2^{-\alpha(n+i)} \leq d\, 2^{-n\alpha}$$

with $d = (1 - 2^{-\alpha})^{-1}$. Similarly

$$|X_t - X_{(k+1)2^{-n}}| \leq d\, 2^{-n\alpha} .$$

Given $t, t+h \in D = \{k2^{-n} \mid n \in \mathbb{N}, k = 0, \dots n-1\}$. Take n so that $2^{-n-1} \leq h < 2^{-n}$ and k so that $k/2^{n+1} \leq t < (k+1)/2^{n+1}$. Then $(k+1)/2^{n+1} \leq t+h \leq (k+3)/2^{n+1}$ and

$$|X_{t+h} - X_t| \leq 2d2^{-(n+1)\alpha} \leq 2dh^\alpha .$$

For almost all ω, this holds for sufficiently small h.

We know now that for almost all ω, the path $X_t(\omega)$ is uniformly continuous on the dense set of **dyadic numbers** $D = \{k/2^n\}$. Such a function can be extended to a continuous function on $[0,1]$ by defining

$$Y_t(\omega) = \lim_{s \in D \to t} X_s(\omega) .$$

Because the inequality in the assumption of the theorem implies $\mathrm{E}[X_t(\omega) - \lim_{s \in D \to t} X_s(\omega)] = 0$ and by Fatou's lemma $\mathrm{E}[Y_t(\omega) - \lim_{s \in D \to t} X_s(\omega)] = 0$ we know that $X_t = Y_t$ almost everywhere. The process Y is therefore a modification of X. Moreover, Y satisfies

$$|Y_t(\omega) - Y_s(\omega)| \leq C(\omega)\ |t-s|^\alpha$$

for all $s, t \in [a, b]$. □

Corollary 4.1.5. Brownian motion exists.

Proof. In one dimension, take the process B_t from above. Since $X_h = B_{t+h} - B_t$ is centered with variance h, the fourth moment is $\mathrm{E}[X_h^4] = \frac{d^4}{dx^4}\exp(-x^2h/2)_{|x=0} = 3h^2$, so that

$$\mathrm{E}[(B_{t+h} - B_t)^4] = 3h^2 .$$

Kolmogorov's lemma (4.1.4) assures the existence of a continuous modification of B.

To define standard Brownian motion in n dimension, we take the joint motion $B_t = (B_t^{(1)}, \dots, B_t^{(n)})$ of n independent one- dimensional Brownian motions. □

Definition. Let B_t be the standard Brownian motion. For any $x \in \mathbb{R}^n$, the process $X_t^x = x + B_t$ is called **Brownian motion started at** x.

The first rigorous construction of Brownian motion was given by **Norbert Wiener** in 1923. By construction of a **Wiener measure** on $C[0,1]$, one has a construction of Brownian motion, where the probability space is directly given by the set of paths. One has then the process $X_t(\omega) = \omega(t)$. We will come to this later. A general construction of such measures is possible given a Markov transition probability function [104]. The construction given here is due to Neveu and goes back to Kakutani. It can be found in Simon's book on functional integration [93] or in the book of Revuz and Yor [83] about continuous martingales and Brownian motion. This construction has the advantage that it can be applied to more general situations.

In McKean's book "Stochastic integrals" [66] one can find Lévy's direct proof of the existence of Brownian motion. Because that proof gives an explicit formula for the Brownian motion process B_t and is so **constructive**, we outline it shortly:

1) Take as a basis in $L^2([0,1]$ the **Haar functions**

$$f_{k,n} := 2^{(n-1)/2}(1_{[(k-1)2^{-n},k2^{-n}]} - 1_{[k2^{-n},(k+1)2^{-n}]})$$

for $\{(k,n) | n \geq 1, k < 2^n \}$ and $f_{0,0} = 1$.

2) Take a family $X_{k,n}$ for $(k,n) \in I = \{(k,n) \mid n \geq 1, k < 2^n, k \text{ odd} \} \cup \{(0,0) \}$ of independent Gaussian random variables.

3) Define

$$B_t = \sum_{(k,n)\in I} X_{k,n} \int_0^t f_{k,n} .$$

4) Prove convergence of the above series.

5) Check

$$\mathrm{E}[B_s B_t] = \sum_{(k,n)\in I} \int_0^s \int_0^t f_{(k,n)} f_\iota = \int_0^1 1_{[0,s]} 1_{[0,t]} = \inf\{s,t\} .$$

6) Extend the definition from $t \in [0,1]$ to $t \in [0,\infty)$ by taking independent Brownian motions $B_t^{(i)}$ and defining $B_t = \sum_{n<[t]} B_{t-n}^{(n)}$, where $[t]$ is the largest integer smaller or equal to t.

4.2 Some properties of Brownian motion

We first want to establish that Brownian motion is unique. To do so, we first have to say, when two processes are the same:

Definition. Two processes X_t on $(\Omega, \mathcal{A}, \mathrm{P})$ and X'_t on $(\Omega', \mathcal{A}', \mathrm{P}')$ are called **indistinguishable**, if there exists an isomorphism $U : \Omega \to \Omega'$ of probability spaces, such that $X'_t(U\omega) = X_t(\omega)$. Indistinguishable processes are considered the same. A special case is if the two processes are defined on the same probability space $(\Omega, \mathcal{A}, \mathrm{P})$ and $X_t(\omega) = Y_t(\omega)$ for almost all ω.

Proposition 4.2.1. Brownian motion is unique in the sense that two standard Brownian motions are indistinguishable.

Proof. The construction of the map $H \to \mathcal{L}^2$ was unique in the sense that if we construct two different processes $X(h)$ and $Y(h)$, then there exists an isomorphism U of the probability space such that $X(h) = Y(U(h))$. The continuity of X_t and Y_t implies then that for almost all ω, $X_t(\omega) = Y_t(U\omega)$. In other words, they are indistinguishable. □

We are now ready to list some symmetries of Brownian motion.

Theorem 4.2.2 (Properties of Brownian motion). The following symmetries exist:
(i) **Time-homogeneity**: For any $s > 0$, the process $\tilde{B}_t = B_{t+s} - B_s$ is a Brownian motion independent of $\sigma(B_u, u \leq s)$.
(ii) **Reflection symmetry**: The process $\tilde{B}_t = -B_t$ is a Brownian motion.
(iii) **Brownian scaling**: For every $c > 0$, the process $\tilde{B}_t = cB_{t/c^2}$ is a Brownian motion.
(iv) **Time inversion**: The process $\tilde{B}_0 = 0, \tilde{B}_t = tB_{1/t}, t > 0$ is a Brownian motion.

Proof. (i),(ii),(iii) In each case, $\tilde{B}_t$ is a continuous centered Gaussian process with continuous paths, independent increments and variance t.

(iv) $\tilde{B}$ is a centered Gaussian process with covariance

$$\mathrm{Cov}[\tilde{B}_s, \tilde{B}_t] = \mathrm{E}[\tilde{B}_s, \tilde{B}_t] = st \cdot \mathrm{E}[B_{1/s}, B_{1/t}] = st \cdot \inf(\frac{1}{s}, \frac{1}{t}) = \inf(s,t) .$$

Continuity of $\tilde{B}_t$ is obvious for $t > 0$. We have to check continuity only for $t = 0$, but since $\mathrm{E}[\tilde{B}_s^2] = s \to 0$ for $s \to 0$, we know that $\tilde{B}_s \to 0$ almost everywhere. □

It follows the **strong law of large numbers for Brownian motion**:

Theorem 4.2.3 (SLLN for Brownian motion). If B_t is Brownian motion, then

$$\lim_{t \to \infty} \frac{1}{t} B_t = 0$$

almost surely.

Proof. From the time inversion property (iv), we see that $t^{-1}B_t = B_{1/t}$ which converges for $t \to \infty$ to 0 almost everywhere, because of the almost everywhere continuity of B_t. □

Definition. A parameterized curve $t \in [0, \infty) \mapsto X_t \in \mathbb{R}^n$ is called **Hölder continuous of order** α if there exists a constant C such that

$$||X_{t+h} - X_t|| \leq C \cdot h^\alpha$$

for all $h > 0$ and all t. A curve which is Hölder continuous of order $\alpha = 1$ is called **Lipshitz continuous**.

The curve is called **locally Hölder continuous of order** α if there exists for each t a constant $C = C(t)$ such that

$$||X_{t+h} - X_t|| \leq C \cdot h^\alpha$$

for all small enough h. For a $\mathbb{R}^d$-valued stochastic process, (local) Hölder continuity holds if for almost all $\omega \in \Omega$ the sample path $X_t(\omega)$ is (local) Hölder continuous for almost all $\omega \in \Omega$.

Proposition 4.2.4. For every $\alpha < 1/2$, Brownian motion has a modification which is locally Hölder continuous of order α.

Proof. It is enough to show it in one dimension because a vector function with locally Hölder continuous component functions is locally Hölder continuous. Since increments of Brownian motion are Gaussian, we have

$$\mathrm{E}[(B_t - B_s)^{2p}] = C_p \cdot |t-s|^p$$

for some constant C_p. Kolmogorov's lemma assures the existence of a modification satisfying locally

$$|B_t - B_s| \leq C\,|t-s|^\alpha, 0 < \alpha < \frac{p-1}{2p} \; .$$

Because p can be chosen arbitrary large, the result follows. □

Because of this proposition, we can assume from now on that all the paths of Brownian motion are locally Hölder continuous of order $\alpha < 1/2$.

Definition. A continuous path $X_t = (X_t^{(1)}, \dots, X_t^{(n)})$ is called **nowhere differentiable**, if for all t, each coordinate function $X_t^{(i)}$ is not differentiable at t.

Theorem 4.2.5 (Wiener). Brownian motion is nowhere differentiable: for almost all ω, the path $t \mapsto X_t(\omega)$ is nowhere differentiable.

Proof. We follow [66]. It is enough to show it in one dimensions. Suppose B_t is differentiable at some point $0 \leq s \leq 1$. There exists then an integer l such that $|B_t - B_s| \leq l(t-s)$ for $t - s > 0$ small enough. But this means that

$$|B_{j/n} - B_{(j-1)/n}| \leq 7\frac{l}{n}$$

for all j satisfying

$$i = [ns] + 1 \leq j \leq [ns] + 4 = i + 3$$

and sufficiently large n so that the set of differentiable paths is included in the set

$$B = \bigcup_{l\geq 1} \bigcup_{m \geq 1} \bigcap_{n \geq m} \bigcup_{0<i\leq n+1} \bigcap_{i<j\leq i+3} \{|B_{j/n} - B_{(j-1)/n}| < 7\frac{l}{n}\} \, .$$

Using Brownian scaling, we show that $\mathrm{P}[B] = 0$ as follows

$$\begin{aligned} & \mathrm{P}[\bigcap_{n \geq m} \bigcup_{0<i\leq n+1} \bigcap_{i<j\leq i+3} \{|B_{j/n} - B_{(j-1)/n}| < 7\frac{l}{n} \}] \\ \leq \; & \liminf_{n\to\infty} n\mathrm{P}[|B_{1/n}| < 7\frac{l}{n}]^3 \\ = \; & \liminf_{n\to\infty} n\mathrm{P}[|B_1| < 7\frac{l}{\sqrt{n}}]^3 \\ \leq \; & \lim_{n\to\infty} \frac{C}{\sqrt{n}} = 0 \; . \end{aligned}$$

□

Remark. This proposition shows especially that we have no Lipshitz continuity of Brownian paths. A slight generalization shows that Brownian motion is not Hölder continuous for any $\alpha \geq 1/2$. One has just to do the same trick with k instead of 3 steps, where $k(\alpha - 1/2) > 1$. The actual modulus of continuity is very near to $\alpha = 1/2$: $|B_t - B_{t+\epsilon}|$ is of the order

$$h(\epsilon) = \sqrt{2\epsilon \log(\frac{1}{\epsilon})} \; .$$

More precisely, $\mathrm{P}[\limsup_{\epsilon\to 0} \sup_{|s-t|\leq\epsilon} \frac{|B_s - B_t|}{h(\epsilon)} = 1] = 1$, as we will see later in theorem (4.4.2).

The covariance of standard Brownian motion was given by $\mathrm{E}[B_s B_t] = \min\{s,t\}$. We constructed it by implementing the Hilbert space $L^2([0,\infty))$ as a Gaussian subspace of $\mathcal{L}^2(\Omega, \mathcal{A}, \mathrm{P})$. We look now at a more general class of Gaussian processes.

Definition. A function $V : T \times T \to \mathbb{R}$ is called **positive semidefinite**, if for all finite sets $\{t_1, \dots, t_d\} \subset T$, the matrix $V_{ij} = V(t_i, t_j)$ satisfies $(u, Vu) \geq 0$ for all vectors $u = (u_1, \dots, u_n)$.

Proposition 4.2.6. The covariance of a centered Gaussian process is positive semidefinite. Any positive semidefinite function V on $T \times T$ is the covariance of a centered Gaussian process X_t.

Proof. The first statement follows from the fact that for all $u = (u_1, \dots, u_n)$

$$\sum_{i,j} V(t_i, t_j) u_i u_j = \mathrm{E}[(\sum_{i=1}^{n} u_i X_{t_i})^2] \geq 0 \; .$$

We introduce for $t \in T$ a formal symbol δ_t. Consider the vector space of finite sums $\sum_{i=1}^n a_i \delta_{t_i}$ with inner product

$$(\sum_{i=1}^{d} a_i \delta_{t_i}, \sum_{j=1}^{d} b_j \delta_{t_j}) = \sum_{i,j} a_i b_j V(t_i, t_j) \; .$$

This is a positive semidefinite inner product. Multiplying out the null vectors $\{||v|| = 0\}$ and doing a completion gives a separable Hilbert space H. Define now as in the construction of Brownian motion the process $X_t = X(\delta_t)$. Because the map $X : H \to \mathcal{L}^2$ preserves the inner product, we have

$$\mathrm{E}[X_t, X_s] = (\delta_s, \delta_t) = V(s,t) .$$

□

Lets look at some examples of Gaussian processes:

Example. The **Ornstein-Uhlenbeck oscillator process** X_t is a one-dimensional process which is used to describe the quantum mechanical oscillator as we will see later. Let $T = \mathbb{R}^+$ and take the function $V(s,t) = \frac{1}{2}e^{-|t-s|}$ on $T \times T$. We first show that V is positive semidefinite: The Fourier transform of $f(t) = e^{-|t|}$ is

$$\int_{\mathbb{R}} e^{ikt} e^{-|t|} dt = \frac{1}{2\pi(k^2+1)} .$$

By Fourier inversion, we get

$$\frac{1}{2\pi} \int_{\mathbb{R}} (k^2+1)^{-1} e^{ik(t-s)} \, dk = \frac{1}{2} e^{-|t-s|} ,$$

and so

$$\begin{aligned} 0 &\leq (2\pi)^{-1} \int_{\mathbb{R}} (k^2+1)^{-1} \sum_j |u_j e^{ikt_j}|^2 \, dk \\ &= \sum_{j,k=1}^{n} u_j u_k \frac{1}{2} e^{-|t_j - t_k|} . \end{aligned}$$

This process has a continuous modification because

$$\mathrm{E}[(X_t - X_s)^2] = (e^{-|t-t|} + e^{-|s-s|} - 2e^{-|t-s|})/2 = (1 - e^{-|t-s|}) \leq |t-s|$$

and Kolmogorov's criterion. The Ornstein-Uhlenbeck is also called the **oscillatory process.**

Proposition 4.2.7. Brownian motion B_t and the Ornstein-Uhlenbeck process O_t are for $t \geq 0$ related by

$$O_t = \frac{1}{\sqrt{2}} e^{-t} B_{e^{2t}} .$$

Proof. Denote by O the Ornstein-Uhlenbeck process and let

$$X_t = 2^{-1/2} e^{-t} B_{e^{2t}} .$$

We want to show that $X = Y$. Both X and O are centered Gaussian, continuous processes with independent increments. To verify that they are the same, we have to show that they have the same covariance. This is a computation:

$$\mathrm{E}[O_t O_s] = \frac{1}{2} e^{-t} e^{-s} \min\{e^{2t}, e^{2s}\} = e^{|s-t|}/2 .$$

□

It follows from this relation that also the Ornstein-Uhlenbeck process is not differentiable almost everywhere. There are also generalized Ornstein-Uhlenbeck processes. The case $V(s,t) = \int_{\mathbb{R}} e^{-ik(t-s)}\, d\mu(k) = \hat{\mu}(t-s)$ with the Cauchy measure $\mu = \frac{1}{2\pi(k^2+1)} dx$ on $\mathbb{R}$ can be generalized to take any symmetric measure μ on $\mathbb{R}$ and let $\hat{\mu}$ denote its Fourier transform $\int_{\mathbb{R}} e^{-ikt}\, d\mu(k)$. The same calculation as above shows that the function $V(s,t) = \hat{\mu}(t-s)$ is positive semidefinite.

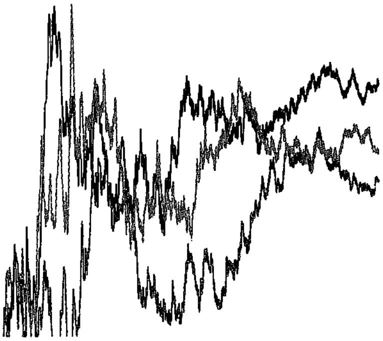


Figure. *Three paths of the Ornstein-Uhlenbeck process.*

Example. **Brownian bridge** is a one-dimensional process with time $T = [0,1]$ and $V(s,t) = s(1-t)$ for $1 \leq s \leq t \leq 1$ and $V(s,t) = V(t,s)$ else. It is also called **tied down process**.

In order to show that V is positive semidefinite, one observes that $X_t = B_s - sB_1$ is a Gaussian process, which has the covariance

$$\mathrm{E}[X_s X_t] = \mathrm{E}[(B_s - sB_1)(B_t - tB_1)] = s + st - 2st = s(1-t) .$$

Since $\mathrm{E}[X_1^2] = 0$, we have $X_1 = 0$ which means that all paths start from 0 at time 0 and end at 1 at time 1.

The realization $X_t = B_s - sB_1$ shows also that X_t has a continuous realization.

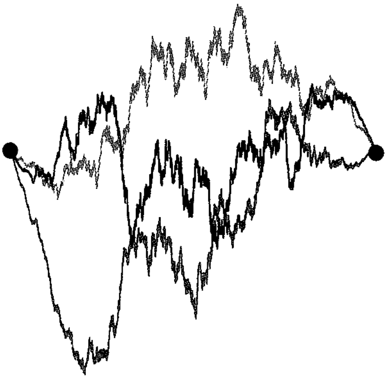


Figure. *Three paths of Brownian bridge.*

Let X_t be the Brownian bridge and let y be a point in $\mathbb{R}^d$. We can consider the Gaussian process $Y_t = ty + X_t$ which describes paths going from 0 at time 0 to y at time 1. The process Y has however no more zero mean. Brownian motion B and Brownian bridge X are related to each other by the formulas:

$$B_t = \tilde{B}_t := (t+1)X_{t/(t+1)}, \; X_t = \tilde{X}_t := (1-t)B_{t/(1-t)} \; .$$

These identities follow from the fact that both are continuous centered Gaussian processes with the right covariance:

$$\begin{aligned} \mathrm{E}[\tilde{B}_s\tilde{B}_t] &= (t+1)(s+1)\min\{\frac{t}{(t+1)}, \frac{s}{(s+1)}\} = \min\{s,t\} = \mathrm{E}[B_sB_t] \; , \\ \mathrm{E}[\tilde{X}_s\tilde{X}_t] &= (1-t)(1-s)\min\{\frac{s}{(1-s)}, \frac{t}{(1-t)}\} = s(1-t) = \mathrm{E}[X_sX_t] \end{aligned}$$

and uniqueness of Brownian motion.

Example. If $V(s,t) = 1_{\{s=t\}}$, we get a Gaussian process which has the property that X_s and X_t are independent, if $s \neq t$. Especially, there is no autocorrelation between different X_s and X_t. This process is called **white noise** or **"great disorder"**. It can not be modified so that $(t,\omega) \mapsto X_t(\omega)$ is measurable: if $(t,\omega) \mapsto X_t(\omega)$ were measurable, then $Y_t = \int_0^t X_s \, ds$ would be measurable too. But then

$$\mathrm{E}[Y_t^2] = \mathrm{E}[(\int_0^t X_s)^2] = \int_0^t \int_0^{t'} \mathrm{E}[X_{s'}X_{s'}] \, ds' \, ds = 0$$

which implies $Y_t = 0$ almost everywhere so that the measure $d\mu(\omega) = X_s(\omega) \, ds$ is zero for almost all ω.

$$t = \mathrm{E}[\int_0^t X_s^2] = \mathrm{E}[\int_0^t X_sX_s \, ds] = \mathrm{E}[\int_0^t X_s \, d\mu(s)] = 0 \; .$$

In a distributional sense, one can see Brownian motion as a solution of the stochastic differential equation and white noise as a generalized mean-square derivative of Brownian motion. We will look at stochastic differential equations later.

Example. Brownian sheet is not a stochastic process with one dimensional time but a **random field**: time $T = \mathbb{R}_+^2$ is two dimensional. Actually, as long as we deal only with Gaussian random variables and do not want to tackle regularity questions, the time T can be quite arbitrary and proposition (4.2.6) stated at the beginning of this section holds true. The Gaussian process with

$$V((s_1, s_2), (t_1, t_2)) = \min(s_1, t_1) \cdot \min(s_2, t_2)$$

is called Brownian sheet. It has similar scaling properties as Brownian motion.

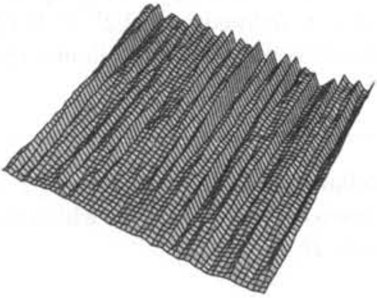


Figure. *Illustrating a sample of a Brownian sheet $B_{t,s}$. Time is two dimensional. Every trace $B_t = B_{t,s_0}$ or $B_t = B_{t,s_0}$ is standard Brownian motion.*

4.3 The Wiener measure

Let $(E, \mathcal{E})$ be a measurable space and let T be a set called "time". A stochastic process on a probability space $(\Omega, \mathcal{A}, \mathrm{P})$ indexed by T and with values in E defines a map

$$\phi : \Omega \to E^T, \omega \mapsto X_t(\omega) .$$

The product space E^T is equipped with the product σ-algebra $\mathcal{E}^T$, which is the smallest algebra for which all the functions X_t are measurable which is the σ-algebra generated by the π-system

$$\{ \prod_{t_1,\dots,t_n}^{n} A_{t_i} = \{x \in E^T, x_{t_i} \in A_{t_i}\} \mid A_{t_i} \in \mathcal{E}\}$$

consisting of cylinder sets. Denote by $Y_t(w) = w(t)$ the coordinate maps on E^T. Because $Y_t \circ \phi$ is measurable for all t, also ϕ is measurable. Denote by P_X the **push-forward measure** of ϕ from $(\Omega, \mathcal{A}, \mathrm{P})$ to $(E^T, \mathcal{E}^T)$ defined by $\mathrm{P}_X[A] = \mathrm{P}[X^{-1}(A)]$. For any finite set $(t_1, \dots, t_n) \subset T$ and all sets $A_i \in \mathcal{E}$, we have

$$\mathrm{P}[X_{t_i} \in A_i, i = 1, \dots, n] = \mathrm{P}_X[Y_{t_i} \in A_i, 1 = 1, \dots n] .$$

One says, the two processes X and Y are **versions of each other**.

Definition. Y is called the **coordinate process** of X and the probability measure P_X is called the **law of** X.

Definition. Two processes X, X' possibly defined on different probability spaces are called **versions of each other** if they have the same law $P_X = P_{X'}$.

One usually does not work with the coordinate process but prefers to work with processes which have some continuity properties. Many processes have versions which are **right continuous** and have left hand limits at every point.

Definition. Let D be a measurable subset of E^T and assume the process has a version X such that almost all paths $X(\omega)$ are in D. Define the probability space $(D, \mathcal{E}^T \cap D, Q)$, where Q is the measure $Q = \phi^*\mathrm{P}$. Obviously, the process Y defined on $(D, \mathcal{E}^T \cap D, Q)$ is another version of X. If D is right continuous with left hand limits, the process is called the **canonical version** of X.

Corollary 4.3.1. Let $\mathrm{E} = \mathbb{R}^d$ and $T = \mathbb{R}^+$. There exists a unique probability measure W on $C(T, E)$ for which the coordinate process Y is the Brownian motion B.

Proof. Let $D = C(T, E) \subset E^T$. Define the measure $W = \phi^* P_X$ and let Y be the coordinate process of B. Uniqueness: assume we have two such measures W, W' and let Y, Y' be the coordinate processes of B on D with respect to W and W'. Since both Y and Y' are versions of X and "being a version" is an equivalence relation, they are also versions of each other. This means that W and W' coincide on a π- system and are therefore the same. □

Definition. If $\mathrm{E} = \mathbb{R}^d$ and $T = [0, \infty)$, the measure W on $C(T, E)$ is called the **Wiener measure**. The probability space $(C(T, E), \mathcal{E}^T \cap C(T, E), W)$ is called the **Wiener space**.

Let $\mathcal{B}'$ be the σ-algebra $\mathcal{E}^T \cap C(T, E)$, which is the Borel σ-algebra restricted to $C(T, E)$. The space $C(T, E)$ carries an other σ-algebra, namely the Borel σ-algebra $\mathcal{B}$ generated by its own topology. We have $\mathcal{B} \subset \mathcal{B}'$, since all closed balls $\{f \in C(T, E) \mid |f - f_0| \leq r\} \in \mathcal{B}$ are in $\mathcal{B}'$. The other relation $\mathcal{B}' \subset \mathcal{B}$ is clear so that $\mathcal{B} = \mathcal{B}'$. The Wiener measure is therefore a **Borel measure**.

Remark. The Wiener measure can also be constructed without Brownian motion and can be used to define Brownian motion. We sketch the idea. Let $S = \dot{\mathbb{R}}^n$ denote the one point compactification of $\mathbb{R}^n$. Define $\Omega = S^{[0,t]}$

be the set of functions from $[0,t]$ to S which is also the set of paths in $\overline{\mathbb{R}^n}$. It is by Tychonov a compact space with the product topology. Define

$$C_{fin}(\Omega) = \{\phi \in C(\Omega, \mathbb{R}) \mid \exists F : \mathbb{R}^n \to \mathbb{R}, \phi(\omega) = F(\omega(t_1), \dots, \omega(t_n))\} .$$

Define also the **Gauss kernel** $p(x,y,t) = (4\pi t)^{-n/2} \exp(-|x-y|^2/4t)$. Define on $C_{fin}(\Omega)$ the functional

$$\begin{aligned}(L\phi)(s_1, \dots, s_m) &= \int_{(\mathbb{R}^n)^m} F(x_1, x_2, \dots, x_m) p(0, x_1, s_1) p(x_1, x_2, s_2) \\ &\quad \cdots p(x_{m-1}, x_m, s_m) \, dx_1 \cdots dx_m\end{aligned}$$

with $s_1 = t_1$ and $s_k = t_k - t_{k-1}$ for $k \geq 2$. Since $L(\phi) \leq |\phi(\omega)|_\infty$, it is a bounded linear functional on the dense linear subspace $C_{fin}(\Omega) \subset C(\Omega)$. It is nonnegative and $L(1) = 1$. By the Hahn Banach theorem, it extends uniquely to a bounded linear functional on $C(\Omega)$. By the Riesz representation theorem, there exists a unique measure μ on $C(\Omega)$ such that $L(\phi) = \int \phi(\omega) \, d\mu(\omega)$. This is the Wiener measure on Ω.

4.4 Lévy's modulus of continuity

We start with an elementary estimate

Lemma 4.4.1.

$$\frac{1}{a} e^{-a^2/2} > \int_a^\infty e^{-x^2/2} \, dx > \frac{a}{a^2+1} e^{-a^2/2} .$$

Proof.

$$\int_a^\infty e^{-x^2/2} \, dx < \int_a^\infty e^{-x^2/2} (x/a) \, dx = \frac{1}{a} e^{-a^2/2} .$$

For the right inequality consider

$$\int_a^\infty \frac{1}{b^2} e^{-b^2/2} \, db < \frac{1}{a^2} \int_a^\infty e^{-x^2/2} \, dx .$$

Integrating by parts of the left hand side of this gives

$$\frac{1}{a} e^{-a^2/2} - \int_a^\infty e^{-x^2/2} \, dx < \frac{1}{a^2} \int_a^\infty e^{-x^2/2} \, dx .$$

□

Theorem 4.4.2 (Lévy's modulus of continuity). If B is standard Brownian motion, then

$$\mathrm{P}[\limsup_{\epsilon \to 0} \sup_{|s-t| \leq \epsilon} \frac{|B_s - B_t|}{h(\epsilon)} = 1] = 1 ,$$

where $h(\epsilon) = \sqrt{2\epsilon \log(1/\epsilon)}$.

Proof. We follow [83]:

(i) Proof of the inequality "≥ 1".
Take $0 < \delta < 1$. Define $a_n = (1-\delta)h(2^{-n}) = (1-\delta)\sqrt{n 2 \log 2}$. Consider

$$\mathrm{P}[A_n] = \mathrm{P}[\max_{1 \leq k \leq 2^n} |B_{k2^{-n}} - B_{(k-1)2^{-n}}| \leq a_n] .$$

Because $B_{k/2^n} - B_{(k-1)/2^n}$ are independent Gaussian random variables, we compute, using the above lemma (4.4.1) and $1 - s < e^{-s}$

$$\begin{aligned}
\mathrm{P}[A_n] &\leq (1 - 2\int_{a_n}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \, dx)^{2^n} \\
&\leq (1 - 2\frac{a_n}{a_n^2+1} e^{-a_n^2/2})^{2^n} \\
&\leq \exp(-2^n \frac{2a_n}{a_n^2+1} e^{-a_n^2/2}) \leq e^{-C\exp(n(1-(1-\delta)^2)/\sqrt{n})} ,
\end{aligned}$$

where C is a constant independent of n. Since $\sum_n \mathrm{P}[A_n] < \infty$, we get by the first Borel-Cantelli that $\mathrm{P}[\limsup_n A_n] = 0$ so that

$$\mathrm{P}[\lim_{n \to \infty} \max_{1 \leq k \leq 2^n} |B_{k2^{-n}} - B_{(k-1)2^n}| \geq h(2^{-n})] = 1 .$$

(ii) Proof of the inequality "≤ 1".
Take again $0 < \delta < 1$ and pick $\epsilon > 0$ such that $(1+\epsilon)(1-\delta) > (1+\delta)$. Define

$$\begin{aligned}
\mathrm{P}[A_n] &= \mathrm{P}[\max_{k=j-i \in K} |B_{j2^{-n}} - B_{i2^{-n}}|/h(k2^{-n}) \geq (1+\epsilon)] \\
&= \mathrm{P}[\bigcup_{k=j-i \in K} \{|B_{j2^{-n}} - B_{i2^{-n}}|] \geq a_{n,k}\} ,
\end{aligned}$$

where

$$K = \{0 < k \leq 2^{n\delta} \}$$

and $a_{n,k} = h(k2^{-n})(1+\epsilon)$.
Using the above lemma, we get with some constants C which may vary

from line to line:

$$\begin{aligned} \mathrm{P}[A_n] &\leq \sum_{k \in K} a_{n,k}^{-1} e^{-a_{n,k}^2/2} \\ &\leq C \cdot \sum_{k \in K} \log(k^{-1}2^n)^{-1/2} e^{-(1+\epsilon)^2 \log(k^{-1}2^n)} \\ &\leq C \cdot 2^{-n(1-\delta)(1+\epsilon)^2} \sum_{k \in K} (\log(k^{-1}2^n))^{-1/2} \qquad (\text{ since } k^{-1} > 2^{-n\delta}) \\ &\leq C \cdot n^{-1/2} 2^{n(\delta-(1-\delta)(1+\epsilon)^2)} . \end{aligned}$$

In the last step was used that there are at most $2^{n\delta}$ points in K and for each of them $\log(k^{-1}2^n) > \log(2^n(1-\delta))$.
We see that $\sum_n \mathrm{P}[A_n]$ converges. By Borel-Cantelli we get for almost every ω an integer $n(\omega)$ such that for $n > n(\omega)$

$$|B_{j2^{-n}} - B_{i2^{-n}}| < (1+\epsilon) \cdot h(k2^{-n}) ,$$

where $k = j - i \in K$. Increase possibly $n(\omega)$ so that for $n > n(\omega)$

$$\sum_{m>n} h(2^{-m}) < \epsilon \cdot h(2^{-(n+1)(1-\delta)}) .$$

Pick $0 \leq t_1 < t_2 \leq 1$ such that $t = t_2 - t_1 < 2^{-n(\omega)(1-\delta)}$. Take next $n > n(\omega)$ such that $2^{-(n+1)(1-\delta)} \leq t < 2^{-n(1-\delta)}$ and write the dyadic development of t_1, t_2:

$$t_1 = i2^{-n} - 2^{-p_1} - 2^{-p_2} \dots , t_2 = j2^{-n} + 2^{-q_1} + 2^{-q_2} \dots$$

with $t_1 \leq i2^{-n} < j2^{-n} \leq t_2$ and $0 < k = j - i \leq t2^n < 2^{n\delta}$. We get

$$\begin{aligned} |B_{t_1}(\omega) - B_{t_2}(\omega)| &\leq |B_{t_1} - B_{i2^{-n}}(\omega)| \\ &\quad +|B_{i2^{-n}}(\omega) - B_{j2^{-n}}(\omega)| \\ &\quad +|B_{j2^{-n}}(\omega) - B_{t_2}| \\ &\leq 2\sum_{p>n}(1+\epsilon)h(2^{-p}) + (1+\epsilon)h(k2^{-n}) \\ &\leq (1+3\epsilon+2\epsilon^2)h(t) . \end{aligned}$$

Because $\epsilon > 0$ was arbitrary, the proof is complete. □

4.5 Stopping times

Stopping times are useful for the construction of new processes, in proofs of inequalities and convergence theorems as well as in the study of return time results. A good source for stopping time results and stochastic process in general is [83].

Definition. A **filtration** of a measurable space $(\Omega, \mathcal{A})$ is an increasing family $(\mathcal{A}_t)_{t \geq 0}$ of sub-σ-algebras of $\mathcal{A}$. A measurable space endowed with a filtration $(\mathcal{A}_t)_{t \geq 0}$ is called a **filtered space**. A process X is called **adapted** to the filtration $\mathcal{A}_t$, if X_t is $\mathcal{A}_t$-measurable for all t.

Definition. A process X on $(\Omega, \mathcal{A}, \mathrm{P})$ defines a natural filtration $\mathcal{A}_t = \sigma(X_s \mid s \leq t)$, the **minimal filtration** of X for which X is adapted. Heuristically, $\mathcal{A}_t$ is the set of events, which may occur up to time t.

Definition. With a filtration we can associate two other filtration by setting for $t > 0$

$$\mathcal{A}_{t^-} = \sigma(\mathcal{A}_s, s < t), \mathcal{A}_{t^+} = \bigcap_{s>t} \mathcal{A}_s \ .$$

For $t = 0$ we can still define $\mathcal{A}_{0^+} = \bigcap_{s>0} \mathcal{A}_s$ and define $\mathcal{A}_{0^-} = \mathcal{A}_0$. Define also $\mathcal{A}_\infty = \sigma(\mathcal{A}_s, s \geq 0)$.

Remark. We always have $\mathcal{A}_{t^-} \subset \mathcal{A}_t \subset \mathcal{A}_{t^+}$ and both inclusions can be strict.

Definition. If $\mathcal{A}_t = \mathcal{A}_{t^+}$ then the filtration $\mathcal{A}_t$ is called **right continuous**. If $\mathcal{A}_t = \mathcal{A}_{t^-}$, then $\mathcal{A}_t$ is **left continuous**. As an example, the filtration $\mathcal{A}_{t^+}$ of any filtration is right continuous.

Definition. A **stopping time** relative to a filtration $\mathcal{A}_t$ is a map $T : \Omega \to [0, \infty]$ such that $\{T \leq t\ \} \in \mathcal{A}_t$.

Remark. If $\mathcal{A}_t$ is right continuous, then T is a stopping time if and only if $\{T < t\ \} \in \mathcal{A}_t$. Also T is a stopping time if and only if $X_t = 1_{(0,T]}(t)$ is adapted. X is then a left continuous adapted process.

Definition. If T is a stopping time, define

$$\mathcal{A}_T = \{A \in \mathcal{A}_\infty \mid A \cap \{T \leq t\} \in \mathcal{A}_t, \forall t\} \ .$$

It is a σ-algebra. As an example, if $T = s$ is constant, then $\mathcal{A}_T = \mathcal{A}_s$. Note also that

$$\mathcal{A}_{T^+} = \{A \in \mathcal{A}_\infty \mid A \cap \{T < t\} \in \mathcal{A}_t, \forall t\} \ .$$

We give examples of stopping times.

Proposition 4.5.1. Let X be the coordinate process on $C(\mathbb{R}_+, E)$, where E is a metric space. Let A be a closed set in E. Then the so called **entry time**

$$T_A(\omega) = \inf\{t \geq 0 \mid X_t(\omega) \in A\ \}$$

is a stopping time relative to the filtration $\mathcal{A}_t = \sigma(\{X_s\ \}_{s \leq t})$.

Proof. Let d be the metric on E. We have

$$\{T_A \leq t\} = \{\inf_{s \in \mathbb{Q}, s \leq t} d(X_s(\omega), A) = 0\ \}$$

which is in $\mathcal{A}_t = \sigma(X_s, x \leq t)$. □

Proposition 4.5.2. Let X be the coordinate process on $D(\mathbb{R}_+, E)$, the space of right continuous functions, where E is a metric space. Let A be an open subset of E. Then the **hitting time**

$$T_A(\omega) = \inf\{t > 0 \mid X_t(\omega) \in A\}$$

is a stopping time with respect to the filtration $\mathcal{A}_{t+}$.

Proof. T_A is a $\mathcal{A}_{t+}$ stopping time if and only if $\{T_A < t\} \in \mathcal{A}_t$ for all t. If A is open and $X_s(\omega) \in A$, we know by the right-continuity of the paths that $X_t(\omega) \in A$ for every $t \in [s, s+\epsilon)$ for some $\epsilon > 0$. Therefore

$$\{T_A < t\} = \{\inf_{s \in \mathbb{Q}, s<t} X_s \in A\} \in \mathcal{A}_t .$$

□

Definition. Let $\mathcal{A}_t$ be a filtration on $(\Omega, \mathcal{A})$ and let T be a stopping time. For a process X, we define a new map X_T on the set $\{T < \infty\}$ by

$$X_T(\omega) = X_{T(\omega)}(\omega) .$$

Remark. We have met this definition already in the case of discrete time but in the present situation, it is not clear whether X_T is measurable. It turns out that this is true for many processes.

Definition. A process X is called **progressively measurable** with respect to a filtration $\mathcal{A}_t$ if for all t, the map $(s, \omega) \mapsto X_s(\omega)$ from $([0,t] \times \Omega, \mathcal{B}([0,t] \times \mathcal{A}_t)$ into $(E, \mathcal{E})$ is measurable.

A progressively measurable process is adapted. For some processes, the inverse holds:

Lemma 4.5.3. An adapted process with right or left continuous paths is progressively measurable.

Proof. Assume right continuity (the argument is similar in the case of left continuity). Write X as the coordinate process $D([0,t], E)$. Denote the map $(s, \omega) \mapsto X_s(\omega)$ with $Y = Y(s, \omega)$. Given a closed ball $U \in \mathcal{E}$. We have to show that $Y^{-1}(U) = \{(s, \omega) \mid Y(s, \omega) \in U\} \in \mathcal{B}([0,t]) \times \mathcal{A}_t$. Given $k = \mathbb{N}$, we define $\mathrm{E}_{0,U} = 0$ and inductively for $k \geq 1$ the k'th hitting time (a stopping time)

$$H_{k,U}(\omega) = \inf\{s \in \mathbb{Q} \mid \mathrm{E}_{k-1,U}(\omega) < s < t,\ X_s \in U\}$$

as well as the k'th exit time (not necessarily a stopping time)

$$\mathrm{E}_{k,U}(\omega) = \inf\{s \in \mathbb{Q} | H_{k,U}(\omega) < s < t,\ X_s \notin U\} .$$

These are countably many measurable maps from $D([0,t],E)$ to $[0,t]$. Then by the right-continuity

$$Y^{-1}(U) = \bigcup_{k=1}^{\infty} \{(s,\omega) \mid H_{k,U}(\omega) \leq s \leq \mathrm{E}_{k,U}(\omega)\}$$

which is in $\mathcal{B}([0,t]) \times \mathcal{A}_t$. □

Proposition 4.5.4. If X is progressively measurable and T is a stopping time, then X_T is $\mathcal{A}_T$-measurable on the set $\{T < \infty\}$.

Proof. The set $\{T < \infty\}$ is itself in $\mathcal{A}_T$. To say that X_T is $\mathcal{A}_T$- measurable on this set is equivalent with $X_T \cdot 1_{\{T \leq t\}} \in \mathcal{A}_t$ for every t. But the map

$$S : (\{T \leq t\}, \mathcal{A}_t \cap \{T \leq t\}) \to ([0,t], \mathcal{B}[0,t])$$

is measurable because T is a stopping time. This means that the map $\omega \mapsto (T(\omega), \omega)$ from $(\Omega, \mathcal{A}_t)$ to $([0,t] \times \Omega, \mathcal{B}([0,t]) \times \mathcal{A}_t)$ is measurable and X_T is the composition of this map with X which is $\mathcal{B}[0,t] \times \mathcal{A}_t$ measurable by hypothesis. □

Definition. Given a stopping time T and a process X, we define the **stopped process** $(X_T)_t(\omega) = X_{T \wedge t}(\omega)$.

Remark. If $\mathcal{A}_t$ is a filtration then $\mathcal{A}_{t \wedge T}$ is a filtration since if T_1 and T_2 are stopping times, then $T_1 \wedge T_2$ is a stopping time.

Corollary 4.5.5. If X is progressively measurable with respect to $\mathcal{A}_t$ and T is a stopping time, then $(X_T)_t = X_{t \wedge T}$ is progressively measurable with respect to $\mathcal{A}_{t \wedge T}$.

Proof. Because $t \wedge T$ is a stopping time, we have from the previous proposition that X_T is $\mathcal{A}_{t \wedge T}$ measurable.
We know by assumption that $\phi : (s,\omega) \mapsto X_s(\omega)$ is measurable. Since also $\psi : (s,\omega) \mapsto (s \wedge T)(\omega)$ is measurable, we know also that the composition $(s,\omega) \mapsto X_T(\omega) = X_{\psi(s,\omega)}(\omega) = \phi(\psi(s,\omega),\omega)$ is measurable. □

Proposition 4.5.6. Every stopping time is the decreasing limit of a sequence of stopping times taking only finitely many values.

Proof. Given a stopping time T, define the discretisation $T_k = +\infty$ if $T \geq k$ and $T_k = q2^{-k}$ if $(q-1)2^{-k} \leq T < q\, 2^{-k}$, $q < 2^k k$. Each T_k is a stopping time and T_k decreases to T. □

Many concepts of classical potential theory can be expressed in an elegant form in a probabilistic language. We give very briefly some examples without proofs, but some hints to the literature.

Let B_t be Brownian motion in $\mathbb{R}^d$ and T_A the hitting time of a set $A \subset \mathbb{R}^d$. Let D be a domain in $\mathbb{R}^d$ with boundary $\delta(D)$ such that the **Green function** $G(x, y)$ exists in D. Such a domain is then called a **Green domain.**

Definition. The **Green function** of a domain D is defined as the fundamental solution satisfying $\Delta G(x, y) = \delta(x-y)$, where $\delta(x-y)$ is the Dirac measure at $y \in D$. Having the fundamental solution G, we can solve the Poisson equation $\Delta u = v$ for a given function v by

$$u = \int_D G(x, y) \cdot v(y)\, dy\ .$$

The Green function can be computed using Brownian motion as follows:

$$G(x, y) = \int_0^\infty g(t, x, y)\, dt\ ,$$

where for $x \in D$,

$$\int_C g(t, x, y)\, dy = \mathrm{P}_x[B_t \in C, T_{\delta D} > t]$$

and P_x is the Wiener measure of B_t starting at the point x.

We can interpret that as follows. To determine $G(x, y)$, consider the killed Brownian motion B_t starting at x, where T is the hitting time of the boundary. $G(x, y)$ is then the probability density, of the particles described by the Brownian motion.

Definition. The classical **Dirichlet problem** for a bounded Green domain $D \in \mathbb{R}^d$ with boundary δD is to find for a given function $f \in C(\delta(D))$, a solution $u \in C(\overline{D})$ such that $\Delta u = 0$ inside D and

$$\lim_{x \to y, x \in D} u(x) = f(y)$$

for every $y \in \delta D$.

This problem can not be solved in general even for domains with piecewise smooth boundaries if $d \geq 3$.

Definition. The following example is called **Lebesgue thorn** or **Lebesgue spine** has been suggested by Lebesgue in 1913. Let D be the inside of a spherical chamber in which a thorn is punched in. The boundary δD is held on constant temperature f, where $f = 1$ at the tip of the thorn y and zero except in a small neighborhood of y. The temperature u inside D is a solution of the Dirichlet problem $\Delta_D u = 0$ satisfying the boundary condition $u = f$ on the boundary δD. But the heat radiated from the thorn is proportional to its surface area. If the tip is sharp enough, a person sitting in the chamber will be cold, no matter how close to the heater. This means $\liminf_{x \to y, x \in D} u(x) < 1 = f(y)$. (For more details, see [43, 46]).

Because of this problem, one has to modify the question and one says, u is a **solution of a modified Dirichlet problem**, if u satisfies $\Delta_D u = 0$ inside D and $\lim_{x \to y, x \in D} u(x) = f(y)$ for all nonsingular points y in the boundary δD. Irregularity of a point y can be defined analytically but it is equivalent with $\mathrm{P}_y[T_{D^c} > 0] = 1$, which means that almost every Brownian particle starting at $y \in \delta D$ will return to δD after positive time.

Theorem 4.5.7 (Kakutani 1944). The solution of the regularized Dirichlet problem can be expressed with Brownian motion B_t and the hitting time T of the boundary:

$$u(x) = \mathrm{E}_x[f(B_T)] .$$

In words, the solution $u(x)$ of the Dirichlet problem is the expected value of the boundary function f at the exit point B_T of Brownian motion B_t starting at x. We have seen in the previous chapter that the discretized version of this result on a graph is quite easy to prove.

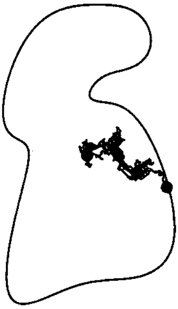


Figure. *To solve the Dirichlet problem in a bounded domain with Brownian motion, start the process at the point x and run it until it reaches the boundary B_T, then compute $f(B_T)$ and average this random variable over all paths ω.*

Remark. Ikeda has discovered that there exists also a probabilistic method for solving the classical von Neumann problem in the case $d = 2$. For more information about this, one can consult [43, 79]. The process for the von Neumann problem is not the process of **killed Brownian motion**, but the process of **reflected Brownian motion**.

Remark. Given the Dirichlet Laplacian Δ of a bounded domain D. One can compute the **heat flow** $e^{-t\Delta}u$ by the following formula

$$(e^{-t\Delta}u)(x) = \mathrm{E}_x[u(B_t); t < T] ,$$

where T is the hitting time of δD for Brownian motion B_t starting at x.

Remark. Let K be a compact subset of a Green domain D. The hitting probability

$$p(x) = \mathrm{P}_x[T_K < T_{\delta D}]$$

is the equilibrium potential of K relative to D. We give a definition of the equilibrium potential later. Physically, the equilibrium potential is obtained by measuring the electrostatic potential, if one is grounding the conducting boundary and charging the conducting set B with a unit amount of charge.

4.6 Continuous time martingales

Definition. Given a filtration $\mathcal{A}_t$ of the probability space $(\Omega, \mathcal{A}, \mathrm{P})$. A real-valued process $X_t \in \mathcal{L}^1$ which is $\mathcal{A}_t$ adapted is called a **submartingale**, if $\mathrm{E}[X_t|\mathcal{A}_s] \geq X_s$, it is called a **supermartingale** if $-X$ is a submartingale and a **martingale**, if it is both a super and sub-martingale. If additionally $X_t \in \mathcal{L}^p$ for all t, we speak of $\mathcal{L}^p$ super or sub-martingales.

We have seen martingales for discrete time already in the last chapter. Brownian motion gives examples with continuous time.

Proposition 4.6.1. Let B_t be standard Brownian motion. Then B_t, $B_t^2 - t$ and $e^{\alpha B_t - \alpha^2 t/2}$ are martingales.

Proof. $B_t - B_s$ is independent of B_s. Therefore

$$\mathrm{E}[B_t \mid \mathcal{A}_s] - B_s = \mathrm{E}[B_t - B_s|\mathcal{A}_s] = \mathrm{E}[B_t - B_s] = 0 .$$

Since by the "extracting knowledge" property

$$\mathrm{E}[B_t B_s \mid \mathcal{A}_s] = B_s \cdot \mathrm{E}[B_t \mid \mathcal{A}_s] = 0 ,$$

we get

$$\begin{aligned} \mathrm{E}[B_t^2 - t \mid \mathcal{A}_s] - (B_s^2 - s) &= \mathrm{E}[B_t^2 - B_s^2 \mid \mathcal{A}_s] - (t - s) \\ &= \mathrm{E}[(B_t - B_s)^2 \mid \mathcal{A}_s] - (t - s) = 0 . \end{aligned}$$

Since Brownian motion begins at any time s new, we have

$$\mathrm{E}[e^{\alpha(B_t-B_s)}|\mathcal{A}_s] = \mathrm{E}[e^{\alpha B_{t-s}}] = e^{\alpha^2(t-s)/2}$$

from which

$$\mathrm{E}[e^{\alpha B_t}|\mathcal{A}_s]e^{-\alpha^2 t/2} = \mathrm{E}[e^{\alpha B_s}]e^{-\alpha^2 s/2}$$

follows. □

As in the discrete case, we remark:

Proposition 4.6.2. If X_t is a $\mathcal{L}^p$-martingale, then $|X_t|^p$ is a submartingale for $p \geq 1$.

Proof. The conditional Jensen inequality gives

$$\mathrm{E}[|X_t|^p|\mathcal{A}_s] \geq |E[X_t|\mathcal{A}_s]|^p = |X_s|^p \, .$$

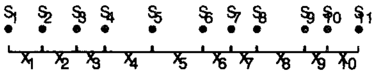
□

Example. Let X_n be a sequence of IID exponential distributed random variables with probability density $f_X(x) = e^{-cx}c$. Let $S_n = \sum_{k=1}^n X_k$. The Poisson process N_t with time $T = \mathbb{R}^+ = [0, \infty)$ is defined as

$$N_t = \sum_{k=1}^{\infty} 1_{S_k \leq t} \, .$$

It is an example of a martingale which is not continuous, This process takes values in $\mathbb{N}$ and measures, how many jumps are necessary to reach t. Since $\mathrm{E}[N_t] = ct$, it follows that $N_t - ct$ is a martingale with respect to the filtration $\mathcal{A}_t = \sigma(N_s, s \leq t)$. It is a right continuous process. We know therefore that it is progressively measurable and that for each stopping time T, also N^T is progressively measurable. See [49] or the last chapter for more information about Poisson processes.

Figure. *The Poisson point process on the line. N_t is the number of events which happen up to time t. It could model for example the number N_t of hits onto a website.*



Proposition 4.6.3. (Interval theorem) The Poisson process has independent increments

$$N_t - N_s = \sum_{n=1}^{\infty} 1_{s < S_n \leq t} \; .$$

Moreover, N_t is Poisson distributed with parameter tc:

$$\mathrm{P}[N_t = k] = \frac{(tc)^k}{k!} e^{-tc} \; .$$

Proof. The proof is done by starting with a Poisson distributed process N_t. Define then

$$S_n(\omega) = \{t \mid N_t = n, N_{t-0} = n-1\}$$

and show that $X_n = S_n - S_{n-1}$ are independent random variables with exponential distribution. □

Remark. Poisson processes on the lattice $\mathbb{Z}^d$ are also called **Brownian motion on the lattice** and can be used to describe Feynman-Kac formulas for **discrete Schrödinger operators**. The process is defined as follows: take X_t as above and define

$$Y_t = \sum_{k=1}^{\infty} Z_k 1_{S_k \leq t} \; ,$$

where Z_n are IID random variables taking values in $\{m \in \mathbb{Z}^d || m| = 1\}$. This means that a particle stays at a lattice site for an exponential time and jumps then to one of the neighbors of n with equal probability. Let P_n be the analog of the Wiener measure on right continuous paths on the lattice and denote with E_n the expectation. The **Feynman-Kac formula** for discrete Schrödinger operators $H = H_0 + V$ is

$$(e^{-itH}u)(n) = e^{2dt}\mathrm{E}_n[u(X_t) i^{N_t} e^{-i\int_0^t V(X_s)\, ds}] \; .$$

4.7 Doob inequalities

We have already established inequalities of Doob for discrete times $T = \mathbb{N}$. By a limiting argument, they hold also for right-continuous submartingales.

Theorem 4.7.1 (Doob's submartingale inequality). Let X be a non-negative right continuous submartingale with time $T = [a, b]$. For any $\epsilon > 0$

$$\epsilon \cdot \mathrm{P}[\sup_{a \leq t \leq b} X_t \geq \epsilon] \leq \mathrm{E}[X_b; \{\sup_{a \leq t \leq b} X_t \geq \epsilon\}] \leq \mathrm{E}[X_b] \; .$$

Proof. Take a countable subset D of T and choose an increasing sequence D_n of finite sets such that $\bigcup_n D_n = D$. We know now that for all n

$$\epsilon \cdot \mathrm{P}[\sup_{t\in D_n} X_t \geq \epsilon] \leq \mathrm{E}[X_b; \{\sup_{t\in D_n} X_t \geq \epsilon\}] \leq \mathrm{E}[X_b] .$$

since $\mathrm{E}[X_t]$ is nondecreasing in t. Going to the limit $n \to \infty$ gives the claim with $T = D$. Since X is right continuous, we get the claim for $T = [a, b]$. □

One often applies this inequality to the non-negative submartingale $|X|$ if X is a martingale.

Theorem 4.7.2 (Doob's L^p inequality). Fix $p > 1$ and q satisfying $p^{-1} + q^{-1} = 1$. Given a non-negative right-continuous submartingale X with time $T = [a, b]$ which is bounded in $\mathcal{L}^p$. Then $X^* = \sup_{t\in T} X_t$ is in $\mathcal{L}^p$ and satisfies

$$||X^*||_p \leq q \cdot \sup_{t\in T} ||X_t||_p .$$

Proof. Take a countable subset D of T and choose an increasing sequence D_n of finite sets such that $\bigcup_n D_n = D$.
We had

$$|| \sup_{t\in D_n} X_t|| \leq q \cdot \sup_{t\in D_n} ||X_t||_p .$$

Going to the limit gives

$$||\sup_{t\in D} X_t|| \leq q \cdot \sup_{t\in D} ||X_t||_p .$$

Since D is dense and X is right continuous we can replace D by T. □

The following inequality measures, how big is the probability that one-dimensional Brownian motion will leave the cone $\{(t, x), |x| \leq a \cdot t\}$.

Theorem 4.7.3 (Exponential inequality). $S_t = \sup_{0\leq s\leq t} B_s$ satisfies for any $a > 0$

$$\mathrm{P}[S_t \geq a \cdot t] \leq e^{-a^2 t/2} .$$

Proof. We have seen in proposition (4.6.1) that $M_t = e^{\alpha B_t - \frac{\alpha^2 t}{2}}$ is a martingale. It is nonnegative. Since

$$\exp(\alpha S_t - \frac{\alpha^2 t}{2}) \leq \exp(\sup_{s\leq t} B_s - \frac{\alpha^2 t}{2}) \leq \sup_{s\leq t} \exp(B_s - \frac{\alpha^2 s}{2}) = \sup_{s\leq t} M_s ,$$

we get with Doob's submartingale inequality (4.7.1)

$$\begin{aligned} \mathrm{P}[S_t \geq at] &\leq \mathrm{P}[\sup_{s \leq t} M_s \geq e^{\alpha a t - \frac{\alpha^2 t}{2}}] \\ &\leq \exp(-\alpha a t + \frac{\alpha^2 t}{2}) \mathrm{E}[M_t] \, . \end{aligned}$$

The result follows from $\mathrm{E}[B_t] = \mathrm{E}[B_0] = 1$ and $\inf_{\alpha > 0} \exp(-\alpha a t + \frac{\alpha^2 t}{2}) = \exp(-\frac{a^2 t}{2})$. □

An other corollary of Doob's maximal inequality will also be useful.

Corollary 4.7.4. For $a, b > 0$,

$$\mathrm{P}[\sup_{s \in [0,1]} (B_s - \frac{\alpha s}{2}) \geq \beta] \leq e^{-\alpha \beta} \, .$$

Proof.

$$\begin{aligned} \mathrm{P}[\sup_{s \in [0,1]} (B_s - \frac{\alpha s}{2}) \geq \beta] &\leq \mathrm{P}[\sup_{s \in [0,1]} (B_s - \frac{\alpha t}{2}) \geq \beta] \\ &= \mathrm{P}[\sup_{s \in [0,1]} (e^{\alpha B_s - \frac{\alpha^2 t}{2}}) \geq e^{\beta \alpha}] \\ &= \mathrm{P}[\sup_{s \in [0,1]} M_s \geq e^{\beta \alpha}] \\ &\leq e^{-\beta \alpha} \sup_{s \in [0,1]} \mathrm{E}[M_s] = e^{-\beta \alpha} \end{aligned}$$

since $\mathrm{E}[M_s] = 1$ for all s. □

4.8 Khintchine's law of the iterated logarithm

Khinchine's law of the iterated logarithm for Brownian motion gives a precise statement about how one-dimensional Brownian motion oscillates in a neighborhood of the origin. As in the law of the iterated logarithm, define

$$\Lambda(t) = \sqrt{2t \log |\log t|} \, .$$

Theorem 4.8.1 (Law of iterated logarithm for Brownian motion).

$$\mathrm{P}[\limsup_{t \to 0} \frac{B_t}{\Lambda(t)} = 1] = 1, \quad \mathrm{P}[\liminf_{t \to 0} \frac{B_t}{\Lambda(t)} = -1] = 1$$

Proof. The second statement follows from the first by changing B_t to $-B_t$.

(i) $\limsup_{s\to 0} \frac{B_s}{\Lambda(s)} \le 1$ almost everywhere:
Take $\theta, \delta \in (0,1)$ and define

$$\alpha_n = (1+\delta)\theta^{-n}\Lambda(\theta^n),\ \beta_n = \frac{\Lambda(\theta^n)}{2} .$$

We have $\alpha_n\beta_n = \log\log(\theta^n)(1+\delta) = \log(n)\log(\theta)$. From corollary (4.7.4), we get

$$\mathrm{P}[\sup_{s\le 1}(B_s - \frac{\alpha_n s}{2}) \ge \beta_n] \le e^{-\alpha_n\beta_n} = Kn^{(-1+\delta)} .$$

The Borel-Cantelli lemma assures

$$\mathrm{P}[\liminf_{n\to\infty} \sup_{s\le 1}(B_s - \frac{\alpha_n s}{2}) < \beta_n] = 1$$

which means that for almost every ω, there is $n_0(\omega)$ such that for $n > n_0(\omega)$ and $s \in [0, \theta^{n-1})$,

$$B_s(\omega) \le \alpha_n\frac{s}{2} + \beta_n \le \alpha_n\frac{\theta^{n-1}}{2} + \beta_n = (\frac{(1+\delta)}{2\theta} + \frac{1}{2})\Lambda(\theta^n) .$$

Since Λ is increasing on a sufficiently small interval $[0,a)$, we have for sufficiently large n and $s \in (\theta^n, \theta^{n-1}]$

$$B_s(\omega) \le (\frac{(1+\delta)}{2\theta} + \frac{1}{2})\Lambda(s) .$$

In the limit $\theta \to 1$ and $\delta \to 0$, we get the claim.

(ii) $\limsup_{s\to 0} \frac{B_s}{\Lambda(s)} \ge 1$ almost everywhere.
For $\theta \in (0,1)$, the sets

$$A_n = \{B_{\theta^n} - B_{\theta^{n+1}} \ge (1-\sqrt{\theta})\Lambda(\theta^n)\}$$

are independent and since $B_{\theta^n} - B_{\theta^{n+1}}$ is Gaussian we have

$$\mathrm{P}[A_n] = \int_a^\infty e^{-u^2/2}\,\frac{du}{\sqrt{2\pi}} > \frac{a}{a^2+1}e^{-a^2/2}$$

with $a = (1-\sqrt{\theta})\Lambda(\theta^n) \le Kn^{-\alpha}$ with some constants K and $\alpha < 1$. Therefore $\sum_n \mathrm{P}[A_n] = \infty$ and by the second Borel-Cantelli lemma,

$$B_{\theta^n} \ge (1-\sqrt{\theta})\Lambda(\theta^n) + B_{\theta^{n+1}} \tag{4.1}$$

for infinitely many n. Since $-B$ is also Brownian motion, we know from (i) that

$$-B_{\theta^{n+1}} < 2\Lambda(\theta^{n+1}) \tag{4.2}$$

for sufficiently large n. Using these two inequalities (4.1) and (4.2) and $\Lambda(\theta^{n+1}) \le 2\sqrt{\theta}\Lambda(\theta^n)$ for large enough n, we get

$$B_{\theta^n} > (1-\sqrt{\theta})\Lambda(\theta^n) - 4\Lambda(\theta^{n+1}) > \Lambda(\theta^n)(1-\sqrt{\theta}-4\sqrt{\theta})$$

for infinitely many n and therefore

$$\liminf_{t\to 0} \frac{B_t}{\Lambda(t)} \geq \limsup_{n\to\infty} \frac{B_{\theta^n}}{\Lambda(\theta^n)} > 1 - 5\sqrt{\theta}\,.$$

The claim follows for $\theta \to 0$. □

Remark. This statement shows also that B_t changes sign infinitely often for $t \to 0$ and that Brownian motion is recurrent in one dimension. One could show more, namely that the set $\{B_t = 0\}$ is a nonempty perfect set with Hausdorff dimension $1/2$ which is in particularly uncountable.

By time inversion, one gets the law of iterated logarithm near infinity:

Corollary 4.8.2.

$$\mathrm{P}[\limsup_{t\to\infty} \frac{B_t}{\Lambda(t)} = 1] = 1, \quad \mathrm{P}[\liminf_{t\to\infty} \frac{B_t}{\Lambda(t)} = -1] = 1\,.$$

Proof. Since $\tilde{B}_t = tB_{1/t}$ (with $\tilde{B}_0 = 0$) is a Brownian motion, we have with $s = 1/t$

$$\begin{aligned} 1 &= \limsup_{s\to 0} \frac{\tilde{B}_s}{\Lambda(s)} = \limsup_{s\to 0} s\frac{B_{1/s}}{\Lambda(s)} \\ &= \limsup_{t\to\infty} \frac{B_t}{t\Lambda(1/t)} = \limsup_{t\to\infty} \frac{B_t}{\Lambda(t)}\,. \end{aligned}$$

The other statement follows again by reflection. □

Corollary 4.8.3. For d-dimensional Brownian motion, one has

$$\mathrm{P}[\limsup_{t\to 0} \frac{B_t}{\Lambda(t)} = 1] = 1, \quad \mathrm{P}[\liminf_{t\to 0} \frac{B_t}{\Lambda(t)} = -1] = 1$$

Proof. Let e be a unit vector in $\mathbb{R}^d$. Then $B_t \cdot e$ is a 1-dimensional Brownian motion since B_t was defined as the product of d orthogonal Brownian motions. From the previous theorem, we have

$$\mathrm{P}[\limsup_{t\to 0} \frac{B_t \cdot e}{\Lambda(t)} = 1] = 1\,.$$

Since $B_t \cdot e \leq |B_t|$, we know that the lim sup is ≥ 1. This is true for all unit vectors and we can even get it simultaneously for a dense set $\{e_n\}_{n\in\mathbb{N}}$

of unit vectors in the unit sphere. Assume the lim sup is $1+\epsilon > 1$. Then, there exists e_n such that

$$\mathrm{P}[\limsup_{t \to 0} \frac{B_t \cdot e_n}{\Lambda(t)} \geq 1 + \frac{\epsilon}{2}] = 1$$

in contradiction to the law of iterated logarithm for Brownian motion. Therefore, we have lim sup $= 1$. By reflection symmetry, lim inf $= -1$. □

Remark. It follows that in d dimensions, the set of limit points of $B_t/\Lambda(t)$ for $t \to 0$ is the entire unit ball $\{|v| \leq 1\}$.

4.9 The theorem of Dynkin-Hunt

Definition. Denote by $I(k,n)$ the interval $[\frac{k-1}{2^n}, \frac{k}{2^n})$. If T is a stopping time, then $T^{(n)}$ denotes its discretisation

$$T^{(n)}(\omega) = \sum_{k=1}^{\infty} 1_{I(k,n)}(T(\omega))\frac{k}{2^n}$$

which is again a stopping time. Define also:

$$\mathcal{A}_{T^+} = \{A \in \mathcal{A}_\infty \mid A \cap \{T < t\} \in \mathcal{A}_t, \forall t\} .$$

The next theorem tells that Brownian motion starts afresh at stopping times.

Theorem 4.9.1 (Dynkin-Hunt). Let T be a stopping time for Brownian motion, then $\tilde{B}_t = B_{t+T} - B_T$ is Brownian motion when conditioned to $\{T < \infty\}$ and $\tilde{B}_t$ is independent of $\mathcal{A}_{T^+}$ when conditioned to $\{T < \infty\}$.

Proof. Let A be the set $\{T < \infty\}$. The theorem says that for every function

$$f(B_t) = g(B_{t+t_1}, B_{t+t_2}, \dots, B_{t+t_n})$$

with $g \in C(\mathbb{R}^n)$

$$\mathrm{E}[f(\tilde{B}_t)1_A] = \mathrm{E}[f(B_t)] \cdot \mathrm{P}[A]$$

and that for every set $C \in \mathcal{A}_{T^+}$

$$\mathrm{E}[f(\tilde{B}_t)1_{A \cap C}] \cdot \mathrm{P}[A] = \mathrm{E}[f(\tilde{B}_t)1_A] \cdot \mathrm{P}[A \cap C] .$$

This two statements are equivalent to the statement that for every $C \in \mathcal{A}_{T^+}$

$$\mathrm{E}[f(\tilde{B}_t) \cdot 1_{A \cap C}] = \mathrm{E}[f(B_t)] \cdot \mathrm{P}[A \cap C] .$$

Let $T^{(n)}$ be the discretisation of the stopping time T and $A_n = \{T^{(n)} < \infty\}$ as well as $A_{n,k} = \{T^{(n)} = k/2^n\}$. Using $A = \{T < \infty\}, \mathrm{P}[\bigcup_{k=1}^{\infty} A_{n,k} \cap C] \to \mathrm{P}[A \cap C]$ for $n \to \infty$, we compute

$$\begin{aligned}
\mathrm{E}[f(\tilde{B}_t)1_{A\cap C}] &= \lim_{n\to\infty} \mathrm{E}[f(B_{T^{(n)}})1_{A_n \cap C}] \\
&= \lim_{n\to\infty} \sum_{k=0}^{\infty} \mathrm{E}[f(B_{k/2^n})1_{A_{n,k}\cap C}] \\
&= \lim_{n\to\infty} \sum_{k=0}^{\infty} \mathrm{E}[f(B_0)] \cdot \mathrm{P}[A_{n,k} \cap C] \\
&= \mathrm{E}[f(B_0)] \lim_{n\to\infty} \mathrm{P}[\bigcup_{k=1}^{\infty} A_{n,k} \cap C] \\
&= \mathrm{E}[f(B_0)1_{A\cap C}] \\
&= \mathrm{E}[f(B_0)] \cdot \mathrm{P}[A \cap C] \\
&= \mathrm{E}[f(B_t)] \cdot \mathrm{P}[A \cap C] \,.
\end{aligned}$$

□

Remark. If $T < \infty$ almost everywhere, no conditioning is necessary and $B_{t+T} - B_T$ is again Brownian motion.

Theorem 4.9.2 (Blumental's zero-one law). For every set $A \in \mathcal{A}_{0^+}$ we have $\mathrm{P}[A] = 0$ or $\mathrm{P}[A] = 1$.

Proof. Take the stopping time T which is identically 0. Now $\tilde{B} = B_{t+T} - B_t = B$. By Dynkin-Hunt's result, we know that $\tilde{B} = B$ is independent of $\mathcal{B}_{T^+} = \mathcal{A}_{0^+}$. Since every $C \in \mathcal{A}_{0^+}$ is $\{B_s, s > 0\}$ measurable, we know that $\mathcal{A}_{0^+}$ is independent to itself. □

Remark. This zero-one law can be used to define regular points on the boundary of a domain $D \in \mathbb{R}^d$. Given a point $y \in \delta D$. We say it is **regular**, if $P_y[T_{\delta D} > 0] = 0$ and **irregular** $P_y[T_{\delta D} > 0] = 1$. This definition turns out to be equivalent to the classical definition in potential theory: a point $y \in \delta D$ is irregular if and only if there exists a barrier function $f : N \to \mathbb{R}$ in a neighborhood N of y. A **barrier function** is defined as a negative sub-harmonic function on $\mathrm{int}(N \cap D)$ satisfying $f(x) \to 0$ for $x \to y$ within D.

4.10 Self-intersection of Brownian motion

Our aim is to prove the following theorem:

Theorem 4.10.1 (Self intersections of random walk). For $d \leq 3$, Brownian motion has infinitely many self intersections with probability 1.

Remark. Kakutani, Dvoretsky and Erdös have shown that for $d > 3$, there are no self-intersections with probability 1. It is known that for $d \leq 2$, there are infinitely many $n-$fold points and for $d \geq 3$, there are no triple points.

Proposition 4.10.2. Let K be a compact subset of $\mathbb{R}^d$ and T the hitting time of K with respect to Brownian motion starting at y. The hitting probability $h(y) = \mathrm{P}[y + B_s \in K, T \leq s < \infty]$ is a harmonic function on $\mathbb{R}^d \setminus K$.

Proof. Let T_δ be the hitting time of $S_\delta = \{|x - y| = \delta\}$. By the law of iterated logarithm, we have $T_\delta < \infty$ almost everywhere. By Dynkin-Hunt, we know that $\tilde{B}_t = B_{t+T_\delta} - B_t$ is again Brownian motion.

If δ is small enough, then $y + B_s \notin K$ for $t \leq T_\delta$. The random variable $B_{T_\delta} \in S_\delta$ has a uniform distribution on S_δ because Brownian motion is rotational symmetric. We have therefore

$$\begin{aligned} h(y) &= \mathrm{P}[y + B_s \in K, s \geq T_\delta] \\ &= \mathrm{P}[y + B_{T_\delta} + \tilde{B} \in K] \\ &= \int_{S_\delta} h(y + x)\, d\mu(x) , \end{aligned}$$

where μ is the normalized Lebesgue measure on S_δ. This equality for small enough δ is the definition of harmonicity. □

Proposition 4.10.3. Let K be a countable union of closed balls. Then $h(K, y) \to 1$ for $y \to K$.

Proof. (i) We show the claim first for one ball $K = B_r(z)$ and let $R = |z-y|$. By Brownian scaling $B_t \sim c \cdot B_{t/c^2}$. The hitting probability of K can only be a function $f(r/R)$ of r/R:

$$\begin{aligned} h(y, K) = \mathrm{P}[y + B_s \in K, T \leq s] &= \mathrm{P}[cy + B_{s/c^2} \in cK, T_K \leq s] \\ &= \mathrm{P}[cy + B_{s/c^2} \in cK, T_{cK} \leq s/c^2] \\ &= \mathrm{P}[cy + B_{\tilde{s}}, T_{cK} \leq \tilde{s}] \\ &= h(cy, cK) . \end{aligned}$$

We have to show therefore that $f(x) \to 1$ as $x \to 1$. By translation invariance, we can fix $y = y_0 = (1, 0, \dots, 0)$ and change K_α, which is a ball of radius α around $(-\alpha, 0, \dots)$. We have

$$h(y_0, K_\alpha) = f(\alpha/(1+\alpha))$$

and take therefore the limit $\alpha \to \infty$

$$\begin{aligned} \lim_{x \to 1} f(x) &= \lim_{\alpha \to \infty} h(y_0, K_\alpha) = h(y_0, \bigcup K_\alpha) \\ &= \mathrm{E}[\inf_{s \geq 0} (B_s)_1 < -1] = 1 \end{aligned}$$

because of the law of iterated logarithm.
(ii) Given $y_n \to y_0 \in K$. Then $y_0 \in K_0$ for some ball K_0.

$$\liminf_{n \to \infty} h(y_n, K) \geq \lim_{n \to \infty} h(y_n, K_0) = 1$$

by (i). □

Definition. Let μ be a probability measure on $\mathbb{R}^3$. Define the **potential theoretical energy** of μ as

$$I(\mu) = \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} |x - y|^{-1} \, d\mu(x) \, d\mu(y) \, .$$

Given a compact set $K \subset \mathbb{R}^3$, the **capacity** of K is defined as

$$(\inf_{\mu \in M(K)} I(\mu))^{-1} \, ,$$

where $M(K)$ is the set of probability measures on K. A measure on K minimizing the energy is called an **equilibrium measure**.

Remark. This definitions can be done in any dimension. In the case $d = 2$, one replaces $|x - y|^{-1}$ by $\log |x - y|^{-1}$. In the case $d \geq 3$, one takes $|x-y|^{-(d-2)}$. The capacity is for $d = 2$ defined as $\exp(-\inf_\mu I(\mu))$ and for $d \geq 3$ as $(\inf_\mu I(\mu))^{-(d-2)}$.

Definition. We say a measure μ_n on $\mathbb{R}^d$ **converges weakly** to μ, if for all continuous functions f, $\int f \, d\mu_n \to \int f \, d\mu$. The set of all probability measures on a compact subset E of $\mathbb{R}^d$ is known to be compact.

The next proposition is part of Frostman's fundamental theorem of **potential theory**. For detailed proofs, we refer to [39, 80].

Proposition 4.10.4. For every compact set $K \subset \mathbb{R}^d$, there exists an equilibrium measure μ on K and the equilibrium potential $\int |x - y|^{-(d-2)} \, d\mu(y)$ rsp. $\int \log(|x - y|^{-1}) \, d\mu(y)$ takes the value $C(K)^{-1}$ on the support K^* of μ.

Proof. (i) (Lower semicontinuity of energy) If μ_n converges to μ, then

$$\liminf_{n\to\infty} I(\mu_n) \geq I(\mu) .$$

(ii) (Existence of equilibrium measure) The existence of an equilibrium measure μ follows from the compactness of the set of probability measures on K and the lower semicontinuity of the energy since a lower semi-continuous function takes a minimum on a compact space. Take a sequence μ_n such that

$$I(\mu_n) \to \inf_{\mu\in M(K)} I(\mu) .$$

Then μ_n has an accumulation point μ and $I(\mu) \leq \inf_{\mu\in M(K)} I(\mu)$.

(iii) (Value of capacity) If the potential $\phi(x)$ belonging to μ is constant on K, then it must take the value $C(K)^{-1}$ since

$$\int \phi(x)\, d\mu(x) = I(\mu) .$$

(iv) (Constancy of capacity) Assume the potential is not constant $C(K)^{-1}$ on K^*. By constructing a new measure on K^* one shows then that one can strictly decrease the energy. This is physically evident if we think of ϕ as the potential of a charge distribution μ on the set K. □

Corollary 4.10.5. Let μ be the equilibrium distribution on K. Then

$$h(y, K) = \phi_\mu \cdot C(K)$$

and therefore $h(y, K) \geq C(K) \cdot \inf_{x\in K} |x - y|^{-1}$.

Proof. Assume first K is a countable union of balls. According to proposition (4.10.2) and proposition (4.10.3), both functions h and $\phi_\mu \cdot C(K)$ are harmonic, zero at ∞ and equal to 1 on $\delta(K)$. They must therefore be equal. For a general compact set K, let $\{y_n\}$ be a dense set in K and let $K_\epsilon = \bigcup_n B_\epsilon(y_n)$. One can pass to the limit $\epsilon \to 0$. Both $h(y, K_\epsilon) \to h(y, K)$ and $\inf_{x\in K_\epsilon} |x - y|^{-1} \to \inf_{x\in K} |x - y|^{-1}$ are clear. The statement $C(K_\epsilon) \to C(K)$ follows from the upper semicontinuity of the capacity: if G_n is a sequence of open sets with $\cap G_n = E$, then $C(G_n) \to C(E)$.
The upper semicontinuity of the capacity follows from the lower semicontinuity of the energy. □

Proposition 4.10.6. Assume, the dimension $d = 3$. For any interval $J = [a, b]$, the set

$$B_J(\omega) = \{B_t(\omega) \mid t \in [a, b]\}$$

has positive capacity for almost all ω.

Proof. We have to find a probability measure $\mu(\omega)$ on $B_I(\omega)$ such that its energy $I(\mu(\omega))$ is finite almost everywhere. Define such a measure by

$$d\mu(A) = |\frac{\{s \in [a, b] \mid B_s \in A\}}{(b-a)}| .$$

Then

$$I(\mu) = \int \int |x-y|^{-1} \, d\mu(x) d\mu(y) = \int_a^b \int_a^b (b-a)^{-1} |B_s - B_t|^{-1} \, dsdt .$$

To see the claim we have to show that this is finite almost everywhere, we integrate over Ω which is by Fubini

$$\mathrm{E}[I(\mu)] = \int_a^b \int_a^b (b-a)^{-1} \mathrm{E}[|B_s - B_t|^{-1}] \, dsdt$$

which is finite since $B_s - B_t$ has the same distribution as $\sqrt{s-t}B_1$ by Brownian scaling and since $\mathrm{E}[|B_1|^{-1}] = \int |x|^{-1} e^{-|x|^2/2} \, dx < \infty$ in dimension $d \geq 2$ and $\int_a^b \int_a^b \sqrt{s-t} \, ds \, dt < \infty$. □

Now we prove the theorem

Proof. We have only to show that in the case $d = 3$. Because Brownian motion projected to the plane is two dimensional Brownian and to the line is one dimensional Brownian motion, the result in smaller dimensions follow.

(i) $\alpha = \mathrm{P}[\bigcup_{t \in [0,1], s \geq 2} B_t = B_s] > 0$.
Proof. Let K be the set $\bigcup_{t \in [0,1]} B_t$. We know that it has positive capacity almost everywhere and that therefore $h(B_s, K) > 0$ almost everywhere. But $h(B_s, K) = \alpha$ since $B_{s+2} - B_s$ is Brownian motion independent of $B_s, 0 \leq s \leq 1$.

(ii) $\alpha_T = \mathrm{P}[\bigcup_{t \in [0,1], 2 \leq T} B_t = B_s] > 0$ for some $T > 0$. Proof. Clear since $\alpha_T \to \alpha$ for $T \to \infty$.
(iii) Proof of the claim. Define the random variables $X_n = 1_{C_n}$ with

$$C_n = \{\omega \mid B_t = B_s, \text{ for some } t \in [nT, nT+1], s \in [nT+2, (n+1)T] \} .$$

They are independent and by the strong law of large numbers $\sum_n X_n = \infty$ almost everywhere. □

Corollary 4.10.7. Any point $B_s(\omega)$ is an accumulation point of self-crossings of $\{B_t(\omega)\}_{t\geq 0}$.

Proof. Again, we have only to treat the three dimensional case. Let $T > 0$ be such that

$$\alpha_T = \mathrm{P}[\bigcup_{t\in[0,1],2\leq T} B_t = B_s] > 0$$

in the proof of the theorem. By scaling,

$$\mathrm{P}[B_t = B_s \mid t \in [0,\beta], s \in [2\beta, T\beta]\,]$$

is independent of β. We have thus self-intersections of the random walk in any interval $[0,b]$ and by translation in any interval $[a,b]$. □

4.11 Recurrence of Brownian motion

We show in this section that like its discrete brother, the random walk, Brownian motion is transient in dimensions $d \geq 3$ and recurrent in dimensions $d \leq 2$.

Lemma 4.11.1. Let T be a finite stopping time and $R_T(\omega)$ be a rotation in $\mathbb{R}^d$ which turns $B_T(\omega)$ onto the first coordinate axis

$$R_T(\omega)B_T(\omega) = (|B_T(\omega)|, 0, \ldots 0)\,.$$

Then $\tilde{B}_t = R_T(B_{t+T} - B_T)$ is again Brownian motion.

Proof. By the Dynkin-Hunt theorem, $\tilde{B}_t = B_{t+T} - B_T$ is Brownian motion and independent of $\mathcal{A}_T$. By checking the definitions of Brownian motion, it follows that if B is Brownian motion, also $R(x)B_t$ is Brownian motion, if $R(x)$ is a random rotation on $\mathbb{R}^d$ independent of B_t. Since R_T is $\mathcal{A}_T$ measurable and $\tilde{B}_t$ is independent of $\mathcal{A}_T$, the claim follows. □

Lemma 4.11.2. Let K_r be the ball of radius r centered at $0 \in \mathbb{R}^d$ with $d \geq 3$. We have for $y \notin K_r$

$$h(y, K_r) = (r/|y|)^{d-2}\,.$$

Proof. Both $h(y, K_r)$ and $(r/|y|)^{d-2}$ are harmonic functions which are 1 at δK_r and zero at infinity. They are the same. □

Theorem 4.11.3 (Escape of Brownian motion in three dimensions). For $d \geq 3$, we have $\lim_{t\to\infty} |B_t| = \infty$ almost surely.

Proof. Define a sequence of stopping times T_n by

$$T_n = \inf\{s > 0 \mid |B_s| = 2^n\} ,$$

which is finite almost everywhere because of the law of iterated logarithm. We know from the lemma (4.11.1) that

$$\tilde{B}_t = R_{T_n}(B_{t+T_n} - B_{T_n})$$

is a copy of Brownian motion. Clearly also $|B_{T_n}| = 2^n$.
We have $B_s \in K_r(0) = \{|x| < r\}$ for some $s > T_n$ if and only if $\tilde{B}_t \in (2^n, 0 \dots, 0) + K_r(0)$ for some $t > 0$.
Therefore using the previous lemma

$$\mathrm{P}[B_s \in K_r(0); s > T_n] = \mathrm{P}[\tilde{B}_t \in (2^n, 0 \dots, 0) + K_r(0); t > 0] = (\frac{r}{2^n})^{d-2}$$

which implies in the case $r2^{-n} < 1$ by the Borel-Cantelli lemma that for almost all ω, $B_s(\omega) \geq r$ for $s > T_n$. Since T_n is finite almost everywhere, we get $\liminf_s |B_s| \geq r$. Since r is arbitrary, the claim follows. □

Brownian motion is recurrent in dimensions $d \leq 2$. In the case $d = 1$, this follows readily from the law of iterated logarithm. First a lemma

---

Lemma 4.11.4. In dimensions $d = 2$, almost every path of Brownian motion hits a ball K_r if $r > 0$: one has $h(y, K) = 1$.

Proof. We know that $h(y) = h(y, K)$ is harmonic and equal to 1 on δK. It is also rotational invariant and therefore $h(y) = a + b \log |y|$. Since $h \in [0, 1]$ we have $h(y) = a$ and so $a = 1$. □

Theorem 4.11.5 (Recurrence of Brownian motion in 1 or 2 dimensions). Let $d \leq 2$ and S be an open nonempty set in $\mathbb{R}^d$. Then the Lebesgue measure of $\{t \mid B_t \in S\}$ is infinite.

Proof. It suffices to take $S = K_r(x_0)$, a ball of radius r around x_0. Since by the previous lemma, Brownian motion hits every ball almost surely, we can assume that $x_0 = 0$ and by scaling that $r = 1$.
Define inductively a sequence of hitting or leaving times T_n, S_n of the annulus $\{1/2 < |x| < 2\}$, where $T_1 = \inf\{t \mid |B_t| = 2\}$ and

$$\begin{aligned} S_n &= \inf\{t > T_n \mid |B_t| = 1/2\} \\ T_n &= \inf\{t > S_{n-1} \mid |B_t| = 2\} . \end{aligned}$$

These are finite stopping times. The Dynkin-Hunt theorem shows that $S_n - T_n$ and $T_n - S_{n-1}$ are two mutually independent families of IID random variables. The Lebesgue measures $Y_n = |I_n|$ of the time intervals

$$I_n = \{t \mid |B_t| \leq 1, \; T_n \leq t \leq T_{n+1} \} ,$$

are independent random variables. Therefore, also $X_n = \min(1, Y_n)$ are independent bounded IID random variables. By the law of large numbers, $\sum_n X_n = \infty$ which implies $\sum_n Y_n = \infty$ and the claim follows from

$$|\{t \in [0, \infty) \mid |B_t| \leq 1 \}| \geq \sum_n T_n .$$

□

Remark. Brownian motion in $\mathbb{R}^d$ can be defined as a diffusion on $\mathbb{R}^d$ with generator $\Delta/2$, where Δ is the Laplacian on $\mathbb{R}^d$. A generalization of Brownian motion to manifolds can be done using the diffusion processes with respect to the Laplace-Beltrami operator. Like this, one can define Brownian motion on the torus or on the sphere for example. See [57].

4.12 Feynman-Kac formula

In quantum mechanics, the Schrödinger equation $i\hbar\dot{u} = Hu$ defines the evolution of the wave function $u(t) = e^{-itH/\hbar}u(0)$ in a Hilbert space $\mathcal{H}$. The operator H is the **Hamiltonian** of the system. We assume, it is a **Schrödinger operator** $H = H_0 + V$, where $H_0 = -\Delta/2$ is the Hamiltonian of a free particle and $V : \mathbb{R}^d \to \mathbb{R}$ is the potential. The free operator H_0 already is not defined on the whole Hilbert space $\mathcal{H} = L^2(\mathbb{R}^d)$ and one restricts H to a vector space $D(H)$ called **domain** containing the in $\mathcal{H}$ dense set $C_0^\infty(\mathbb{R}^d)$ of all smooth functions which are zero at infinity. Define

$$D(A^*) = \{u \in \mathcal{H} \mid v \mapsto (Av, u) \text{ is a bounded linear functional on } D(A)\}.$$

If $u \in D(A^*)$, then there exists a unique function $w = A^*u \in \mathcal{H}$ such that $(Av, u) = (v, w)$ for all $u \in D(A)$. This defines the **adjoint** A^* of A with domain $D(A^*)$.

Definition. A linear operator $A : D(A) \subset \mathcal{H} \to \mathcal{H}$ is called **symmetric** if $(Au, v) = (u, Av)$ for all $u, v \in D(A)$ and **self-adjoint**, if it is symmetric and $D(A) = D(A^*)$.

Definition. A sequence of bounded linear operators A_n converges **strongly** to A, if $A_n u \to Au$ for all $u \in \mathcal{H}$. One writes $A = s - \lim_{n\to\infty} A_n$.

Define $e^A = 1 + A + A^2/2! + A^3/3! + \cdots$. We will use the fact that a self-adjoint operator defines a one parameter family of unitary operators $t \mapsto e^{itA}$ which is strongly continuous. Moreover, e^{itA} leaves the domain $D(A)$ of A invariant. For more details, see [81, 7].

Theorem 4.12.1 (Trotter product formula). Given self-adjoint operators A, B defined on $D(A), D(B) \subset \mathcal{H}$. Assume $A + B$ is self-adjoint on $D = D(A) \cap D(B)$, then

$$e^{it(A+B)} = s - \lim_{n\to\infty} (e^{itA/n} e^{itB/n})^n .$$

If A, B are bounded from below, then

$$e^{-t(A+B)} = s - \lim_{n\to\infty} (e^{-tA/n} e^{-tB/n})^n .$$

Proof. Define

$$S_t = e^{it(A+B)}, V_t = e^{itA}, W_t = e^{itB}, U_t = V_t W_t$$

and $v_t = S_t v$ for $v \in D$. Because $A+B$ is self-adjoint on D, one has $v_t \in D$. Use a telescopic sum to estimate

$$\begin{aligned} ||(S_t - U_{t/n}^n)v|| &= ||\sum_{j=0}^{n-1} U_{t/n}^j (S_{t/n} - U_{t/n}) S_{t/n}^{n-j-1} v|| \\ &\leq n \sup_{0\leq s\leq t} ||(S_{t/n} - U_{t/n}) v_s|| . \end{aligned}$$

We have to show that this goes to zero for $n \to \infty$. Given $u \in D = D(A) \cap D(B)$,

$$\lim_{s\to 0} \frac{S_s - 1}{s} u = i(A + B)u = \lim_{s\to 0} \frac{U_s - 1}{s} u$$

so that for each $u \in D$

$$\lim_{n\to\infty} n \cdot ||(S_{t/n} - U_{t/n})u|| = 0 . \tag{4.3}$$

The linear space D with norm $|||u||| = ||(A + B)u|| + ||u||$ is a Banach space since $A + B$ is self-adjoint on D and therefore closed. We have a bounded family $\{n(S_{t/n} - U_{t/n})\}_{n\in\mathbb{N}}$ of bounded operators from D to $\mathcal{H}$. The principle of uniform boundedness states that

$$||n(S_{t/n} - U_{t/n})u|| \leq C \cdot |||u||| .$$

An $\epsilon/3$ argument shows that the limit (4.3) exists uniformly on compact subsets of D and especially on $\{v_s\}_{s\in[0,t]} \subset D$ and so $n \sup_{0\leq s\leq t} ||(S_{t/n} - U_{t/n})v_s|| = 0$. The second statement is proved in exactly the same way. □

Remark. Trotter's product formula generalizes the Lie product formula

$$\lim_{n\to\infty} (\exp(\frac{A}{n})\exp(\frac{B}{n}))^n = \exp(A+B)$$

for finite dimensional matrices A, B, which is a special case.

Corollary 4.12.2. (Feynman 1948) Assume $H = H_0 + V$ is self-adjoint on $D(H)$. Then

$$e^{-itH}u(x_0) = \lim_{n\to\infty} (\frac{2\pi it}{n})^{-d/2} \int_{(\mathbb{R}^d)^n} e^{iS_n(x_0,x_1,x_2,\dots,x_n,t)} u(x_n) \, dx_1 \dots dx_n$$

where

$$S_n(x_0, x_1, \dots, x_n, t) = \frac{t}{n}\sum_{i=1}^{n} \frac{1}{2}(\frac{|x_i - x_{i-1}|}{t/n})^2 - V(x_i) \; .$$

Proof. (Nelson) From $\dot{u} = -iH_0 u$, we get by Fourier transform $\dot{\hat{u}} = i\frac{|k|^2}{2}\hat{u}$ which gives $\hat{u}_t(k) = \exp(i\frac{|k|^2}{2}t)\hat{u}_0(k)$ and by inverse Fourier transform

$$e^{-itH_0}u(x) = u_t(x) = (2\pi it)^{-d/2}\int_{\mathbb{R}^d} e^{i\frac{|x-y|^2}{2t}} u(y) \, dy \; .$$

The Trotter product formula

$$e^{-it(H_0+V)} = s - \lim_{n\to\infty} (e^{itH_0/n}e^{itV/n})^n \quad .$$

gives now the claim. □

Remark. We did not specify the set of potentials, for which $H_0 + V$ can be made self-adjoint. For example, $V \in C_0^\infty(\mathbb{R}^\nu)$ is enough or $V \in L^2(\mathbb{R}^3) \cap L^\infty(\mathbb{R}^3)$ in three dimensions.

We have seen in the above proof that e^{-itH_0} has the integral kernel $\tilde{P}_t(x,y) = (2\pi it)^{-d/2}e^{i\frac{|x-y|^2}{2t}}$. The same Fourier calculation shows that e^{-tH_0} has the integral kernel

$$P_t(x,y) = (2\pi t)^{-d/2}e^{-\frac{|x-y|^2}{2t}} \; ,$$

where g_t is the density of a Gaussian random variable with variance t. Note that even if $u \in L^2(\mathbb{R}^d)$ is only defined almost everywhere, the function $u_t(x) = e^{-tH_0}u(x) = \int P_t(x-y)u(y)dy$ is continuous and defined everywhere.

Lemma 4.12.3. Given $f_1, \dots, f_n \in L^\infty(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$ and $0 < s_1 < \cdots < s_n$. Then

$$(e^{-t_1 H_0} f_1 \cdots e^{-t_n H_0} f_n)(0) = \int f_1(B_{s_1}) \cdots f_n(B_{s_n}) \, dB,$$

where $t_1 = s_1, t_i = s_i - s_{i-1}, i \geq 2$ and the f_i on the left hand side are understood as multiplication operators on $L^2(\mathbb{R}^d)$.

Proof. Since B_{s_1}, $B_{s_2} - B_{s_1}, \dots B_{s_n} - B_{s_{n-1}}$ are mutually independent Gaussian random variables of variance $t_1, t_2, \dots, t_n$, their joint distribution is

$$P_{t_1}(0, y_1) P_{t_2}(0, y_2) \dots P_{t_n}(0, y_n) \, dy$$

which is after a change of variables $y_1 = x_1, y_i = x_i - x_{i-1}$

$$P_{t_1}(0, x_1) P_{t_2}(x_1, x_2) \dots P_{t_n}(x_{n-1}, x_n) \, dx \, .$$

Therefore,

$$\begin{aligned}
& \int f_1(B_{s_1}) \cdots f_n(B_{s_n}) \, dB \\
& \int_{(\mathbb{R}^d)^n} P_{t_1}(0, y_1) P_{t_2}(0, y_2) \dots P_{t_n}(0, y_n) f_1(y_1) \dots f_n(y_n) \, dy \\
= & \int_{(\mathbb{R}^d)^n} P_{t_1}(0, x_1) P_{t_2}(x_1, x_2) \dots P_{t_n}(x_{n-1}, x_n) f_1(x_1) \dots f_n(x_n) \, dx \\
= & \ (e^{-t_1 H_0} f_1 \cdots e^{-t_n H_0} f_n)(0) \, .
\end{aligned}$$

□

Denote by dB the Wiener measure on $C([0, \infty), \mathbb{R}^d)$ and with dx the Lebesgue measure on $\mathbb{R}^d$. We define also an **extended Wiener measure** $dW = dx \times dB$ on $C([0, \infty), \mathbb{R}^d)$ on all paths $s \mapsto W_s = x + B_s$ starting at $x \in \mathbb{R}^d$.

Corollary 4.12.4. Given $f_0, f_1, \dots, f_n \in L^\infty(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$ and $0 < s_1 < \cdots < s_n$. Then

$$\int f_0(W_{s_0}) \cdots f_n(W_{s_n}) \, dW = (\overline{f_0}, e^{-t_1 H_0} f_1 \cdots e^{-t_n H_0} f_n) \, .$$

Proof. (i) Case $s_0 = 0$. From the above lemma, we have after the dB integration that

$$\begin{aligned}\int f_0(W_{s_0}) \cdots f_n(W_{s_n})\, dW &= \int_{\mathbb{R}^d} f_0(x) e^{-t_1 H_0} f_1(x) \cdots e^{-t_n H_0} f_n(x)\, dx \\ &= (\overline{f}_0, e^{-t_1 H_0} f_1 \cdots e^{-t_n H_0} f_n) .\end{aligned}$$

(ii) In the case $s_0 > 0$ we have from (i) and the dominated convergence theorem

$$\begin{aligned}\int & f_0(W_{s_0}) \cdots f_n(W_{s_n})\, dW \\ = & \lim_{R\to\infty} \int_{\mathbb{R}^d} 1_{\{|x|<R\}}(W_0) \\ & f_0(W_{s_0}) \cdots f_n(W_{s_n})\, dW \\ = & \lim_{R\to\infty} (\overline{f}_0 e^{-s_0 H_0} 1_{\{|x|<R\}}, e^{-t_1 H_0} f_1 \cdots e^{-t_n H_0} f_n(x)) \\ = & (\overline{f}_0, e^{-t_1 H_0} f_1 \cdots e^{-t_n H_0} f_n) .\end{aligned}$$

□

We prove now the Feynman-Kac formula for Schrödinger operators of the form $H = H_0 + V$ with $V \in C_0^\infty(\mathbb{R}^d)$. Because V is continuous, the integral $\int_0^t V(W_s(\omega))\, ds$ can be taken for each ω as a limit of Riemann sums and $\int_0^t V(W_s)\, ds$ certainly is a random variable.

Theorem 4.12.5 (Feynman-Kac formula). Given $H = H_0 + V$ with $V \in C_0^\infty(\mathbb{R}^d)$, then

$$(f, e^{-tH} g) = \int \overline{f}(W_0) g(W_t) e^{-\int_0^t V(W_s))\, ds}\, dW .$$

Proof. (Nelson) By the Trotter product formula

$$(f, e^{-tH} g) = \lim_{n\to\infty} (f, (e^{-tH_0/n} e^{-tV/n})^n g)$$

so that by corollary (4.12.4)

$$(f, e^{-tH} g) = \lim_{n\to\infty} \int \overline{f}(W_0) g(W_t) \exp(-\frac{t}{n} \sum_{j=0}^{n-1} V(W_{tj/n}))\, dW \tag{4.4}$$

and since $s \mapsto W_s$ is continuous, we have almost everywhere

$$\frac{t}{n} \sum_{j=0}^{n-1} V(W_{tj/n}) \to \int_0^t V(W_s)\, ds .$$

The integrand on the right hand side of (4.4) is dominated by

$$|f(W_0)| \cdot |g(W_t)| \cdot e^{t||V||_\infty}$$

which is in $L^1(dW)$ because again by corollary (4.12.4),

$$\int |f(W_0)| \cdot |g(W_t)| \, dW = (|f|, e^{-tH_0}|g|) < \infty .$$

The dominated convergence theorem leads us now to the claim. □

Remark. The formula can be extended to larger classes of potentials like potentials V which are locally in L^1. The selfadjointness, which needed in Trotter's product formula, is assured if $V \in L^2 \cap L^p$ with $p > d/2$. Also Trotter's product formula allows further generalizations [93, 31].

Why is the Feynman-Kac formula useful?

- One can use Brownian motion to study Schrödinger semigroups. It allows for example to give an easy proof of the ArcSin-law for Brownian motion.
- One can treat operators with magnetic fields in a unified way.
- Functional integration is a way of quantization which generalizes to more situations.
- It is useful to study ground states and ground state energies under perturbations.
- One can study the classical limit $\hbar \to 0$.

4.13 The quantum mechanical oscillator

The one-dimensional Schrödinger operator

$$H = H_0 + U = -\frac{1}{2}\frac{d^2}{dx^2} + \frac{1}{2}x^2 - \frac{1}{2}$$

is the Hamiltonian of the **quantum mechanical oscillator**. It is a quantum mechanical system which can be solved explicitly like its classical analog, which has the Hamiltonian $H(x,p) = \frac{1}{2}p^2 + \frac{1}{2}x^2 - \frac{1}{2}$.

One can write

$$H = AA^* - 1 = A^*A ,$$

with

$$A^* = \frac{1}{\sqrt{2}}(x - \frac{d}{dx}), \; A = \frac{1}{\sqrt{2}}(x + \frac{d}{dx}) .$$

The first order operator A^* is also called **particle creation operator** and A, the **particle annihilation operator**. The space C_0^∞ of smooth functions of compact support is dense in $L^2(\mathbb{R})$. Because for all $u, v \in C_0^\infty(\mathbb{R})$

$$(Au, v) = (u, A^*v)$$

the two operators are adjoint to each other. The vector

$$\Omega_0 = \frac{1}{\pi^{1/4}} e^{-x^2/2}$$

is a unit vector because Ω_0^2 is the density of a $N(0, 1/\sqrt{2})$ distributed random variable. Because $A\Omega_0 = 0$, it is an eigenvector of $H = A^*A$ with eigenvalue $1/2$. It is called the **ground state** or **vacuum state** describing the system with no particle. Define inductively the n-particle states

$$\Omega_n = \frac{1}{\sqrt{n}} A^* \Omega_{n-1}$$

by creating an additional particle from the $(n-1)$-particle state Ω_{n-1} .

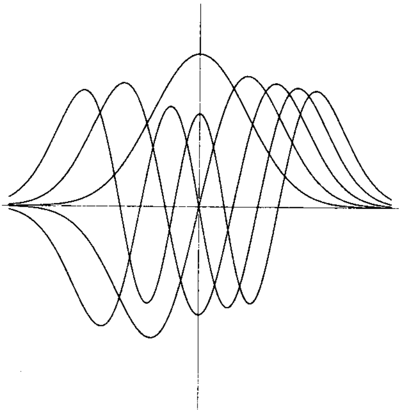


Figure. *The first Hermit functions Ω_n. They are unit vectors in $L^2(\mathbb{R})$ defined by*

$$\Omega_n(x) = \frac{H_n(x)\omega_0(x)}{\sqrt{2^n n!}} ,$$

where $H_n(x)$ are **Hermite polynomials**, *$H_0(x) = 1, H_1(x) = 2x, H_2(x) = 4x^2 - 2, H_3(x) = 8x^3 - 12x, \dots$.*

Theorem 4.13.1 (Quantum mechanical oscillator). The following properties hold:
a) The functions are orthonormal $(\Omega_n, \Omega_m) = \delta_{n,m}$.
b) $A\Omega_n = \sqrt{n}\Omega_{n-1}, A^*\Omega_n = \sqrt{n+1}\Omega_{n+1}$.
c) $(n - \frac{1}{2})$ are the eigenvalues of H

$$H = (A^*A - \frac{1}{2})\Omega_n = (n - \frac{1}{2})\Omega_n$$

.
d) The functions Ω_n form a basis in $L^2(\mathbb{R})$.

Proof. Denote by $[A, B] = AB - BA$ the commutator of two operators A and B. We check first by induction the formula

$$[A, (A^*)^n] = n \cdot (A^*)^{n-1} .$$

For $n = 1$, this means $[A, A^*] = 1$. The induction step is

$$\begin{aligned}[A,(A^*)^n] &= [A,(A^*)^{n-1}]A^* + (A^*)^{n-1}[A,A^*] \\ &= (n-1)(A^*)^{n-1} + (A^*)^{n-1} = n(A^*)^{n-1} .\end{aligned}$$

a) Also

$$((A^*)^n\Omega_0, (A^*)^m\Omega_0) = n!\delta_{mn} \cdot$$

can be proven by induction. For $n = 0$ it follows from the fact that Ω_0 is normalized. The induction step uses $[A,(A^*)^n] = n \cdot (A^*)^{n-1}$ and $A\Omega_0 = 0$:

$$\begin{aligned}((A^*)^n\Omega_0, (A^*)^m\Omega_0) &= (A(A^*)^n\Omega_0(A^*)^{m-1}\Omega_0) \\ &= ([A,(A^*)^n]\Omega_0(A^*)^{m-1}\Omega_0) \\ &= n((A^*)^{n-1}\Omega_0, (A^*)^{m-1}\Omega_0) .\end{aligned}$$

If $n < m$, then we get from this 0 after n steps, while in the case $n = m$, we obtain $((A^*)^n\Omega_0, (A^*)^n\Omega_0) = n \cdot ((A^*)^{n-1}\Omega_0, (A^*)^{n-1}\Omega_0)$, which is by induction $n(n-1)!\delta_{n-1,n-1} = n!$.

b) $A^*\Omega_n = \sqrt{n+1} \cdot \Omega_{n+1}$ is the definition of Ω_n.

$$A\Omega_n = \frac{1}{\sqrt{n!}}A(A^*)^n\Omega_0 = \frac{1}{\sqrt{n!}}n\Omega_0 = \sqrt{n}\Omega_{n-1} .$$

c) This follows from *b*) and the definition $\Omega_n = \frac{1}{\sqrt{n}}A^*\Omega_{n-1}$.

d) Part a) shows that $\{\Omega_n\}_{n=0}^\infty$ it is an orthonormal set in $L^2(\mathbb{R})$. In order to show that they span $L^2(\mathbb{R})$, we have to verify that they span the dense set

$$\mathcal{S} = \{f \in C_0^\infty(\mathbb{R}) \mid x^m f^{(n)}(x) \to 0, |x| \to \infty, \forall m, n \in \mathbb{N}\}$$

called the **Schwarz space**. The reason is that by the Hahn-Banach theorem, a function f must be zero in $L^2(\mathbb{R})$ if it is orthogonal to a dense set. So, lets assume $(f, \Omega_n) = 0$ for all n. Because $A^* + A = \sqrt{2}x$

$$0 = \sqrt{n!2^n}\,(f,\Omega_n) = (f,(A^*)^n\Omega_0) = (f,(A^*+A)^n\Omega_0) = 2^{n/2}\,(f,x^n\Omega_0)$$

we have

$$\begin{aligned}(f\Omega_0)\hat{}(k) &= \int_{-\infty}^{\infty} f(x)\Omega_0(x)e^{ikx}\,dx \\ &= (f,\Omega_0 e^{ikx}) = (f, \sum_{n\geq 0}\frac{(ikx)^n}{n!}\Omega_0) \\ &= \sum_{n\geq 0}\frac{(ik)^n}{n!}(f,x^n\Omega_0) = 0 .\end{aligned}$$

and so $f\Omega_0 = 0$. Since $\Omega_0(x)$ is positive for all x, we must have $f = 0$. This finishes the proof that we have a complete basis. □

Remark. This gives a complete solution to the quantum mechanical harmonic oscillator. With the eigenvalues $\{\lambda_n = n - 1/2\}_{n=0}^{\infty}$ and the complete set of eigenvectors Ω_n one can solve the **Schrödinger equation**

$$i\hbar \frac{d}{dt} u = Hu$$

by writing the function $u(x) = \sum_{n=0}^{\infty} u_n \Omega_n(x)$ as a sum of eigenfuctions, where $u_n = (u, \Omega_n)$. The solution of the Schrödinger equation is

$$u(t,x) = \sum_{n=0}^{\infty} u_n e^{i\hbar(n-1/2)t} \Omega_n(x) .$$

Remark. The formalism of particle creation and annihilation operators can be extended to some potentials of the form $U(x) = q^2(x) - q'(x)$ the operator $H = -D^2/2 + U/2$ can then be written as $H = A^*A$, where

$$A^* = \frac{1}{\sqrt{2}}(q(x) - \frac{d}{dx}),\ A = \frac{1}{\sqrt{2}}(q(x) + \frac{d}{dx}) .$$

The oscillator is the special case $q(x) = x$. See [12]. The **Bäcklund transformation** $H = A^*A \mapsto \tilde{H} = AA^*$ is in the case of the harmonic oscillator the map $H \mapsto H + 1$ has the effect that it replaces U with $\tilde{U} = U - \partial_x^2 \log \Omega_0$, where Ω_0 is the lowest eigenvalue. The new operator $\tilde{H}$ has the same spectrum as H except that the lowest eigenvalue is removed. This procedure can be reversed and to create "soliton potentials" out of the vacuum. It is also natural to use the language of **super-symmetry** as introduced by Witten: take two copies $\mathcal{H}_f \oplus \mathcal{H}_b$ of the Hilbert space where "f" stands for Fermion and "b" for Boson. With

$$Q = \begin{bmatrix} 0 & A^* \\ A & 0 \end{bmatrix}, P = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix},$$

one can write $H \oplus \tilde{H} = Q^2$, $P^2 = 1$, $QP + PQ = 0$ and one says (H, P, Q) has super-symmetry. The operator Q is also called a **Dirac operator**. A super-symmetric system has the property that nonzero eigenvalues have the same number of bosonic and fermionic eigenstates. This implies that $\tilde{H}$ has the same spectrum as H except that lowest eigenvalue can disappear.

Remark. In **quantum field theory**, there exists a process called **canonical quantization**, where a quantum mechanical system is extended to a quantum field. Particle annihilation and creation operators play an important role.

4.14 Feynman-Kac for the oscillator

We want to treat perturbations $L = L_0 + V$ of the harmonic oscillator L_0 with an similar Feynman-Kac formula. The calculation of the integral kernel $p_t(x,y)$ of e^{-tL_0} satisfying

$$(e^{-tL_0}f)(x) = \int_{\mathbb{R}} p_t(x,y)f(y)\,dy$$

is slightly more involved than in the case of the free Laplacian. Let Ω_0 be the ground state of L_0 as in the last section.

Lemma 4.14.1. Given $f_0, f_1, \dots, f_n \in L^\infty(\mathbb{R})$ and $-\infty < s_0 < s_1 < \cdots < s_n < \infty$. Then

$$(\Omega_0, f_0 e^{-t_1 L_0} f_1 \cdots e^{-t_n L_0} f_n \Omega_0) = \int f_0(Q_{s_0}) \cdots f_n(Q_{s_n})\,dQ\ ,$$

where $t_0 = s_0, t_i = s_i - s_{i-1}, i \geq 1$.

Proof. The Trotter product formula for $L_0 = H_0 + U$ gives

$$\begin{aligned} & (\Omega_0, f_0 e^{-t_1 L_0} f_1 \cdots e^{-t_n L_0} f_n \Omega_0) \\ = & \lim_{m=(m_1,\dots,m_n), m_i \to \infty} (\Omega_0, f_0 (e^{-t_1 H_0/m_1} e^{-t_1 U/m_1})^{m_1} f_1 \cdots e^{-t_n H_0} f_n \Omega_0) \\ = & \int f_0(x_0) \cdots f_n(x_n)\,dG_m(x,y) \end{aligned}$$

and G_m is a measure. Since e^{-tH_0} has a Gaussian kernel and e^{-tU} is a multiple of a Gaussian density and integrals are Gaussian, the measure dG_m is Gaussian converging to a Gaussian measure dG. Since $L_0(x\Omega_0) = x\Omega_0$ and $(x\Omega_0, x\Omega_0) = 1/2$ we have

$$\int x_i x_j\,dG = (x\Omega_0, e^{-(s_j - s_i)} L_0 x \Omega_0) = \frac{1}{2} e^{-(s_j - s_i)}$$

which shows that dG is the joint probability distribution of $Q_{s_0}, \dots Q_{s_n}$. The claim follows. □

Theorem 4.14.2 (Mehler formula). The kernel $p_t(x,y)$ of L_0 is given by the **Mehler formula**

$$p_t(x,y) = \frac{1}{\sqrt{\pi\sigma^2}} \exp\left(-\frac{(x^2+y^2)(1+e^{-2t}) - 4xye^{-t}}{2\sigma^2}\right) .$$

with $\sigma^2 = (1 - e^{-2t})$.

Proof. We have

$$(f, e^{-tL_0}g) = \int f(y)\Omega_0^{-1}(y)g(x)\Omega_0^{-1}(x)\, dG(x,y) = \int f(y)p_t(x,y)\, dy$$

with the Gaussian measure dG having covariance

$$A = \frac{1}{2}\begin{bmatrix} 1 & e^{-t} \\ e^{-t} & 1 \end{bmatrix} .$$

We get Mehler's formula by inverting this matrix and using that the density is

$$(2\pi)\det(A)^{-1/2}e^{-((x,y),A(x,y))} .$$

□

Definition. Let dQ be **the Wiener measure** on $C(\mathbb{R})$ belonging to the oscillator process Q_t.

Theorem 4.14.3 (Feynman-Kac for oscillator process). Given $L = L_0 + V$ with $V \in C_0^\infty(\mathbb{R})$, then

$$(f\Omega_0, e^{-iL}g\Omega_0) = \int \overline{f}(Q_0)g(Q_t)e^{-\int_0^t V(Q_s))\, ds}\, dQ$$

for all $f, g \in L^2(\mathbb{R}, \Omega_0^2 dx)$.

Proof. By the Trotter product formula

$$(f\Omega_0, e^{-iL}g\Omega_0) = \lim_{n\to\infty}(f\Omega_0, (e^{-tL_0/n}e^{-tV/n})^n g\Omega_0)$$

so that

$$(f\Omega_0, e^{-iL}g\Omega_0) = \lim_{n\to\infty}\int \overline{f}(Q_0)g(Q_t)\exp(-\frac{t}{n}\sum_{j=0}^{n-1}V(Q_{tj/n}))\, dQ . \quad (4.5)$$

and since Q is continuous, we have almost everywhere

$$\frac{t}{n}\sum_{j=0}^{n-1}V(Q_{tj/n}) \to \int_0^t V(Q_s)\, ds .$$

The integrand on the right hand side of (4.5) is dominated by

$$|f(Q_0)||g(Q_t)|e^{t||V||_\infty}$$

which is in $L^1(dQ)$ since

$$\int |f(Q_0)||g(Q_t)|\, dQ = (\Omega_0|f|, e^{-tL_0}\Omega_0|g|) < \infty .$$

The dominated convergence theorem gives the claim. □

4.15 Neighborhood of Brownian motion

The Feynman-Kac formula can be used to understand the Dirichlet Laplacian of a domain $D \subset \mathbb{R}^d$. For more details, see [93].

Example. Let D be an open set in $\mathbb{R}^d$ such that the Lebesgue measure $|D|$ is finite and the Lebesgue measure of the boundary $|\delta D|$ is zero. Denote by H_D the Dirichlet Laplacian $-\Delta/2$. Denote by $k_D(E)$ the number of eigenvalues of H_D below E. This function is also called the integrated density of states. Denote with K_d the unit ball in $\mathbb{R}^d$ and with $|K_d| = \text{Vol}(K_d) = \pi^{d/2}\Gamma(\frac{d}{2}+1)^{-1}$ its volume. **Weyl's formula** describes the asymptotic behavior of $k_D(E)$ for large E:

$$\lim_{E\to\infty} \frac{k_D(E)}{E^{d/2}} = \frac{|K_d| \cdot |D|}{2^{d/2}\pi^d} .$$

It shows that one can read off the volume of D from the spectrum of the Laplacian.

Example. Put n ice balls $K_{j,n}, 1 \leq j \leq n$ of radius r_n into a glass of water so that $n \cdot r_n = \alpha$. In order to know, how good this ice cools the water it is good to know the lowest eigenvalue E_1 of the Dirichlet Laplacian H_D since the motion of the temperature distribution u by the heat equation $\dot{u} = H_D u$ is dominated by $e^{-t\text{E}_1}$. This motivates to compute the lowest eigenvalue of the domain $D \setminus \bigcup_{j=1}^n K_{j,n}$. This can be done exactly in the limit $n \to \infty$ and when ice $K_{j,n}$ is randomly distributed in the glass. Mathematically, this is described as follows:
Let D be an open bounded domain in $\mathbb{R}^d$. Given a sequence $x = (x_1, x_2, \dots)$ which is an element in $D^{\mathbb{N}}$ and a sequence of radii $r_1, r_2, \dots$, define

$$D_n = D \setminus \bigcup_{i=1}^n \{|x - x_i| \leq r_n\} .$$

This is the domain D with n points balls $K_{j,n}$ with center $x_1, \dots x_n$ and radius r_n removed. Let $H(x,n)$ be the Dirichlet Laplacian on D_n and $\text{E}_k(x,n)$ the k-th eigenvalue of $H(x,n)$ which are random variable $\text{E}_k(n)$ in x, if $D^{\mathbb{N}}$ is equipped with the product Lebesgue measure. One can show that in the case $nr_n \to \alpha$

$$\text{E}_k(n) \to \text{E}_k(0) + 2\pi\alpha|D|^{-1}$$

in probability. Random impurities produce a constant shift in the spectrum. For the physical system with the crushed ice, where the crushing makes $nr_n \to \infty$, there is much better cooling as one might expect.

Definition. Let $\mathcal{W}_\delta(t)$ be the set

$$\{x \in \mathbb{R}^d \mid |x - B_t(\omega)| \leq \delta, \text{ for some } s \in [0,t]\} .$$

It is of course dependent on ω and just a δ-neighborhood of the Brownian path $B_{[0,t]}(\omega)$. This set is called **Wiener sausage** and one is interested in the expected volume $|\mathcal{W}_\delta(t)|$ of this set as $\delta \to 0$. We will look at this problem a bit more closely in the rest of this section.

Figure. *A sample of Wiener sausage in the plane $d = 2$. A finite path of Brownian motion with its neighborhood $\mathcal{W}_\delta$.*

Lets first prove a lemma, which relates the Dirichlet Laplacian $H_D = -\Delta/2$ on D with Brownian motion.

Lemma 4.15.1. Let D be a bounded domain in $\mathbb{R}^d$ containing 0 and $p_D(x, y, t)$, the integral kernel of e^{-tH}, where H is the Dirichlet Laplacian on D. Then

$$\mathrm{E}[B_s \in D; 0 \leq s \leq t] = 1 - \int p_D(0, x, t)\, dx\ .$$

Proof. (i) It is known that the Dirichlet Laplacian can be approximated in the strong resolvent sense by operators $H_0 + \lambda V$, where $V = 1_{D^c}$ is the characteristic function of the exterior D^c of D. This means that

$$(H_0 + \lambda \cdot V)^{-1} u \to (H_D - z)^{-1} u, \lambda \to \infty$$

for z outside $[0, \infty)$ and all $u \in C_c^\infty(\mathbb{R}^d)$.

(ii) Since Brownian paths are continuous, we have $\int_0^t V(B_s)\, ds > 0$ if and only if $B_s \in C^c$ for some $s \in [0, t]$. We get therefore

$$e^{-\lambda \int_0^t V(B_s)\, ds} \to 1_{\{B_s \in D^c\}}$$

point wise almost everywhere.

Let u_n be a sequence in C_c^∞ converging point wise to 1. We get with the dominated convergence theorem, using (*i*) and (*ii*) and Feynman-Kac

$$\begin{aligned}
\mathrm{E}[B_s \in D^c; 0 \leq s \leq t] &= \lim_{n\to\infty} \mathrm{E}[u_n(B_s) \in D^c; 0 \leq s \leq t] \\
&= \lim_{n\to\infty} \lim_{\lambda\to\infty} \mathrm{E}[e^{-\lambda \int_0^t V(B_s)\, ds} u_n(B_t)] \\
&= \lim_{n\to\infty} \lim_{\lambda\to\infty} e^{-t(H_0+\lambda\cdot V)} u_n(0) \\
&= \lim_{n\to\infty} e^{-tH_D} u_n(0) \\
&= \lim_{n\to\infty} \int p_D(0,x,t) u_n(0)\, dx = \int p_D(0,x,t)\, dx \; .
\end{aligned}$$

□

Theorem 4.15.2 (Spitzer). In three dimensions $d = 3$,

$$\mathrm{E}[|\mathcal{W}_\delta(t)|] = 2\pi\delta t + 4\delta^2\sqrt{2\pi t} + \frac{4\pi}{3}\delta^3 \; .$$

Proof. Using Brownian scaling,

$$\begin{aligned}
\mathrm{E}[|\mathcal{W}_{\lambda\delta}(\lambda^2 t)|] &= \mathrm{E}[|\{|x - B_s| \leq \lambda\delta, 0 \leq s \leq \lambda^2 t\}|] \\
&= \mathrm{E}[|\{|\frac{x}{\lambda} - \frac{B_{\tilde{s}\lambda^2}}{\lambda}| \leq \delta, 0 \leq \tilde{s} = s/\lambda^2 \leq t\}|] \\
&= \mathrm{E}[|\{|\frac{x}{\lambda} - B_{\tilde{s}}| \leq \delta, 0 \leq \tilde{s} \leq t\}|] \\
&= \lambda^3 \cdot \mathrm{E}[|\mathcal{W}_\delta(t)|] \; ,
\end{aligned}$$

so that one assume without loss of generality that $\delta = 1$: knowing $\mathrm{E}[|\mathcal{W}_1(t)|]$, we get the general case with the formula $\mathrm{E}[|\mathcal{W}_\delta(t)|] = \delta^3 \cdot \mathrm{E}[|\mathcal{W}_1(\delta^{-2}t)|]$.

Let K be the closed unit ball in $\mathbb{R}^d$. Define the hitting probability

$$f(x,t) = \mathrm{P}[x + B_s \in K; 0 \leq s \leq t] \; .$$

We have

$$\mathrm{E}[|\mathcal{W}_1(t)|] = \int_{\mathbb{R}^d} f(x,t)\, dx \; .$$

Proof.

$$\begin{aligned}
\mathrm{E}[|\mathcal{W}_1(t)|] &= \int\int \mathrm{P}[x \in \mathcal{W}_1(t)]\, dx\, dB \\
&= \int\int \mathrm{P}[B_s - x \in K; 0 \leq s \leq t]\, dx\, dB \\
&= \int\int \mathrm{P}[B_s - x \in K; 0 \leq s \leq t]\, dB\, dx \\
&= \int f(x,t)\, dx \; .
\end{aligned}$$

The hitting probability is radially symmetric and can be computed explicitly in terms of $r = |x|$: for $|x| \geq 1$, one has

$$f(x,t) = \frac{2}{r\sqrt{2\pi t}} \int_0^\infty e^{-\frac{(|x|+z-1)^2}{2t}} \, dz \, .$$

Proof. The kernel of e^{-tH} satisfies the heat equation

$$\partial_t p(x,0,t) = (\Delta/2)p(x,0,t)$$

inside D. From the previous lemma follows that $\dot{f} = (\Delta/2)f$, so that the function $g(r,t) = rf(x,t)$ satisfies $\dot{g} = \frac{\partial^2}{2(\partial r)^2} g(r,t)$ with boundary condition $g(r,0) = 0, g(1,t) = 1$. We compute

$$\int_{|x|\geq 1} f(x,t) \, dx = 2\pi t + 4\sqrt{2\pi t}$$

and $\int_{|x|\leq 1} f(x,t) \, dx = 4\pi/3$ so that

$$\mathrm{E}[|\mathcal{W}_1(t)| = 2\pi t + 4\sqrt{2\pi t} + 4\pi/3 \, .$$

□

Corollary 4.15.3. In three dimensions, one has:

$$\lim_{\delta\to 0} \frac{1}{\delta} \mathrm{E}[|\mathcal{W}_\delta(t)|] = 2\pi t$$

and

$$\lim_{t\to\infty} \frac{1}{t} \cdot \mathrm{E}[|\mathcal{W}_\delta(t)|] = 2\pi\delta \, .$$

Proof. The proof follows immediately from Spitzer's theorem (4.15.2). □

Remark. If Brownian motion were one-dimensional, then $\delta^{-2}\mathrm{E}[|\mathcal{W}_\delta(t)|]$ would stay bounded as $\delta \to 0$. The corollary shows that the Wiener sausage is quite "fat". Brownian motion is rather "two-dimensional".

Remark. Kesten, Spitzer and Wightman have got stronger results. It is even true that $\lim_{\delta\to 0} |W_\delta(t)|/t = 2\pi\delta$ and $\lim_{t\to\infty} |W_\delta(t)|/t = 2\pi\delta$ for almost all paths.

4.16 The Ito integral for Brownian motion

We start now to develop stochastic integration first for Brownian motion and then more generally for continuous martingales. Lets start with a motivation. We know by theorem (4.2.5) that almost all paths of Brownian motion are not differentiable. The usual Lebesgue-Stieltjes integral

$$\int_0^t f(B_s)\dot{B}_s \, ds$$

can therefore not be defined. We are first going to see, how a stochastic integral can still be constructed. Actually, we were already dealing with a special case of stochastic integrals, namely with Wiener integrals $\int_0^t f(B_s) \, dB_s$, where f is a function on $C([0,\infty], \mathbb{R}^d)$ which can contain for example $\int_0^t V(B_s) \, ds$ as in the Feynman-Kac formula. But the result of this integral was a **number** while the stochastic integral, we are going to define, will be a **random variable**.

Definition. Let B_t be the one-dimensional Brownian motion process and let f be a function $f : \mathbb{R} \to \mathbb{R}$. Define for $n \in \mathbb{N}$ the random variable

$$J_n(f) = \sum_{m=1}^{2^n} f(B_{(m-1)2^{-n}})(B_{m2^{-n}} - B_{(m-1)2^{-n}}) =: \sum_{m=1}^{2^n} J_{n,m}(f) \, .$$

We will use later for $J_{n,m}(f)$ also the notation $f(B_{t_{m-1}})\delta_n B_{t_m}$, where $\delta_n B_t = B_t - B_{t-2^{-n}}$.

Remark. We have earlier defined the discrete stochastic integral for a previsible process C and a martingale X

$$(\int C \, dX)_n = \sum_{m=1}^{n} C_m(X_m - X_{m-1}) \, .$$

If we want to take for C a function of X, then we have to take $C_m = f(X_{m-1})$. This is the reason, why we have to take the differentials $\delta_n B_{t_m}$ to "stick out into future".

The stochastic integral is a limit of discrete stochastic integrals:

Lemma 4.16.1. If $f \in C^1(\mathbb{R})$ such that f, f' are bounded on $\mathbb{R}$, then $J_n(f)$ converges in $\mathcal{L}^2$ to a random variable

$$\int_0^1 f(B_s) \, dB = \lim_{n \to \infty} J_n$$

satisfying

$$|| \int_0^1 f(B_s) \, dB||_2^2 = \mathrm{E}[\int_0^1 f(B_s)^2 \, ds] \, .$$

Proof. (i) For $i \neq j$ we have $\mathrm{E}[J_{n,i}(f)J_{n,j}(f)] = 0$.
Proof. For $j > i$, there is a factor $B_{j2^{-n}} - B_{(j-1)2^{-n}}$ of $J_{n,i}(f)J_{n,j}(f)$ independent of the rest of $J_{n,i}(f)J_{n,j}(f)$ and the claim follows from $\mathrm{E}[B_{j2^{-n}} - B_{(j-1)2^{-n}}] = 0$.

(ii) $\mathrm{E}[J_{n,m}(f)^2] = \mathrm{E}[f(B_{(m-1)2^{-n}})^2]2^{-n}$.
Proof. $f(B_{(m-1)/2^n})$ is independent of $(B_{m2^{-n}} - B_{(m-1)2^{-n}})^2$ which has expectation 2^{-n}.

(iii) From *(ii)* follows

$$||J_n(f)||_2 = \sum_{m=1}^{2^n} \mathrm{E}[f(B_{(m-1)2^{-n}})^2]2^{-n} \, .$$

(iv) The claim: J_n converges in $\mathcal{L}^2$.
Since $f \in C^1$, there exists $C = ||f'||_\infty^2$ and this gives $|f(x) - f(y)|^2 \leq C \cdot |x-y|^2$. We get

$$\begin{aligned} & ||J_{n+1}(f) - J_n(f)||_2^2 \\ = \; & \sum_{m=1}^{2^n-1} \mathrm{E}[(f(B_{(2m+1)2^{-(n+1)}}) - f(B_{(2m)2^{-(n+1)}}))^2]2^{-(n+1)} \\ \leq \; & C\sum_{m=1}^{2^n-1} \mathrm{E}[(B_{(2m+1)2^{-(n+1)}} - B_{(2m)2^{-(n+1)}})^2]2^{-(n+1)} \\ = \; & C \cdot 2^{-n-2} \, , \end{aligned}$$

where the last equality followed from the fact that $\mathrm{E}[(B_{(2m+1)2^{-(n+1)}} - B_{(2m)2^{-(n+1)}})^2] = 2^{-n}$ since B is Gaussian. We see that J_n is a Cauchy sequence in $\mathcal{L}^2$ and has therefore a limit.

(v) The claim $|| \int_0^1 f(B_s) \, dB||_2^2 = \mathrm{E}[\int_0^1 f(B_s)^2 \, ds]$.
Proof. Since $\sum_m f(B_{(m-1)2^{-n}})^2 2^{-n}$ converges point wise to $\int_0^1 f(B_s)^2 \, ds$, (which exists because f and B_s are continuous), and is dominated by $||f||_\infty^2$, the claim follows since J_n converges in $\mathcal{L}^2$. □

We can extend the integral to functions f, which are locally L^1 and bounded near 0. We write $L^p_{loc}(\mathbb{R})$ for functions f which are in $L^p(I)$ when restricted to any finite interval I on the real line.

Corollary 4.16.2. $\int_0^1 f(B_s) \, dB$ exists as a $\mathcal{L}^2$ random variable for $f \in L^1_{loc}(\mathbb{R}) \cap L^\infty(-\epsilon, \epsilon)$ and any $\epsilon > 0$.

Proof. (i) If $f \in L^1_{loc}(\mathbb{R}) \cap L^\infty(-\epsilon, \epsilon)$ for some $\epsilon > 0$, then

$$\mathrm{E}[\int_0^1 f(B_s)^2 \, ds] = \int_0^1 \int_{\mathbb{R}} \frac{f(x)^2}{\sqrt{2\pi s}} e^{-x^2/2s} \, dx \, ds \; < \infty .$$

(ii) If $f \in L^1_{loc}(\mathbb{R}) \cap L^\infty(-\epsilon, \epsilon)$, then for almost every $B(\omega)$, the limit

$$\lim_{a \to \infty} \int_0^1 1_{[-a,a]}(B_s) f(B_s)^2 \, ds$$

exists point wise and is finite.
Proof. B_s is continuous for almost all ω so that $1_{[-a,a]}(B_s)f(B)$ is independent of a for large a. The integral $\mathrm{E}[\int_0^1 1_{[-a,a]}(B_s)f(B_s)^2 \, ds]$ is bounded by $\mathrm{E}[f(B_s)^2 \, ds] < \infty$ by (i).

(iii) The claim.
Proof. Assume $f \in L^1_{loc}(\mathbb{R}) \cap L^\infty(-\epsilon, \epsilon)$. Given $f_n \in C^1(\mathbb{R})$ with $1_{[-a,a]} f_n \to f$ in $L^2(\mathbb{R})$.
By the dominated convergence theorem, we have

$$\int 1_[- a, a] f_n(B_s) \, dB \to \int 1_[- a, a] f(B_s) \, dB$$

in $\mathcal{L}^2$. Since by (ii), the $\mathcal{L}^2$ bound is independent of a, we can also pass to the limit $a \to \infty$. □

Definition. This integral is called an **Ito integral**. Having the one-dimensional integral allows also to set up the integral in higher dimensions: with Brownian motion in $\mathbb{R}^d$ and $f \in L^2_{loc}(\mathbb{R}^d)$ define the integral $\int_0^1 f(B_s) \, dB_s$ component wise.

Lemma 4.16.3. For $n \to \infty$,

$$\sum_{j=1}^{2^n} J_{n,j}(1)^2 = \sum_{j=1}^{2^n} (B_{j/2^n} - B_{(j-1)/2^n})^2 \to 1 .$$

Proof. By definition of Brownian motion, we know that for fixed n, $J_{n,j}$ are $N(0, 2^{-n})$-distributed random variables and so

$$\mathrm{E}[\sum_{j=1}^{2^n} J_{n,j}(1)^2] = 2^n \cdot \mathrm{Var}[B_{j/2^n} - B_{(j-1)/2^n}] = 2^n 2^{-n} = 1 .$$

Now, $X_j = 2^n J_{n,j}$ are IID $N(0,1)$-distributed random variables so that by the law of large numbers

$$\frac{1}{2^n} \sum_{j=1}^{2^n} X_j \to 1$$

for $n \to \infty$. □

The formal rules of integration do not hold for this integral. We have for example in one dimension:

$$\int_0^1 B_s \, dB = \frac{1}{2}(B_1^2 - 1) \neq \frac{1}{2}(B_1^2 - B_0^2) .$$

Proof. Define

$$\begin{aligned} J_n^- &= \sum_{m=1}^{2^n} f(B_{(m-1)2^{-n}})(B_{m2^{-n}} - B_{(m-1)2^{-n}}) , \\ J_n^+ &= \sum_{m=1}^{2^n} f(B_{m2^{-n}})(B_{m2^{-n}} - B_{(m-1)2^{-n}}) . \end{aligned}$$

The above lemma implies that $J_n^+ - J_n^- \to 1$ almost everywhere for $n \to \infty$ and we check also $J_n^+ + J_n^- = B_1^2$. Both of these identities come from cancellations in the sum and imply together the claim. □

We mention now some trivial properties of the stochastic integral.

Theorem 4.16.4 (Properties of the Ito integral). Here are some basic properties of the Ito integral:
(1) $\int_0^t f(B_s) + g(B_s) \, dB_s = \int_0^t f(B_s) \, dB_s + \int_0^t g(B_s) \, dB_s$.
(2) $\int_0^t \lambda \cdot f(B_s) \, dB_s = \lambda \cdot \int_0^t f(B_s) \, dB_s$.
(3) $t \mapsto \int_0^t f(B_s) \, dB_s$ is a continuous map from $\mathbb{R}^+$ to $\mathcal{L}^2$.
(4) $\mathrm{E}[\int_0^t f(B_s) \, dB_s] = 0$.
(5) $\int_0^t f(B_s) \, dB_s$ is $\mathcal{A}_t$ measurable.

Proof. (1) and (2) follow from the definition of the integral.
For (3) define $X_t = \int_0^t f(B_s) \, dB$. Since

$$\begin{aligned} ||X_t - X_{t+\epsilon}||_2^2 &= \mathrm{E}[\int_t^{t+\epsilon} f(B_s)^2 \, ds] \\ &= \int_t^{t+\epsilon} \int_{\mathbb{R}} \frac{f(x)^2}{\sqrt{2\pi s}} e^{-x^2/2s} \, dx \, ds \to 0 \end{aligned}$$

for $\epsilon \to 0$, the claim follows.
(4) and (5) can be seen by verifying it first for elementary functions f. □

It will be useful to consider an other generalizations of the integral.

Definition. If $dW = dxdB$ is the Wiener measure on $\mathbb{R}^d \times C([0, \infty)$, define

$$\int_0^t f(W_s) \, dW_s = \int_{\mathbb{R}^d} \int_0^t f(x + B_s) \, dB_s \, dx .$$

Definition. Assume f is also time dependent so that it is a function on $\mathbb{R}^d \times \mathbb{R}$. As long as $\mathrm{E}[\int_0^1 |f(B_s,s)|^2 \, ds] < \infty$, we can also define the integral

$$\int_0^t f(B_s,s) \, ds \, .$$

The following formula is useful for understanding and calculating stochastic integrals. It is the "fundamental theorem for stochastic integrals" and allows to do "change of variables" in stochastic calculus similarly as the fundamental theorem of calculus does for usual calculus.

Theorem 4.16.5 (Ito's formula). For a C^2 function $f(x)$ on $\mathbb{R}^d$

$$f(B_t) - f(B_0) = \int_0^t \nabla f(B_s) \cdot \, dB_s + \frac{1}{2} \int_0^t \Delta f(B_s) \, ds \, .$$

If B_s would be an ordinary path in $\mathbb{R}^d$ with velocity vector $dB_s = \dot{B}_s \, ds$, then we had

$$f(B_t) - f(B_0) = \int_0^t \nabla f(B_s) \cdot \, \dot{B}_s \, ds$$

by the fundamental theorem of line integrals in calculus. It is a bit surprising that in the stochastic setup, a second derivative Δf appears in a first order differential. One writes sometimes the formula also in the differential form

$$df = \nabla f \, dB + \frac{1}{2} \Delta f \, dt \, .$$

Remark. We cite [11]: "Ito's formula is now the bread and butter of the "quant" department of several major financial institutions. Models like that of **Black-Scholes** constitute the basis on which a modern business makes decisions about how everything from stocks and bonds to pork belly futures should be priced. Ito's formula provides the link between various stochastic quantities and differential equations of which those quantities are the solution." For more information on the Black-Scholes model and the famous Black-Scholes formula, see [16].
It is not much more work to prove a more general formula for functions $f(x,t)$, which can be time-dependent too:

Theorem 4.16.6 (Generalized Ito formula). Given a function $f(x,t)$ on $\mathbb{R}^d \times [0,t]$ which is twice differentiable in x and differentiable in t. Then

$$f(B_t,t) - f(B_0,0) = \int_0^t \nabla f(B_s,s) \cdot \, dB_s + \frac{1}{2} \int_0^t \Delta f(B_s,s) \, ds + \int_0^t \dot{f}(B_s,s) \, ds \, .$$

In differential notation, this means

$$df = \nabla f \, dB + (\frac{1}{2}\Delta f + \dot{f}) \, dt \, .$$

Proof. By a change of variables, we can assume $t = 1$. For each n, we discretized time

$$\{0 < 2^{-n} < \dots, t_k = k \cdot 2^{-n}, \cdots, 1\}$$

and define $\delta_n B_{t_k} = B_{t_k} - B_{t_{k-1}}$. We write

$$\begin{aligned} f(B_1, 1) \quad - \quad & f(B_0, 0) = \sum_{k=1}^{2^n} (\nabla f)(B_{t_{k-1}}, t_{k-1}) \delta_n B_{t_k} \\ + \quad & \sum_{k=1}^{2^n} f(B_{t_k}, t_{k-1}) - f(B_{t_{k-1}}, t_{k-1}) - (\nabla f)(B_{t_{k-1}}, t_{k-1}) \delta_n B_{t_k} \\ + \quad & \sum_{k=1}^{2^n} f(B_{t_k}, t_k) - f(B_{t_k}, t_{k-1}) \\ = \quad & I_n + II_n + III_n \, . \end{aligned}$$

(i) By definition of the Ito integral, the first sum I_n converges in $\mathcal{L}^2$ to $\int_0^1 (\nabla f)(B_s, s) \, dB_s$.

(ii) If $p > 2$, we have $\sum_{k=1}^{2^n} |\delta_n B_{t_k}|^p \to 0$ for $n \to \infty$.
Proof. $\delta_n B_{t_k}$ is a $N(0, 2^{-n})$-distributed random variable so that

$$\mathrm{E}[|\delta_n B_{t_k}|^p] = (2\pi)^{-1/2} 2^{-(np)/2)} \int_{-\infty}^{\infty} |x|^p e^{-x^2/2} \, dx = C 2^{-(np)/2} \, .$$

This means

$$\mathrm{E}[\sum_{k=1}^{2^n} |\delta_n B_{t_k}|^p] = C 2^n 2^{-(np)/2}$$

which goes to zero for $n \to \infty$ and $p > 2$.

(iii) $\sum_{k=1}^{2^n} \mathrm{E}[(B_{t_k} - B_{t_{k-1}})^4] \to 0$ follows from (ii). We have therefore

$$\begin{aligned} \sum_{k=1}^{2^n} \mathrm{E}[g(B_{t_k}, t_k)^2 ((B_{t_k} - B_{t_{k-1}})^2 - 2^{-n})^2] \quad & \leq \quad C \sum_{k=1}^{2^n} \mathrm{Var}[(B_{t_k} - B_{t_{k-1}})^2] \\ & \leq \quad C \sum_{k=1}^{2^n} \mathrm{E}[(B_{t_k} - B_{t_{k-1}})^4] \to 0 \, . \end{aligned}$$

(iv) Using a Taylor expansion

$$f(x) = f(y) - \nabla f(y)(x - y) - \frac{1}{2} \sum_{i,j} \partial_{x_i x_j} f(y)(x - y)_i (x - y)_j + O(|x - y|^3 \, ,$$

we get for $n \to \infty$

$$II_n - \sum_{k=1}^{2^n} \frac{1}{2} \sum_{i,j} \partial_{x_i x_j} f(B_{t_{k-1}}, t_{k-1}) (\delta_n B_{t_k})_i (\delta_n B_{t_k})_j \to 0$$

in $\mathcal{L}^2$. Since

$$\sum_{k=1}^{2^n} \frac{1}{2} \partial_{x_i x_j} f(B_{t_{k-1}}, t_{k-1}) [(\delta_n B_{t_k})_i (\delta_n B_{t_k})_j - \delta_{ij} 2^{-n}]$$

goes to zero in $\mathcal{L}^2$ (applying (ii) for $g = \partial_{x_i x_j} f$ and note that $(\delta_n B_{t_k})_i$ and $(\delta_n B_{t_k})_j$ are independent for $i \neq j$), we have therefore

$$II_n \to \frac{1}{2} \int_0^t \Delta f(B_s, s) \, ds$$

in $\mathcal{L}^2$.

(v) A Taylor expansion with respect to t

$$f(x,t) - f(x,s) - \dot{f}(x,s)(t-s) + O((t-s)^2)$$

gives

$$III_n \to \int_0^t \dot{f}(B_s, s) \, ds$$

in $\mathcal{L}^1$ because $s \to f(B_s, s)$ is continuous and III_n is a Riemann sum approximation. □

Example. Consider the function

$$f(x,t) = e^{\alpha x - \alpha^2 t/2} .$$

Because this function satisfies the heat equation $\dot{f} + f''/2 = 0$, we get from Ito's formula

$$f(B_t, t) - f(B_0, t) = \alpha \int_0^t f(B_s, s) \cdot dB_s .$$

We see that for functions satisfying the heat equation $\dot{f} + f''/2 = 0$ Ito's formula reduces to the usual rule of calculus. If we make a power expansion in α of

$$\int_0^t e^{\alpha B_s - \alpha^2 s/2} \, dB = \frac{1}{\alpha} e^{\alpha B_s - \alpha^2 s/2} - \frac{1}{\alpha} ,$$

we get other formulas like

$$\int_0^t B_s \, dB = \frac{1}{2}(B_t^2 - t) .$$

Wick ordering.
There is a notation used in **quantum field theory** developed by **Gian-Carlo Wick** at about the same time as Ito's invented the integral. This **Wick ordering** is a map on polynomials $\sum_{i=1}^n a_i x^i$ which leave monomials (polynomials of the form $x^n + a_{n-1}x^{n-1}\cdots$) invariant.

Definition. Let

$$\Omega_n(x) = \frac{H_n(x)\Omega_0(x)}{\sqrt{2^n n!}}$$

be the n'-th eigenfunction of the quantum mechanical oscillator. Define

$$: x^n := \frac{1}{2^{n/2}} H_n(\frac{x}{\sqrt{2}})$$

and extend the definition to all polynomials by linearity. The Polynomials $: x^n :$ are orthogonal with respect to the measure $\Omega_0^2 dy = \pi^{-1/2} e^{-y^2}\ dy$ because we have seen that the eigenfunctions Ω_n are orthonormal.

Example. Here are the first **Wick powers**:

$$\begin{aligned} : x : &= x \\ : x^2 : &= x^2 - 1 \\ : x^3 : &= x^3 - 3x \\ : x^4 : &= x^4 - 6x^2 + 3 \\ : x^5 : &= x^5 - 10x^3 + 15x \ . \end{aligned}$$

Definition. The multiplication operator $Q : f \mapsto xf$ is called the **position operator**. By definition of the creation and annihilation operators one has $Q = \frac{1}{\sqrt{2}}(A + A^*)$.

The following formula indicates, why Wick ordering has its name and why it is useful in quantum mechanics:

Proposition 4.16.7. As operators, we have the identity

$$: Q^n := \frac{1}{2^{n/2}} : (A + A^*)^n := \frac{1}{2^{n/2}} \sum_{j=0}^{n} \binom{n}{j} (A^*)^j A^{n-j} \ .$$

Definition. Define $L = \sum_{j=0}^n \binom{n}{j} (A^*)^j A^{n-j}$.

Proof. Since we know that Ω_n forms a basis in L^2, we have only to verify that $: Q^n : \Omega_k = 2^{-n/2} L \Omega_k$ for all k. From

$$\begin{aligned} 2^{-1/2}[Q, L] &= [A + A^*, \sum_{j=0}^{n} \binom{n}{j} (A^*)^j A^{n-j}] \\ &= \sum_{j=0}^{n} \binom{n}{j} j (A^*)^{j-1} A^{n-j} - (n-j)(A^*)^j A^{n-j-1} \\ &= 0 \end{aligned}$$

we obtain by linearity $[H_k(\sqrt{2}Q), L]$. Because $: Q^n : \Omega_0 = 2^{-n/2}(n!)^{1/2}\Omega_n = 2^{-n/2}(A^*)^n\Omega_0 = 2^{-n/2}L\Omega_0$, we get

$$\begin{aligned} 0 &= (: Q^n : -2^{-n/2}L)\Omega_0 \\ &= (k!)^{-1/2} H_k(\sqrt{s}Q)(: Q^n : -2^{-n/2}L)\Omega_0 \\ &= (: Q^n : -2^{-n/2}L)(k!)^{-1/2} H_k(\sqrt{s}Q)\Omega_0 \\ &= (: Q^n : -2^{-n/2}L)\Omega_k \ . \end{aligned}$$

□

Remark. The new ordering made the operators A, A^* behaves as if A, B would commutate. even so they don't: they satisfy the commutation relations $[A, A^*] = 1$:

The fact that stochastic integration is relevant to quantum mechanics can be seen from the following formula for the Ito integral:

Theorem 4.16.8 (Ito Integral of B^n)**.** Wick ordering makes the Ito integral behave like an ordinary integral.

$$\int_0^t : B_s^n : \, dB_s = \frac{1}{n+1} : B_t^{n+1} : \ .$$

Remark. Notation can be important to make a concept appear natural. An other example, where an adaption of notation helps is **quantum calculus**, "calculus without taking limits" [44], where the derivative is defined as $D_q f(x) = d_q f(x)/d_q(x)$ with $d_q f(x) = f(qx) - f(x)$. One can see that $D_q x^n = [n] x^{n-1}$, where $[n] = \frac{q^n - 1}{q-1}$. The limit $q \to 1$ corresponds to the classical limit case $\hbar \to 0$ of quantum mechanics.

Proof. By rescaling, we can assume that $t = 1$.
We prove all these equalities simultaneously by showing

$$\int_0^1 : e^{\alpha B_s} : \, dB = \alpha^{-1} : e^{\alpha B_1} : -\alpha^{-1} \ .$$

The generating function for the Hermite polynomials is known to be

$$\sum_{n=0}^{\infty} H_n(x)\frac{\alpha^n}{n!} = e^{\alpha\sqrt{2}x-\frac{\alpha^2}{2}} .$$

(We can check this formula by multiplying it with Ω_0, replacing x with $x/\sqrt{2}$ so that we have

$$\sum_{n=0}^{\infty} \frac{\Omega_n(x)\alpha^n}{(n!)^{1/2}} = e^{\alpha x-\frac{\alpha^2}{2}-\frac{x^2}{2}} .$$

If we apply A^* on both sides, the equation goes onto itself and we get after k such applications of A^* that that the inner product with Ω_k is the same on both sides. Therefore the functions must be the same.)
This means

$$: e^{\alpha x} := \sum_{j=0}^{\infty} \frac{\alpha^n : x^n :}{n!} = e^{\alpha x - \frac{1}{2}\alpha^2} .$$

Since the right hand side satisfies $\dot{f} + f''/2 = 2$, the claim follows from the Ito formula for such functions. □

We can now determine all the integrals $\int B_s^n \, dB$:

$$\begin{aligned}
\int_0^t 1 \, dB &= B_t \\
\int_0^t B_s \, dB &= \frac{1}{2}(B_t^2 - 1) \\
\int_0^t B_s^2 \, dB &= \int_0^t : B_s^2 : +1 \, dB = B_t + \frac{1}{3}(: B_t :^3) = B_t + \frac{1}{3}(B_t^3 - 3B_t)
\end{aligned}$$

and so on.

Stochastic integrals for the oscillator and the Brownian bridge process. Let $Q_t = e^{-t}B_{e^{2t}}/\sqrt{2}$ the oscillator process and $A_t = (1-t)B_{t/(1-t)}$ the Brownian bridge. If we define new discrete differentials

$$\begin{aligned}
\delta_n Q_{t_k} &= Q_{t_{k+1}} - e^{-(t_{k+1}-t_k)}Q_{t_k} \\
\delta_n A_{t_k} &= A_{t_{k+1}} - A_{t_k} + \frac{t_{k+1}-t_k}{(1-t)}A_{t_k}
\end{aligned}$$

the stochastic integrals can be defined as in the case of Brownian motion as a limit of discrete integrals.

Feynman-Kac formula for Schrödinger operators with magnetic fields. Stochastic integrals appear in the Feynman-Kac formula for particles moving in a magnetic field. Let $A(x)$ be a vector potential in $\mathbb{R}^3$ which gives the magnetic field $B(x) = \mathrm{curl}(A)$. Quantum mechanically, a particle moving in an magnetic field together with an external field is described by the Hamiltonian

$$H = (i\nabla + A)^2 + V .$$

In the case $A = 0$, we get the usual Schrödinger operator. The Feynman-Kac formula is the Wiener integral

$$e^{-tH}u(0) = \int e^{-F(B,t)}u(B_t)\,dB ,$$

where $F(B,t)$ is a stochastic integral.

$$F(B,t) = i\int a(B_s)\,dB + \frac{i}{2}\int_0^t \mathrm{div}(A)\,ds + \int_0^t V(B_s)\,ds .$$

4.17 Processes of bounded quadratic variation

We develop now the stochastic Ito integral with respect to general martingales. Brownian motion B will be replaced by a martingale M which are assumed to be in $\mathcal{L}^2$. The aim will be to define an integral

$$\int_0^t K_s\,dM_s ,$$

where K is a progressively measurable process which satisfies some boundedness condition.

Definition. Given a right-continuous function $f : [0,\infty) \to \mathbb{R}$. For each finite subdivision

$$\Delta = \{0 = t_0, t_1, \dots, t = t_n\}$$

of the interval $[0,t]$ we define $|\Delta| = \sup_{i=1}^r |t_{i+1} - t_i|$ called the **modulus** of Δ. Define

$$||f||_\Delta = \sum_{i=0}^{n-1} |f_{t_{i+1}} - f_{t_i}| .$$

A function with finite total variation $||f||_t = \sup_\Delta ||f||_\Delta < \infty$ is called a function of **finite variation**. If $\sup_t |f|_t < \infty$, then f is called of **bounded variation**. One abbreviates, bounded variation with BV.

Example. Differentiable C^1 functions are of finite variation. Note that for functions of finite variations, V_t can go to ∞ for $t \to \infty$ but if V_t stays bounded, we have a function of bounded variation. Monotone and bounded functions are of finite variation. Sums of functions of bounded variation are of bounded variation.

Remark. Every function of finite variation can be written as $f = f^+ - f^-$, where $f^\pm$ are both positive and increasing. Proof: define $f^\pm = (\pm f_t + ||f||_t)/2$.

Remark. Functions of bounded variation are in one to one correspondence to Borel measures on $[0, \infty)$ by the Stieltjes integral $\int_0^t |df| = f_t^+ + f_t^-$.

Definition. A process X_t is called **increasing** if the paths $X_t(\omega)$ are finite, right-continuous and increasing for almost all $\omega \in \Omega$. A process X_t is called of **finite variation**, if the paths $X_t(\omega)$ are finite, right-continuous and of finite variation for almost all $\omega \in \Omega$.

Remark. Every bounded variation process A can be written as $A_t = A_t^+ - A_t^-$, where $A_t^\pm$ are increasing. The process $V_t = \int_0^t |dA|_s = A_t^+ + A_t^-$ is increasing and we get for almost all $\omega \in \Omega$ a measure called the **variation** of A.

If X_t is a bounded $\mathcal{A}_t$-adapted process and A is a process of bounded variation, we can form the Lebesgue-Stieltjes integral

$$(X \cdot A)_t(\omega) = \int_0^t X_s(\omega)\, dA_s(\omega) .$$

We would like to define such an integral for martingales. The problem is:

Proposition 4.17.1. A continuous martingale M is never of finite variation, unless it is constant.

Proof. Assume M is of finite variation. We show that it is constant.

(i) We can assume without loss of generality that M is of bounded variation.
Proof. Otherwise, we can look at the martingale M^{S_n}, where S_n is the stopping time $S_n = \inf\{s \mid V_s \geq n\}$ and V_t is the variation of M on $[0, t]$.

(ii) We can also assume also without loss of generality that $M_0 = 0$.

(iii) Let $\Delta = \{t_0 = 0, t_1, \dots, t_n = t\}$ be a subdivision of $[0, t]$. Since M is a martingale, we have by Pythagoras

$$\begin{aligned}
\mathrm{E}[M_t^2] &= \mathrm{E}[\sum_{i=0}^{k-1} (M_{t_{i+1}}^2 - M_{t_i}^2)] \\
&= \mathrm{E}[\sum_{i=1}^{k-1} (M_{t_{i+1}} - M_{t_i})(M_{t_{i+1}} + M_{t_i})] \\
&= \mathrm{E}[\sum_{i=1}^{k-1} (M_{t_{i+1}} - M_{t_i})^2]
\end{aligned}$$

and so

$$\mathrm{E}[M_t^2] \leq \mathrm{E}[V_t(\sup_i |M_{t_{i+1}} - M_{t_i}|] \leq K \cdot \mathrm{E}[\sup_i |M_{t_{i+1}} - M_{t_i}|] \,.$$

If the modulus $|\Delta|$ goes to zero, then the right hand side goes to zero since M is continuous. Therefore $M = 0$. □

Remark. This proposition applies especially for Brownian motion and underlines the fact that the stochastic integral could not be defined point wise by a Lebesgue-Stieltjes integral.

Definition. If $\Delta = \{t_0 = 0 < t_1 < \dots\}$ is a subdivision of $\mathbb{R}^+ = [0, \infty)$ with only finitely many points $\{t_0, t_1, \dots, t_k\}$ in each interval $[0, t]$, we define for a process X

$$T_t^\Delta = T_t^\Delta(X) = (\sum_{i=0}^{k-1} (X_{t_{i+1}} - X_{t_i})^2) + (X_t - X_{t_k})^2 \,.$$

The process X is called of **finite quadratic variation**, if there exists a process $< X, X >$ such that for each t, the random variable T_t^Δ converges in probability to $< X, X >_t$ as $|\Delta| \to 0$.

Theorem 4.17.2 (Doob-Meyer decomposition). Given a continuous and bounded martingale M of finite quadratic variation. Then $< M, M >$ is the unique continuous increasing adapted process vanishing at zero such that $M^2 - < M, M >$ is a martingale.

Remark. Before we enter the not so easy proof given in [83], let us mention the corresponding result in the discrete case (see theorem (3.5.1), where M^2 was a submartingale so that M^2 could be written uniquely as a sum of a martingale and an increasing previsible process.

Proof. Uniqueness follows from the previous proposition: if there would be two such continuous and increasing processes A, B, then $A - B$ would be a continuous martingale with bounded variation (if A and B are increasing they are of bounded variation) which vanishes at zero. Therefore $A = B$.

(i) $M_t^2 - T_t^\Delta(M)$ is a continuous martingale.
Proof. For $t_i < s < t_{i+1}$, we have from the martingale property using that $(M_{t_{i+1}} - M_s)^2$ and $(M_s - M_{t_i})^2$ are independent,

$$\mathrm{E}[(M_{t_{i+1}} - M_{t_i})^2 \mid \mathcal{A}_s] = \mathrm{E}[(M_{t_{i+1}} - M_s)^2 | \mathcal{A}_s] + (M_s - M_{t_i})^2 \,.$$

This implies with $0 = t_0 < t_1 < \cdots < t_l < s < t_{l+1} < \cdots < t_k < t$ and using orthogonality

$$\begin{aligned} \mathrm{E}[T_t^\Delta(M) - T_s^\Delta(M)|\mathcal{A}_s] &= \mathrm{E}[\sum_{j=l}^{k}(M_{t_{j+1}} - M_{t_j})^2|\mathcal{A}_s] \\ &+ \mathrm{E}[(M_t - M_{t_k})^2|\mathcal{A}_s] + \mathrm{E}[(M_s - M_{t_l})^2|\mathcal{A}_s] \\ &= \mathrm{E}[(M_t - M_s)^2|\mathcal{A}_s] = \mathrm{E}[M_t^2 - M_s^2|\mathcal{A}_s] \, . \end{aligned}$$

This implies that $M_t^2 - T_t^\Delta(M)$ is a continuous martingale.

(ii) Let C be a constant such that $|M| \leq C$ in $[0, a]$. Then $\mathrm{E}[T_a^\Delta] \leq 4C^2$, independent of the subdivision $\Delta = \{t_0, \dots, t_n\}$ of $[0, a]$.
Proof. The previous computation in (i) gives for $s = 0$, using $T_0^\Delta(M) = 0$

$$\mathrm{E}[T_t^\Delta(M)|\mathcal{A}_0] = \mathrm{E}[M_t^2 - M_0^2|\mathcal{A}_0] \leq \mathrm{E}[(M_t - M_0)(M_t + M_0)] \leq 4C^2 \, .$$

(iii) For any subdivision Δ, one has $\mathrm{E}[(T_a^\Delta)^2] \leq 48C^4$.
Proof. We can assume $t_n = a$. Then

$$\begin{aligned} (T_a^\Delta(M))^2 &= (\sum_{k=1}^{n}(M_{t_k} - M_{t_{k-1}})^2)^2 \\ &= 2\sum_{k=1}^{n}(T_a^\Delta - T_{t_k}^\Delta)(T_{t_k}^\Delta - T_{t_{k-1}}^\Delta) + \sum_{k=1}^{n}(M_{t_k} - M_{t_{k-1}})^4 \, . \end{aligned}$$

From (i), we have

$$\mathrm{E}[T_a^\Delta - T_{t_k}^\Delta|\mathcal{A}_{t_k}] = \mathrm{E}[(M_a - M_{t_k})^2 \mid \mathcal{A}_{t_k}]$$

and consequently, using (ii)

$$\begin{aligned} \mathrm{E}[(T_a^\Delta)^2] &= 2\sum_{k=1}^{n}\mathrm{E}[(M_a - M_{t_k})^2(T_{t_k}^\Delta - T_{t_{k+1}}^\Delta)] + \sum_{k=1}^{n}\mathrm{E}[(M_{t_k} - M_{t_{k-1}})^4] \\ &\leq \mathrm{E}[(2\sup_k |M_a - M_{t_k}|^2 + \sup_k |M_{t_k} - M_{t_{k-1}}|^2)T_a^\Delta] \\ &\leq 12C^2\mathrm{E}[T_a^\Delta] \leq 48C^4 \, . \end{aligned}$$

(iii) For fixed $a > 0$ and subdivisions Δ_n of $[0, a]$ satisfying $|\Delta_n| \to 0$, the sequence $T_a^{\Delta_n}$ has a limit in $\mathcal{L}^2$.
Proof. Given two subdivisions Δ', Δ'' of $[0, a]$, let Δ be the subdivision obtained by taking the union of the points of Δ' and Δ''. By (i), the process $X = T^{\Delta'} - T^{\Delta''}$ is a martingale and by (i) again, applied to the martingale X instead of M we have, using $(x + y)^2 \leq 2(x^2 + y^2)$

$$\mathrm{E}[X_a^2] = \mathrm{E}[(T_a^{\Delta'} - T_a^{\Delta''})^2] = \mathrm{E}[T_a^\Delta(X)] \leq 2(E[T_a^\Delta(T^{\Delta'})] + \mathrm{E}[T_a^\Delta(T^{\Delta''})]) \, .$$

We have therefore only to show that $\mathrm{E}[T_a^\Delta(T^{\Delta'})] \to 0$ for $|\Delta'| + |\Delta''| \to 0$. Let s_k be in Δ and t_m the rightmost point in Δ' such that $t_m \leq s_k < s_{k+1} \leq t_{m+1}$. We have

$$\begin{aligned} T_{s_{k+1}}^{\Delta'} - T_{s_k}^{\Delta'} &= (M_{s_{k+1}} - M_{t_m})^2 - (M_{s_k} - M_{t_m})^2 \\ &= (M_{s_{k+1}} - M_{s_k})(M_{s_{k+1}} + M_{s_k} - 2M_{t_m}) \end{aligned}$$

and so

$$T_a^\Delta(T^{\Delta'}) \leq (\sup_k |M_{s_{k+1}} + M_{s_k} - 2M_{t_m}|^2)T_a^\Delta \, .$$

By the Cauchy Schwarz-inequality

$$\mathrm{E}[T_a^\Delta(T^{\Delta'})] \leq \mathrm{E}[\sup_k |M_{s_{k+1}} + M_{s_k} - 2M_{t_m}|^4]^{1/2}\mathrm{E}[(T_a^\Delta)^2]^{1/2}$$

and the first factor goes to 0 as $|\Delta| \to 0$ and the second factor is bounded because of (iii).

(iv) There exists a sequence of $\Delta_n \subset \Delta_{n+1}$ such that $T_t^{\Delta_n}(M)$ converges uniformly to a limit $\langle M, M \rangle$ on $[0, a]$.
Proof. Doob's inequality applied to the discrete time martingale $T^{\Delta_n} - T^{\Delta_m}$ gives

$$\mathrm{E}[\sup_{t \leq a} |T_t^{\Delta_n} - T_t^{\Delta_m}|^2] \leq 4\mathrm{E}[(T_a^{\Delta_n} - T_a^{\Delta_m})^2] \, .$$

Choose the sequence Δ_n such that Δ_{n+1} is a refinement of Δ_n and such that $\bigcup_n \Delta_n$ is dense in $[0, a]$, we can achieve that the convergence is uniform. The limit $\langle M, M \rangle$ is therefore continuous.

(v) $\langle M, M \rangle$ is increasing.
Proof. Take $\Delta_n \subset \Delta_{n+1}$. For any pair $s < t$ in $\bigcup_n \Delta_n$, we have $T_s^{\Delta_n}(M) \leq T_t^{\Delta_n}(M)$ if n is so large that Δ_n contains both s and t. Therefore $\langle M, M \rangle$ is increasing on $\bigcup_n \Delta_n$, which can be chosen to be dense. The continuity of M implies that $\langle M, M \rangle$ is increasing everywhere. □

Remark. The assumption of boundedness for the martingales is not essential. It holds for general martingales and even more generally for so called **local martingales**, stochastic processes X for which there exists a sequence of bounded stopping times T_n increasing to ∞ for which X^{T_n} are martingales.

Corollary 4.17.3. Let M, N be two continuous martingales with the same filtration. There exists a unique continuous adapted process $\langle M, N \rangle$ of finite variation which is vanishing at zero and such that

$$MN - \langle M, N \rangle$$

is a martingale.

Proof. Uniqueness follows again from the fact that a finite variation martingale must be zero. To get existence, use the parallelogram law

$$\langle M, N \rangle = \frac{1}{4}(\langle M + N, M + N \rangle - \langle M - N, M - N \rangle) \, .$$

This is vanishing at zero and of finite variation since it is a sum of two processes with this property.
We know that $M^2 - \langle M,M\rangle, N^2 - \langle N,N\rangle$ and so that $(M \pm N)^2 - \langle M \pm N, M \pm N\rangle$ are martingales. Therefore

$$\begin{aligned} & (M+N)^2 - \langle M+N, M+N\rangle - (M-N)^2 - \langle M-N, M-N\rangle \\ = \; & 4MN - \langle M+N, M+N\rangle - \langle M-N, M-N\rangle \, . \end{aligned}$$

and $MN - \langle M,N\rangle$ is a martingale. □

Definition. The process $\langle M,N\rangle$ is called the **bracket** of M and N and $\langle M,M\rangle$ the increasing process of M.

Example. If $B = (B^{(1)}, \dots, B^{(d)})$ is Brownian motion, then $\langle < B^{(i)}, B^{(j)}\rangle = \delta_{ij} t$ as we have computed in the proof of the Ito formula in the case $t = 1$. It can be shown that every martingale M which has the property that

$$\langle M^{(i)}, M^{(j)}\rangle = \delta_{ij} \cdot t$$

must be Brownian motion. This is **Lévy's characterization of Brownian motion**.

Remark. If M is a martingale vanishing at zero and $\langle M,M\rangle = 0$, then $M = 0$. Since $M_t^2 - \langle M,M\rangle_t$ is a martingale vanishing at zero, we have $\mathrm{E}[M_t^2] = \mathrm{E}[\langle M,M\rangle_t]$.

Remark. Since we have got $\langle M,M\rangle$ as a limit of processes T_t^Δ, we could also write $\langle M,N\rangle$ as such a limit.

4.18 The Ito integral for martingales

In the last section, we have defined for two continuous martingales M,N, the bracket process $\langle M,N\rangle$. Because $\langle M,M\rangle$ was increasing, it was of finite variation and therefore also $\langle M,N\rangle$ is of finite variation. It defines a random measure $d\langle M,N\rangle$.

Theorem 4.18.1 (Kunita-Watanabe inequality). Let M,N be two continuous martingales and H,K two measurable processes. Then for all $p,q \geq 1$ satisfying $1/p + 1/q = 1$, we have for all $t \leq \infty$

$$\begin{aligned} \mathrm{E}[\int_0^t |H_s|\,|K_s|\,|d\langle M,N\rangle_s|] \;\; & \leq \;\; ||(\int_0^t H_s^2 d\langle M,M\rangle)^{1/2}||_p \\ & \cdot \;\; ||(\int_0^t K_s^2 d\langle N,N\rangle)^{1/2}||_q \, . \end{aligned}$$

Proof. (i) Define $\langle M,N\rangle_s^t = \langle M,N\rangle_t - \langle M,N\rangle_s$. Claim: almost surely

$$|\langle M,N\rangle_s^t| \leq (\langle M,M\rangle_s^t)^{1/2}(\langle N,N\rangle_s^t)^{1/2} .$$

Proof. For fixed r, the random variable

$$\langle M,M\rangle_s^t + 2r\langle M,N\rangle_s^t + r^2\langle N,N\rangle_s^t = \langle M+rN, M+rN\rangle_s^t$$

is positive almost everywhere and this stays true simultaneously for a dense set of $r \in \mathbb{R}$. Since M,N are continuous, it holds for all r. The claim follows, since $a+2rb+cr^2 \geq 0$ for al! $r \geq 0$ with nonnegative a,c implies $b \leq \sqrt{a}\sqrt{c}$.

(ii) To prove the claim, it is, using Hölder's inequality, enough to show almost everywhere, the inequality

$$\int_0^t |H_s|\,|K_s|\; d|\langle M,N\rangle|_s \leq (\int_0^t H_s^2 d\langle M,M\rangle)^{1/2} \cdot (\int_0^t K_s^2 d\langle N,N\rangle)^{1/2}$$

holds. By taking limits, it is enough to prove this for $t < \infty$ and bounded K,H. By a density argument, we can also assume the both K and H are step functions $H = \sum_{i=1}^n H_i 1_{J_i}$ and $K = \sum_{i=1}^n K_i 1_{J_i}$, where $J_i = [t_i, t_{i+1})$.

(iii) We get from (i) for step functions H,K as in (ii)

$$\begin{aligned} |\int_0^t H_s K_s d\langle M,N\rangle_s| &\leq \sum_i |H_i K_i||\langle M,N\rangle_{t_i}^{t_{i+1}}| \\ &\leq \sum_i |H_i K_i|((\langle M,M\rangle_{t_i}^{t_{i+1}})^{1/2}((\langle M,M\rangle_{t_i}^{t_{i+1}})^{1/2} \\ &\leq (\sum_i H_i^2 \langle M,M\rangle_{t_i}^{t_{i+1}})^{1/2}(\sum_i K_i^2 \langle N,N\rangle_{t_i}^{t_{i+1}})^{1/2} \\ &= (\int_0^t H_s^2 d\langle M,M\rangle)^{1/2} \cdot (\int_0^t K_s^2 d\langle N,N\rangle)^{1/2} , \end{aligned}$$

where we have used Cauchy-Schwarz inequality for the summation over i. □

Definition. Denote by $\mathcal{H}^2$ the set of $\mathcal{L}^2$-martingales which are $\mathcal{A}_t$-adapted and satisfy

$$||M||_{\mathcal{H}_2} = (\sup_t \mathrm{E}[M_t^2])^{1/2} < \infty .$$

Call H^2 the subset of continuous martingales in $\mathcal{H}^2$ and with H_0^2 the subset of continuous martingales which are vanishing at zero.
Given a martingale $M \in \mathcal{H}^2$, we define $\mathcal{L}^2(M)$ the space of progressively measurable processes K such that

$$||K||_{\mathcal{L}^2(M)}^2 = E[\int_0^\infty K_s^2 d\langle M,M\rangle_s] < \infty .$$

Both $\mathcal{H}^2$ and $\mathcal{L}^2(M)$ are Hilbert spaces.

Lemma 4.18.2. The space H^2 of continuous $\mathcal{L}^2$ martingales is closed in $\mathcal{H}^2$ and so a Hilbert space. Also H_0^2 is closed in H^2 and is therefore a Hilbert space.

Proof. Take a sequence $M^{(n)}$ in H^2 converging to $M \in \mathcal{H}^2$. By Doob's inequality

$$\mathrm{E}[(\sup_t |M_t^n - M_t|)^2] \le 4||M^{(n)} - M||_{\mathcal{H}^2}^2 .$$

We can extract a subsequence, for which $\sup_t |M_t^{(n_k)} - M_t|$ converges point wise to zero almost everywhere. Therefore $M \in H^2$. The same argument shows also that H_0^2 is closed. □

Proposition 4.18.3. Given $M \in H^2$ and $K \in \mathcal{L}^2(M)$. There exists a unique element $\int_0^t KdM \in H_0^2$ such that

$$< \int_0^t KdM, N >= \int_0^t Kd\langle M, N\rangle$$

for every $N \in H^2$. The map $K \mapsto \int_0^t KdM$ is an isometry form $\mathcal{L}^2(M)$ to H_0^2.

Proof. We can assume $M \in H_0$ since in general, we define $\int_0^t K\ dM = \int_0^t K\ d(M - M_0)$.

(i) By the Kunita-Watanabe inequality, we have for every $N \in H_0^2$

$$|\mathrm{E}[\int_0^t K_s d\langle M, N\rangle_s]| \le ||N||_{\mathcal{H}^2} \cdot ||K||_{\mathcal{L}^2(M)} .$$

The map

$$N \mapsto \mathrm{E}[(\int_0^t K_s)\ d\langle M, N\rangle_s]$$

is therefore a linear continuous functional on the Hilbert space H_0^2. By Riesz representation theorem, there is an element $\int K\ dM \in H_0^2$ such that

$$\mathrm{E}[(\int_0^t K_s\ dM_s)N_t] = \mathrm{E}[\int_0^t K_s d\langle M, N\rangle_s]$$

for every $N \in H_0^2$.

(ii) Uniqueness. Assume there exist two martingales $L, L' \in H_0^2$ such that $\langle L, N\rangle = \langle L', N\rangle$ for all $N \in H_0^2$. Then, in particular, $\langle L - L', L - L'\rangle = 0$, from which $L = L'$ follows.

(iii) The integral $K \mapsto \int_0^t K\, dM$ is an isometry because

$$\begin{aligned} \|\int_0^t K\, dM\|_{\mathcal{H}_0}^2 &= \mathrm{E}[(\int_0^\infty K_s\, dM_s)^2] \\ &= \mathrm{E}[\int_0^\infty K_s^2\, d\langle M, M\rangle] \\ &= \|K\|_{\mathcal{L}^2(M)}^2 . \end{aligned}$$

□

Definition. The martingale $\int_0^t K_s\ dM_s$ is called the **Ito integral** of the progressively measurable process K with respect to the martingale M. We can take especially, $K = f(M)$, since continuous processes are progressively measurable. If we take $M = B$, Brownian motion, we get the already familiar Ito integral.

Definition. An $\mathcal{A}_t$ adapted right-continuous process is called a **local martingale** if there exists a sequence T_n of increasing stopping times with $T_n \to \infty$ almost everywhere, such that for every n, the process $X^{T_n} 1_{\{T_n > 0\}}$ is a uniformly integrable $\mathcal{A}_t$-martingale. Local martingales are more general than martingales. Stochastic integration can be defined more generally for local martingales.

We show now that Ito's formula holds also for general martingales. First, a special case, the integration by parts formula.

Theorem 4.18.4 (Integration by parts). Let X, Y be two continuous martingales. Then

$$X_t Y_t - X_0 Y_0 = \int_0^t X_s\, dY_s + \int_0^t Y_s dX_s + \langle X, Y\rangle_t$$

and especially

$$X_t^2 - X_0^2 = 2\int_0^t X_s\, dX_s + \langle X, X\rangle_t \ .$$

Proof. The general case follows from the special case by polarization: use the special case for $X \pm Y$ as well as X and Y.
The special case is proved by discretisation: let $\Delta = \{t_0, t_1, \dots, t_n\}$ be a finite discretisation of $[0, t]$. Then

$$\sum_{i=1}^n (X_{t_{i+1}} - X_{t_i})^2 = X_t^2 - X_0^2 - 2\sum_{i=1}^n X_{t_i}(X_{t_{i+1}} - X_{t_i}) \ .$$

Letting $|\Delta|$ going to zero, we get the claim. □

Theorem 4.18.5 (Ito formula for martingales). Given vector martingales $M = (M^{(1)}, \dots, M^{(d)})$ and X and a function $f \in C^2(\mathbb{R}^d, \mathbb{R})$. Then

$$f(X_t)-f(X_0) = \int_0^t \nabla f(X)\, dM_t + \frac{1}{2} \sum_{ij} \int_0^t \delta_{x_i}\delta_{x_j} f_{x_i x_j}(X_s)\, d\langle M_t^{(i)}, M_t^{(j)}\rangle \, .$$

Proof. It is enough to prove the formula for polynomials. By the integration by parts formula, we get the result for functions $f(x) = x_i g(x)$, if it is established for a function g. Since it is true for constant functions, we are done by induction. □

Remark. The usual Ito formula in one dimensions is a special case

$$f(X_t) - f(X_0) = \int_0^t f'(X_s)\, dB_s + \frac{1}{2} \int_0^t f''(X_s)\, ds \, .$$

In one dimension and if $M_t = B_t$ is Brownian motion and X_t is a martingale, we have We will use it later, when dealing with stochastic differential equations. It is a special case, because $\langle B_t, B_t \rangle = t$, so that $d\langle B_t, B_t \rangle = dt$.

Example. If $f(x) = x^2$, this formula gives for processes satisfying $X_0 = 0$

$$X_t^2/2 = \int_0^t X_s\, dB_s + \frac{1}{2} t \, .$$

This formula integrates the stochastic integral $\int_0^t X_s\, dB_s = X_t^2/2 - t/2$.

Example. If $f(x) = \log(x)$, the formula gives

$$\log(X_t/X_0) = \int_0^t dB_s/X_s - \frac{1}{2} \int_0^t ds/X_s^2 \, .$$

4.19 Stochastic differential equations

We have seen earlier that if B_t is Brownian motion, then $X = f(B,t) = e^{\alpha B_t - \alpha^2 t/2}$ is a martingale. In the last section we learned using Ito's formula and and $\frac{1}{2}\Delta f + \dot{f} = 0$ that

$$\int_0^t \alpha X_s\, dM_s = X_t - 1 \, .$$

We can write this in differential form as

$$dX_t = \alpha X_t\, dM_t, X_0 = 1 \, .$$

This is an example of a **stochastic differential equation** (SDE) and one would use the notation

$$\frac{dX}{dM} = \alpha X$$

if it would not lead to confusion with the corresponding ordinary differential equation, where M is not a stochastic process but a variable and where the solution would be $X = e^{\alpha B}$. Here, the solution is the stochastic process $X_t = e^{\alpha B_t - \alpha^2 t/2}$.

Definition. Let B_t be Brownian motion in $\mathbb{R}^d$. A **solution of a stochastic differential equation**

$$dX_t = f(X_t, B_t) \cdot dB_t + g(X_t)\, dt ,$$

is a $\mathbb{R}^d$-valued process X_t satisfying

$$X_t = \int_0^t f(X_s, B_s) \cdot dB_s + \int_0^t g(X_s)\, ds ,$$

where $f : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ and $g : \mathbb{R}^d \times \mathbb{R}^+ \to \mathbb{R}^d$.

As for ordinary differential equations, where one can easily solve separable differential equations $dx/dt = f(x) + g(t)$ by integration, this works for stochastic differential equations. However, to integrate, one has to use an adapted substitution. The key is **Ito's formula** (4.18.5) which holds for martingales and so for solutions of stochastic differential equations which is in one dimensions

$$f(X_t) - f(X_0) = \int_0^t f'(X_s)\, dX_s + \frac{1}{2} \int_0^t f''(X_s)\, d\langle X_s, X_s \rangle .$$

The following "multiplication table" for the product $\langle \cdot, \cdot \rangle$ and the differentials dt, dB_t can be found in many books of stochastic differential equations [2, 46, 66] and is useful to have in mind when solving actual stochastic differential equations:

	dt	dB_t
dt	0	0
dB_t	0	t

Example. The linear ordinary differential equation $dX/dt = rX$ with solution $X_t = e^{rt} X_0$ has a stochastic analog. It is called the **stochastic population model**. We look for a stochastic process X_t which solves the SDE

$$\frac{dX_t}{dt} = rX_t + \alpha X_t \zeta_t .$$

Separation of variables gives

$$\frac{dX}{X} = rtdt + \alpha \zeta dt$$

and integration with respect to t

$$\int_0^t \frac{dX_t}{X_t} = rt + \alpha B_t \; .$$

In order to compute the stochastic integral on the left hand side, we have to do a change of variables with $f(X) = \log(x)$. Looking up the multiplication table:

$$\langle dX_t, dX_t \rangle = \langle rX_t dt + \alpha X_t dB_t, rX_t dt + \alpha^2 X_t dB_t \rangle = \alpha^2 X_t^2 dt \; .$$

Ito's formula in one dimensions

$$f(X_t) - f(X_0) = \int_0^t f'(X_s) \, dX_s + \frac{1}{2} \int_0^t f''(X_s) \langle X_s, X_s \rangle$$

gives therefore

$$\log(X_t/X_0) = \int_0^t dX_s/X_s - \frac{1}{2} \int_0^t \alpha^2 ds$$

so that $\int_0^t dX_s/X_s = \alpha^2 t/2 + \log(X_t/X_0)$. Therefore,

$$\alpha^2 t/2 + \log(X_t/X_0) = rt + \alpha B_t$$

and so $X_t = X_0 e^{rt - \alpha^2 t/2 + \alpha B_t}$. This process is called **geometric Brownian motion**. We see especially that $\dot{X} = X/2 + X\xi$ has the solution $X_t = e^{B_t}$.

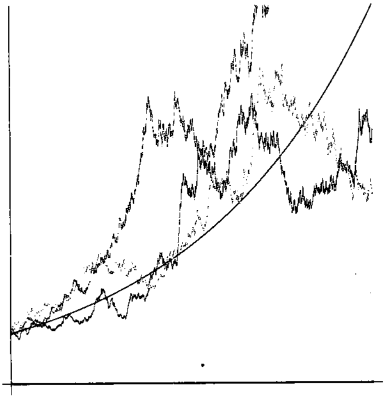


Figure. *Solutions to the stochastic population model for $r > 0$.*

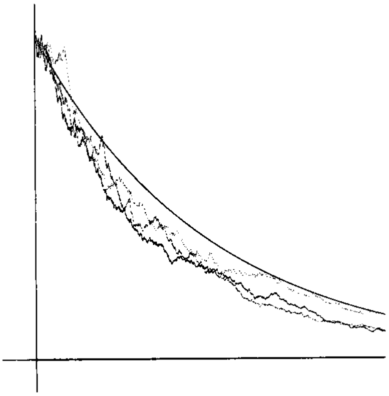


Figure. *Solutions to the stochastic population model for $r < 0$.*

Remark. The stochastic population model is also important when modeling financial markets. In that area the constant r is called the **percentage drift** or **expected gain** and α is called the **percentage volatility**. The **Black-Scholes model** makes the assumption that the stock prices evolves according to geometric Brownian motion.

Example. In principle, one can study stochastic versions of any differential equation. An example from physics is when a particle move in a possibly time-dependent force field $F(x,t)$ with **friction** b for which the equation without noise is

$$\ddot{x} = -b\dot{x} + F(x,t) .$$

If we add white noise, we get a stochastic differential equation

$$\ddot{x} = -b\dot{x} + F(x,t) + \alpha\zeta(t) .$$

For example, with $X = \dot{x}$ and $F = 0$, the function $v(t)$ satisfies the stochastic differential equation

$$\frac{dX_t}{dt} = -bX_t + \alpha\zeta_t ,$$

which has the solution

$$X_t = e^{-bt} + \alpha B_t .$$

With a time dependent force $F(x,t)$, already the differential equation without noise can not be given closed solutions in general. If the friction constant b is noisy, we obtain

$$\frac{dX_t}{dt} = (-b + \alpha\zeta_t)X_t$$

which is the stochastic population model treated in the previous example.

Example. Here is a list of stochastic differential equations with solutions. We again use the notation of **white noise** $\zeta(t) = \frac{dB}{dt}$ which is a generalized function in the following table. The notational replacement $dB_t = \zeta_t dt$ is quite popular for more applied sciences like **engineering** or **finance**.

Stochastic differential equation	Solution
$\frac{d}{dt}X_t = 1\zeta(t)$	$X_t = B_t$
$\frac{d}{dt}X_t = B_t\zeta(t)$	$X_t =: B_t^2 : /2 = (B_t^2 - 1)/2$
$\frac{d}{dt}X_t = B_t^2\zeta(t)$	$X_t =: B_t^3 : /3 = (B_t^3 - 3B_t)/3$
$\frac{d}{dt}X_t = B_t^3\zeta(t)$	$X_t =: B_t^4 : /4 = (B_t^4 - 6B_t^2 + 3)/4$
$\frac{d}{dt}X_t = B_t^4\zeta(t)$	$X_t =: B_t^5 : /5 = (B_t^5 - 10B_t^3 + 15B_t)/5$
$\frac{d}{dt}X_t = \alpha X_t\zeta(t)$	$X_t = e^{\alpha B_t - \alpha^2 t/2}$
$\frac{d}{dt}X_t = rX_t + \alpha X_t\zeta(t)$	$X_t = e^{rt + \alpha B_t - \alpha^2 t/2}$

Remark. Because the Ito integral can be defined for any continuous martingale, Brownian motion could be replaced by an other continuous martingale M leading to other classes of stochastic differential equations. A solution must then satisfy

$$X_t = \int_0^t f(X_s, M_s, s) \cdot dM_s + \int_0^t g(X_s, s)\, ds .$$

Example.

$$X_t = e^{\alpha M_t - \alpha^2 \langle X,X \rangle_t/2}$$

is a solution of $dX_t = \alpha M_t dM_t$, $M_0 = 1$.

Remark. Stochastic differential equations were introduced by Ito in 1951. Differential equations with a different integral came from Stratonovich but there are formulas which relating them with each other. So, it is enough to consider the Ito integral. Both versions of stochastic integration have advantages and disadvantages. Kunita shows in his book [55] that one can view solutions as stochastic flows of diffeomorphisms. This brings the topic into the framework of ergodic theory.

For ordinary differential equations $\dot{x} = f(x,t)$, one knows that unique solutions exist locally if f is Lipshitz continuous in x and continuous in t. The proof given for 1-dimensional systems generalizes to differential equations in arbitrary Banach spaces. The idea of the proof is a Picard iteration of an operator which is a contraction. Below, we give a detailed proof of this existence theorem for ordinary differential equations. For stochastic differential equations, one can do the same. We will do such an iteration on the Hilbert space $H^2_{[0,t]}$ of $\mathcal{L}^2$ martingales X having finite norm

$$||X||_T = \mathrm{E}[\sup_{t \leq T} X_t^2] .$$

We will need the following version of Doob's inequality:

Lemma 4.19.1. Let X be a $\mathcal{L}^p$ martingale with $p \geq 1$. Then

$$\mathrm{E}[\sup_{s \leq t} |X_s|^p] \leq (\frac{p}{p-1})^p \cdot \mathrm{E}[|X_t|^p] .$$

Proof. We can assume without loss of generality that X is bounded. The general result follows by approximating X by $X \wedge k$ with $k \to \infty$. Define $X^* = \sup_{s \leq t} |X_s|^p$. From Doob's inequality

$$\mathrm{P}[X \geq \lambda] \leq \mathrm{E}[|X_t| \cdot 1_{X^* \geq \lambda}]$$

we get

$$\begin{aligned}
\mathrm{E}[|X^*|^p] &= \mathrm{E}[\int_0^{X^*} p\lambda^{p-1} \, d\lambda] \\
&= \mathrm{E}[\int_0^{\infty} p\lambda^{p-1} 1_{\{X^* \geq \lambda\}} \, d\lambda] \\
&= \mathrm{E}[\int_0^{\infty} p\lambda^{p-1} \mathrm{P}[X^* \geq \lambda] \, d\lambda] \\
&\leq \mathrm{E}[\int_0^{\infty} p\lambda^{p-1} \mathrm{E}[|X_t| \cdot 1_{X^* \geq \lambda}] \, d\lambda] \\
&= p\mathrm{E}[|X_t| \int_0^{X^*} \lambda^{p-2} \, d\lambda \\
&= \frac{p}{p-1} \mathrm{E}[|X_t| \cdot (X^*)^{p-1}] .
\end{aligned}$$

Hölder's inequality gives

$$\mathrm{E}[|X^*|^p] \leq \frac{p}{p-1}\mathrm{E}[(X^*)^p]^{(p-1)/p}\mathrm{E}[|X_t|^p]^{1/p}$$

and the claim follows. □

Theorem 4.19.2 (Local existence and uniqueness of solutions). Let M be a continuous martingale. Assume $f(x,t)$ and $g(x,t)$ are continuous in t and Lipshitz continuous in x. Then there exists $T > 0$ and a unique solution X_t of the SDE

$$dX = f(x,t)\, dM + g(x,t)\, ds$$

with initial condition $X_0 = X_0$.

Proof. Define the operator

$$\mathcal{S}(X) = \int_0^t f(s, X_s)\, dM_s + \int_0^t g(s, X_s)\, ds$$

on $\mathcal{L}^2$-processes. Write $\mathcal{S}(X) = \mathcal{S}_1(X) + \mathcal{S}_2(X)$. We will show that on some time interval $(0,T]$, the map $\mathcal{S}$ is a contraction and that $\mathcal{S}^n(X)$ converges in the metric $|||X - Y|||_T = \mathrm{E}[\sup_{s \leq T}(X_s - Y_s)^2]$, if T is small enough to a unique fixed point. It is enough that for $i = 1, 2$

$$|||\mathcal{S}_i(X) - \mathcal{S}_i(Y)|||_T \leq (1/4) \cdot ||X - Y||_T$$

then $\mathcal{S}$ is a contraction

$$|||\mathcal{S}(X) - \mathcal{S}(Y)|||_T \leq (1/2) \cdot ||X - Y||_T \, .$$

By assumption, there exists a constant K, such that

$$|f(t,w) - f(t,w')| \leq K \cdot \sup_{s \leq 1} |w - w'| \, .$$

(i) $|||\mathcal{S}_1(X) - \mathcal{S}_1(Y)|||_T = ||| \int_0^t f(s, X_s) - f(s, Y_s)\, dM_s |||_T \leq (1/4) \cdot |||X - Y|||_T$ for T small enough.

Proof. By the above lemma for $p = 2$, we have

$$\begin{aligned}
|||\mathcal{S}_1(X) - \mathcal{S}_1(Y)|||_T &= \mathrm{E}[\sup_{t \leq T}(\int_0^t f(s,X) - f(s,Y)\, dM_s)^2] \\
&\leq 4\mathrm{E}[(\int_0^T f(t,X) - f(t,Y)\, dM_t)^2] \\
&= 4\mathrm{E}[(\int_0^T f(t,X) - f(t,Y))^2\, d\langle M, M\rangle_t] \\
&\leq 4K^2\mathrm{E}[\int_0^T \sup_{s \leq t} |X_s - Y_s|^2\, dt] \\
&= 4K^2 \int_0^T |||X - Y|||_s\, ds \\
&\leq (1/4) \cdot |||X - Y|||_T \,,
\end{aligned}$$

where the last inequality holds for T small enough.

(ii) $|||\mathcal{S}_2(X) - \mathcal{S}_2(Y)|||_T = ||| \int_0^t g(s,X_s) - g(s,Y_s)\, ds|||_T \leq (1/4) \cdot |||X - Y|||_T$ for T small enough. This is proved for differential equations in Banach spaces.
The two estimates (i) and (ii) prove the claim in the same way as in the classical Cauchy-Picard existence theorem. □

Appendix. In this Appendix, we add the existence of solutions of ordinary differential equations in Banach spaces. Let $\mathcal{X}$ be a Banach space and I an interval in $\mathbb{R}$. The following lemma is useful for proving existence of fixed points of maps.

Lemma 4.19.3. Let $X = \overline{B_r(x_0)} \subset \mathcal{X}$ and assume ϕ is a differentiable map $\mathcal{X} \to \mathcal{X}$. If for all $x \in X$, $||D\phi(x)|| \leq |\lambda| < 1$ and

$$||\phi(x_0) - x_0|| \leq (1 - \lambda) \cdot r$$

then ϕ has exactly one fixed point in X.

Proof. The condition $||x - x_0|| < r$ implies that

$$||\phi(x) - x_0|| \leq ||\phi(x) - \phi(x_0)|| + ||\phi(x_0) - x_0|| \leq \lambda r + (1 - \lambda)r = r \,.$$

The map ϕ maps therefore the ball X into itself. Banach's fixed point theorem applied to the complete metric space X and the contraction ϕ implies the result. □

Let f be a map from $I \times \mathcal{X}$ to $\mathcal{X}$. A differentiable map $u : \mathcal{J} \to \mathcal{X}$ of an open ball $J \subset I$ in $\mathcal{X}$ is called a **solution of the differential equation**

$$\dot{x} = f(t,x)$$

if we have for all $t \in J$ the relation

$$\dot{u}(t) = f(t, u(t)) .$$

Theorem 4.19.4 (Cauchy-Picard Existence theorem). Let $f : I \times \mathcal{X} \to \mathcal{X}$ be continuous in the first coordinate and locally Lipshitz continuous in the second. Then, for every $(t_0, x_0) \in I \times \mathcal{X}$, there exists an open interval $J \subset I$ with midpoint t_0, such that on J, there exists exactly one solution of the differential equation $\dot{x} = f(t, x)$.

Proof. There exists an interval $J(t_0, a) = (t_0 - a, t_0 + a) \subset I$ and a ball $B(x_0, b)$, such that

$$M = \sup\{||f(t,x)|| \mid (t,x) \in J(t_0,a) \times B(x_0,b)\}$$

as well as

$$k = \sup\{\frac{||f(t,x_1) - f(t,x_2)||}{||x_1 - x_2||} \mid (t,x_1),(t,x_2) \in J(t_0,a) \times B(x_0,b), x_1 \neq x_2\}$$

are finite. Define for $r < a$ the Banach space

$$\mathcal{X}_r = C(\overline{J}(t_0,r), \mathcal{X}) = \{y : \overline{J}(t_0,r) \to \mathcal{X},\ y \text{ continuous}\}$$

with norm

$$||y|| = \sup_{t \in \overline{J}(t_0,r)} ||y(t)||$$

Let $V_{r,b}$ be the open ball in $\mathcal{X}_r$ with radius b around the constant map $t \mapsto x_0$. For every $y \in V_{r,b}$ we define

$$\phi(y) : t \mapsto x_0 + \int_{t_0}^{t} f(s, y(s)) ds$$

which is again an element in $\mathcal{X}_r$. We prove now, that for r small enough, ϕ is a contraction. A fixed point of ϕ is then a solution of the differential equation $\dot{x} = f(t,x)$, which exists on $J = J_r(t_0)$. For two points $y_1, y_2 \in V_r$, we have by assumption

$$||f(s, y_1(s)) - f(s, y_2(s))|| \leq k \cdot ||y_1(s) - y_2(s)|| \leq k \cdot ||y_1 - y_2||$$

for every $s \in \overline{J}_r$. Thus, we have

$$\begin{aligned} ||\phi(y_1) - \phi(y_2)|| &= ||\int_{t_0}^{t} f(s, y_1(s)) - f(s, y_2(s))\, ds|| \\ &\leq \int_{t_0}^{t} ||f(s, y_1(s)) - f(s, y_2(s))||\, ds \\ &\leq kr \cdot ||y_1 - y_2|| . \end{aligned}$$

On the other hand, we have for every $s \in \overline{J}_r$

$$||f(s, y(s))|| \leq M$$

and so

$$||\phi(x_0) - x_0|| = ||\int_{t_0}^{t} f(s, x_0(s))\, ds|| \leq \int_{t_0}^{t} ||f(s, x_0(s))||\, ds \leq M \cdot r\,.$$

We can apply the above lemma, if $kr < 1$ and $Mr < b(1 - kr)$. This is the case, if $r < b/(M + kb)$. By choosing r small enough, we can get the contraction rate as small as we wish. □

Definition. A set X with a distance function $d(x, y)$ for which the following properties
(i) $d(y, x) = d(x, y) \geq 0$ for all $x, y \in X$.
(ii) $d(x, x) = 0$ and $d(x, y) > 0$ for $x \neq y$.
(iii) $d(x, z) \leq d(x, y) + d(y, z)$ for all x, y, z. hold is called a **metric space**.

Example. The plane $\mathbb{R}^2$ with the usual distance $d(x, y) = |x - y|$. An other metric is the Manhattan or taxi metric $d(x, y) = |x_1 - y_1| + |x_2 - y_2|$.

Example. The set $C([0, 1])$ of all continuous functions $x(t)$ on the interval $[0, 1]$ with the distance $d(x, y) = \max_t |x(t) - y(t)|$ is a metric space.

Definition. A map $\phi : X \to X$ is called a **contraction**, if there exists $\lambda < 1$ such that $d(\phi(x), \phi(y)) \leq \lambda \cdot d(x, y)$ for all $x, y \in X$. The map ϕ shrinks the distance of any two points by the contraction factor λ.

Example. The map $\phi(x) = \frac{1}{2}x + (1, 0)$ is a contraction on R^2.

Example. The map $\phi(x)(t) = \sin(t)x(t) + t$ is a contraction on $C([0, 1])$ because $|\phi(x)(t) - \phi(y)(t)| = |\sin(t)| \cdot |x(t) - y(t)| \leq \sin(1) \cdot |x(t) - y(t)|$.

Definition. A **Cauchy sequence** in a metric space (X, d) is defined to be a sequence which has the property that for any $\epsilon > 0$, there exists n_0 such that $|x_n - x_m| \leq \epsilon$ for $n \geq n_0, m \geq n_0$.
A metric space in which every Cauchy sequence converges to a limit is called **complete**.

Example. The n-dimensional Euclidean space

$$(\mathbb{R}^n, d(x, y) = |x - y| = \sqrt{x_1^2 + \cdots + x_n^2})$$

is complete. The set of rational numbers with the usual distance

$$(\mathbb{Q}, d(x, y) = |x - y|)$$

is not complete.

Example. The space $C[0,1]$ is complete: given a Cauchy sequence x_n, then $x_n(t)$ is a Cauchy sequence in R for all t. Therefore $x_n(t)$ converges point wise to a function $x(t)$. This function is continuous: take $\epsilon > 0$, then $|x(t) - x(s)| \leq |x(t) - x_n(t)| + |x_n(t) - y_n(s)| + |y_n(s) - y(s)|$ by the triangle inequality. If s is close to t, the second term is smaller than $\epsilon/3$. For large n, $|x(t) - x_n(t)| \leq \epsilon/3$ and $|y_n(s) - y(s)| \leq \epsilon/3$. So, $|x(t) - x(s)| \leq \epsilon$ if $|t - s|$ is small.

Theorem 4.19.5 (Banachs fixed point theorem). A contraction ϕ in a complete metric space (X, d) has exactly one fixed point in X.

Proof. (i) We first show by induction that

$$d(\phi^n(x), \phi^n(y)) \leq \lambda^n \cdot d(x,y)$$

for all n.

(ii) Using the triangle inequality and $\sum_k \lambda^k = (1-\lambda)^{-1}$, we get for all $x \in X$,

$$d(x, \phi^n x) \leq \sum_{k=0}^{n-1} d(\phi^k x, \phi^{k+1} x) \leq \sum_{k=0}^{n-1} \lambda^k d(x, \phi(x)) \leq \frac{1}{1-\lambda} \cdot d(x, \phi(x)) \; .$$

(iii) For all $x \in X$ the sequence $x_n = \phi^n(x)$ is a Cauchy sequence because by (i),(ii),

$$d(x_n, x_{n+k}) \leq \lambda^n \cdot d(x, x_k) \leq \lambda^n \cdot \frac{1}{1-\lambda} \cdot d(x, x_1) \; .$$

By completeness of X it has a limit $\tilde{x}$ which is a fixed point of ϕ.

(iv) There is only one fixed point. Assume, there were two fixed points $\tilde{x}, \tilde{y}$ of ϕ. Then

$$d(\tilde{x}, \tilde{y}) = d(\phi(\tilde{x}), \phi(\tilde{y})) \leq \lambda \cdot d(\tilde{x}, \tilde{y}) \; .$$

This is impossible unless $\tilde{x} = \tilde{y}$. □

Chapter 5

Selected Topics

5.1 Percolation

Definition. Let e_i be the standard basis in the lattice $\mathbb{Z}^d$. Denote with $\mathbb{L}^d$ the Cayley graph of $\mathbb{Z}^d$ with the generators $A = \{e_1, \dots, e_d\}$. This graph $\mathbb{L}^d = (V, E)$ has the lattice $\mathbb{Z}^d$ as vertices. The edges or bonds in that graph are straight line segments connecting **neighboring points** x, y. Points satisfying $|x - y| = \sum_{i=1}^{d} |x_i - y_i| = 1$.

Definition. We declare each bond of $\mathbb{L}^d$ to be **open** with probability $p \in [0, 1]$ and **closed** otherwise. Bonds are open ore closed independently of all other bonds. The product measure P_p is defined on the probability space $\Omega = \prod_{e \in E}\{0, 1\}$ of all **configurations**. We denote expectation with respect to P_p with $\mathrm{E}_p[\cdot]$.

Definition. A **path** in $\mathbb{L}^d$ is a sequence of vertices $(x_0, x_1, \dots, x_n)$ such that $(x_i, x_{i+1}) = e_i$ are bonds of $\mathbb{L}^d$. Such a path has **length** n and **connects** x_0 with x_n. A path is called **open** if all its edges are open and **closed** if all its edges are closed. Two subgraphs of $\mathbb{L}^d$ are **disjoint** if they have no edges and no vertices in common.

Definition. Consider the **random subgraph** of $\mathbb{L}^d$ containing the vertex set $\mathbb{Z}^d$ and only open edges. The connected components of this graph are called **open clusters**. If it is finite, an open cluster is also called a **lattice animal**. Call $C(x)$ the open cluster containing the vertex x. By translation invariance, the distribution of $C(x)$ is independent of x and we can take $x = 0$ for which we write $C(0) = C$.

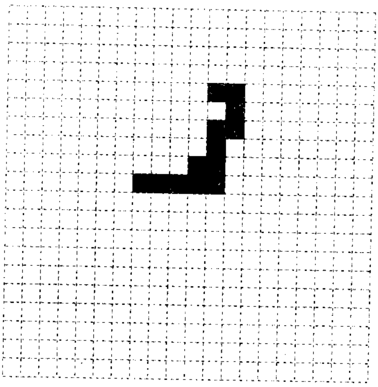


Figure. *A lattice animal.*

Definition. Define the **percolation probability** $\theta(p)$ being the probability that a given vertex belongs to an infinite open cluster.

$$\theta(p) = \mathrm{P}[|C| = \infty] = 1 - \sum_{n=1}^{\infty} \mathrm{P}[|C| = n] \, .$$

One of the goals of **bond percolation** theory is to study the function $\theta(p)$.

Lemma 5.1.1. There exists a critical value $p_c = p_c(d)$ such that $\theta(p) = 0$ for $p < p_c$ and $\theta(p) > 0$ for $p > p_c$. The value $d \mapsto p_c(d)$ is non-increasing with respect to the dimension $p_c(d+1) \leq p_c(d)$.

Proof. The function $p \mapsto \theta(p)$ is non-decreasing and $\theta(0) = 0, \theta(1) = 1$. We can therefore define

$$p_c = \inf\{p \in [0,1] \mid \theta(p) > 0\} .$$

The graph $\mathbb{Z}^d$ can be embedded into the graph $\mathbb{Z}^{d'}$ for $d < d'$ by realizing $\mathbb{Z}^d$ as a linear subspace of $\mathbb{Z}^{d'}$ parallel to a coordinate plane. Any configuration in $\mathbb{L}^{d'}$ projects then to a configuration in $\mathbb{L}^d$. If the origin is in an infinite cluster of $\mathbb{Z}^d$, then it is also in an infinite cluster of $\mathbb{Z}^{d'}$. □

Remark. The one-dimensional case $d = 1$ is not interesting because $p_c = 1$ there. Interesting phenomena are only possible in dimensions $d > 1$. The planar case $d = 2$ is already very interesting.

Definition. A **self-avoiding random walk** in $\mathbb{L}^d$ is the process S_T obtained by stopping the ordinary random walk S_n with stopping time

$$T(\omega) = \inf\{n \in \mathbb{N} \mid \omega(n) = \omega(m), m < n\} .$$

Let $\sigma(n)$ be the number of self-avoiding paths in $\mathbb{L}^d$ which have length n. The **connective constant** of $\mathbb{L}^d$ is defined as

$$\lambda(d) = \lim_{n \to \infty} \sigma(n)^{1/n} .$$

Remark. The exact value of $\lambda(d)$ is not known. But one has the elementary estimate $d < \lambda(d) < 2d - 1$ because a self-avoiding walk can not reverse direction and so $\sigma(n) \leq 2d(2d-1)^{n-1}$ and a walk going only forward in each direction is self-avoiding. For example, it is known that $\lambda(2) \in [2.62002, 2.69576]$ and numerical estimates makes one believe that the real value is 2.6381585. The number c_n of self-avoiding walks of length n in $\mathbb{L}^2$ is for small values $c_1 = 4, c_2 = 12, c_3 = 36, c_4 = 100, c_5 = 284, c_6 = 780, c_7 = 2172, \dots$. Consult [62] for more information on the self-avoiding random walk.

Theorem 5.1.2 (Broadbent-Hammersley theorem). If $d > 1$, then

$$0 < \lambda(d)^{-1} \leq p_c(d) \leq p_c(2) < 1 \, .$$

Proof. (i) $p_c(d) \geq \lambda(d)^{-1}$.
Let $N(n) \leq \sigma(n)$ be the number of open self-avoiding paths of length n in $\mathbb{L}^n$. Since any such path is open with probability p^n, we have

$$\mathrm{E}_p[N(n)] = p^n \sigma(n) \, .$$

If the origin is in an infinite open cluster, there must exist open paths of all lengths beginning at the origin so that

$$\theta(p) \leq \mathrm{P}_p[N(n) \geq 1] \leq \mathrm{E}_p[N(n)] = p^n \sigma(n) = (p\lambda(d) + o(1))^n$$

which goes to zero for $p < \lambda(p)^{-1}$. This shows that $p_c(d) \geq \lambda(d)^{-1}$.

(ii) $p_c(2) < 1$.
Denote by $\mathbb{L}^2_*$ the dual graph of $\mathbb{L}^2$ which has as vertices the faces of $\mathbb{L}^2$ and as vertices pairs of faces which are adjacent. We can realize the vertices as $\mathbb{Z}^2 + (1/2, 1/2)$. Since there is a bijective relation between the edges of $\mathbb{L}^2$ and $\mathbb{L}^2_*$ and we declare an edge of $\mathbb{L}^2_*$ to be open if it crosses an open edge in $\mathbb{L}^2$ and closed, if it crosses a closed edge. This defines bond percolation on $\mathbb{L}^2_*$.

The fact that the origin is in the interior of a closed circuit of the dual lattice if and only if the open cluster at the origin is finite follows from the **Jordan curve theorem** which assures that a closed path in the plane divides the plane into two disjoint subsets.

Let $\rho(n)$ denote the number of closed circuits in the dual which have length n and which contain in their interiors the origin of $\mathbb{L}^2$. Each such circuit contains a self-avoiding walk of length $n-1$ starting from a vertex of the form $(k+1/2, 1/2)$, where $0 \leq k < n$. Since the number of such paths γ is at most $n\sigma(n-1)$, we have

$$\rho(n) \leq n\sigma(n-1)$$

and with $q = 1 - p$

$$\sum_{\gamma} \mathrm{P}[\gamma \text{ is closed}] \leq \sum_{n=1}^{\infty} q^n n\sigma(n-1) = \sum_{n=1}^{\infty} qn(q\lambda(2) + o(1))^{n-1}$$

which is finite if $q\lambda(2) < 1$. Furthermore, this sum goes to zero if $q \to 0$ so that we can find $0 < \delta < 1$ such that for $p > \delta$

$$\sum_{\gamma} \mathrm{P}[\gamma \text{ is closed}] \leq 1/2.$$

We have therefore

$$\mathrm{P}[|C| = \infty] = \mathrm{P}[\text{no } \gamma \text{ is closed}] \geq 1 - \sum_{\gamma} \mathrm{P}[\gamma \text{ is closed}] \geq 1/2$$

so that $p_c(2) \leq \delta < 1$. □

Remark. We will see below that even $p_c(2) < 1 - \lambda(2)^{-1}$. It is however known that $p_c(2) = 1/2$.

Definition. The parameter set $p < p_c$ is called the **sub-critical phase**, the set $p > p_c$ is the **supercritical phase**.

Definition. For $p < p_c$, one is also interested in the **mean size** of the open cluster

$$\chi(p) = \mathrm{E}_p[|C|] \, .$$

For $p > p_c$, one would like to know the **mean size** of the finite clusters

$$\chi^f(p) = \mathrm{E}_p[|C| \mid |C| < \infty] \, .$$

It is known that $\chi(p) < \infty$ for $p < p_c$ but only conjectured that $\chi^f(p) < \infty$ for $p > p_c$.

An interesting question is whether there exists an open cluster at the critical point $p = p_c$. The answer is known to be no in the case $d = 2$ and generally believed to be no for $d \geq 3$. For p near p_c it is believed that the percolation probability $\theta(p)$ and the mean size $\chi(p)$ behave as powers of $|p - p_c|$. It is conjectured that the following critical exponents

$$\begin{aligned}
\gamma &= -\lim_{p \nearrow p_c} \frac{\log \chi(p)}{\log |p - p_c|} \\
\beta &= \lim_{p \searrow p_c} \frac{\log \theta(p)}{\log |p - p_c|} \\
\delta^{-1} &= -\lim_{n \to \infty} \frac{\log \mathrm{P}_{p_c}[|C| \geq n}{\log n}] \, .
\end{aligned}$$

exist.

Percolation deals with a **family** of probability spaces $(\Omega, \mathcal{A}, \mathrm{P}_p)$, where $\Omega = \{0,1\}^{\mathbb{L}^d}$ is the set of configurations with product σ-algebra $\mathcal{A}$ and product measure $\mathrm{P}_p = (p, 1-p)^{\mathbb{L}^d}$.

Definition. There exists a natural partial ordering in Ω coming from the ordering on $\{0,1\}$: we say $\omega \leq \omega'$, if $\omega(e) \leq \omega'(e)$ for all bonds $e \in \mathbb{L}^2$. We call a random variable X on $(\Omega, \mathcal{A}, \mathrm{P})$ **increasing** if $\omega \leq \omega'$ implies $X(\omega) \leq X(\omega')$. It is called **decreasing** if $-X$ is increasing. As usual, this notion can also be defined for measurable sets $A \in \mathcal{A}$: a set A is **increasing** if 1_A is increasing.

Lemma 5.1.3. If X is a increasing random variable in $\mathcal{L}^1(\Omega, P_q) \cap \mathcal{L}^1(\Omega, \mathrm{P}_p)$, then

$$\mathrm{E}_p[X] \leq \mathrm{E}_q[X]$$

if $p \leq q$.

Proof. If X depends only on a single bond e, we can write $\mathrm{E}_p[X] = pX(1) + (1-p)X(0)$. Because X is assumed to be increasing, we have $\frac{d}{dp}\mathrm{E}_p[X] = X(1) - X(0) \geq 0$ which gives $\mathrm{E}_p[X] \leq \mathrm{E}_q[X]$ for $p \leq q$. If X depends only on finitely many bonds, we can write it as a sum $X = \sum_{i=1}^d X_i$ of variables X_i which depend only on one bond and get again

$$\frac{d}{dp}\mathrm{E}_p[X] = \sum_{i=1}^n (X_i(1) - X_i(0)) \geq 0 .$$

In general we approximate every random variable in $\mathcal{L}^1(\Omega, \mathrm{P}_p) \cap \mathcal{L}^1(\Omega, P_q)$ by step functions which depend only on finitely many coordinates X_i. Since $\mathrm{E}_p[X_i] \to \mathrm{E}_p[X]$ and $\mathrm{E}_q[X_i] \to \mathrm{E}_q[X]$, the claim follows. □

The following **correlation inequality** is named after Fortuin, Kasterleyn and Ginibre (1971).

Theorem 5.1.4 (FKG inequality). For increasing random variables $X, Y \in \mathcal{L}^2(\Omega, \mathrm{P}_p)$, we have

$$\mathrm{E}_p[XY] \geq \mathrm{E}_p[X] \cdot \mathrm{E}_p[Y] .$$

Proof. As in the proof of the above lemma, we prove the claim first for random variables X which depend only on n edges $e_1, e_2, \dots, e_n$ and proceed by induction.

(i) The claim, if X and Y only depend on one edge e.
We have

$$(X(\omega) - X(\omega')(Y(\omega) - Y(\omega')) \geq 0$$

since the left hand side is 0 if $\omega(e) = \omega'(e)$ and if $1 = \omega(e) = \omega'(e) = 0$, both factors are nonnegative since X, Y are increasing, if $0 = \omega(e) = \omega'(e) = 1$ both factors are non-positive since X, Y are increasing. Therefore

$$\begin{aligned} 0 &\leq \sum_{\sigma,\sigma' \in \{0,1\}} (X(\omega) - X(\omega'))(Y(\omega) - Y(\omega'))\mathrm{P}_p[\omega(e) = \sigma]\mathrm{P}_p[\omega(e) = \sigma'] \\ &= 2(\mathrm{E}_p[XY] - \mathrm{E}_p[X]\mathrm{E}_p[Y]) . \end{aligned}$$

(ii) Assume the claim is known for all functions which depend on k edges with $k < n$. We claim that it holds also for X, Y depending on n edges $e_1, e_2, \dots, e_n$.
Let $\mathcal{A}_k = \mathcal{A}(e_1, \dots e_k)$ be the σ-algebra generated by functions depending only on the edges e_k. The random variables

$$X_k = \mathrm{E}_p[X|\mathcal{A}_k], Y_k = \mathrm{E}_p[Y|\mathcal{A}_k]$$

depend only on the $e_1, \dots, e_k$ and are increasing. By induction,

$$\mathrm{E}_p[X_{n-1}Y_{n-1}] \geq \mathrm{E}_p[X_{n-1}]\mathrm{E}_p[Y_{n-1}] .$$

By the tower property of conditional expectation, the right hand side is $\mathrm{E}_p[X]\mathrm{E}_p[Y]$. For fixed $e_1, \dots, e_{n-1}$, we have $(XY)_{n-1} \geq X_{n-1}Y_{n-1}$ and so

$$\mathrm{E}_p[XY] = \mathrm{E}_p[(XY)_{n-1}] \geq \mathrm{E}_p[X_{n-1}Y_{n-1}] .$$

(iii) Let X, Y be arbitrary and define $X_n = \mathrm{E}_p[X|\mathcal{A}_n]$, $Y_n = \mathrm{E}_p[Y|\mathcal{A}_n]$. We know from *(ii)* that $\mathrm{E}_p[X_nY_n] \geq \mathrm{E}_p[X_n]\mathrm{E}_p[Y_n]$. Since $X_n = \mathrm{E}[X|\mathcal{A}_n]$ and $Y_n = \mathrm{E}[X|\mathcal{A}_n]$ are martingales which are bounded in $\mathcal{L}^2(\Omega, \mathrm{P}_p)$, Doob's convergence theorem (3.5.4) implies that $X_n \to X$ and $Y_n \to Y$ in $\mathcal{L}^2$ and therefore $\mathrm{E}[X_n] \to \mathrm{E}[X]$ and $\mathrm{E}[Y_n] \to \mathrm{E}[Y]$. By the Schwarz inequality, we get also in $\mathcal{L}^1$ or the $\mathcal{L}^2$ norm in $(\Omega, \mathcal{A}, \mathrm{P}_p)$

$$\begin{aligned} ||X_nY_n - XY||_1 &\leq ||(X_n - X)Y_n||_1 + ||X(Y_n - Y)||_1 \\ &\leq ||X_n - X||_2||Y_n||_2 + ||X||_2||Y_n - Y||_2 \\ &\leq C(||X_n - X||_2 + ||Y_n - Y||_2) \to 0 \end{aligned}$$

where $C = \max(||X||_2, ||Y||_2)$ is a constant. This means $\mathrm{E}_p[X_nY_n] \to \mathrm{E}_p[XY]$. □

Remark. It follows immediately that if A, B are increasing events in Ω, then $\mathrm{P}_p[A \cap B] \geq \mathrm{P}_p[A] \cdot \mathrm{P}_p[B]$.

Example. Let Γ_i be families of paths in $\mathbb{L}^d_*$ and let A_i be the event that some path in Γ_i is open. Then A_i are increasing events and so after applying the inequality k times, we get

$$\mathrm{P}_p[\bigcap_{i=1}^k A_i] \geq \prod_{i=1}^k \mathrm{P}_p[A_i] .$$

We show now, how this inequality can be used to give an explicit bound for the critical percolation probability p_c in $\mathbb{L}^2$. The following corollary belongs still to the theorem of Broadbent-Hammersley.

Corollary 5.1.5.

$$p_c(2) \leq (1 - \lambda(2)^{-1}) .$$

Proof. Given any integer $N \in \mathbb{N}$, define the events

$$\begin{aligned} F_N &= \{\exists \text{ no closed path of length} \leq N \text{ in } \mathbb{L}^d_*\} \\ G_N &= \{\exists \text{ no closed path of length} > N \text{ in } \mathbb{L}^d_*\} . \end{aligned}$$

We know that $F_N \cap G_N \subset \{|C| = \infty\}$. Since F_N and G_N are both increasing, the correlation inequality says $\mathrm{P}_p[F_N \cap G_N] \geq \mathrm{P}_p[F_N] \cdot \mathrm{P}_p[G_N]$. We deduce

$$\theta(p) = \mathrm{P}_p[|C| = \infty] = \mathrm{P}_p[F_N \cap G_N] \geq \mathrm{P}_p[F_N] \cdot \mathrm{P}_p[G_N] .$$

If $(1-p)\lambda(2) < 1$, then we know that

$$\mathrm{P}_p[G_N^c] \leq \sum_{n=N}^{\infty} (1-p)^n n\sigma(n-1)$$

which goes to zero for $N \to \infty$. For N large enough, we have therefore $\mathrm{P}_p[G_N] \geq 1/2$. Since also $\mathrm{P}_p[F_N] > 0$, it follows that $\theta_p > 0$, if $(1-p)\lambda(2) < 1$ or $p < (1 - \lambda(2)^{-1})$ which proves the claim. □

Definition. Given $A \in \mathcal{A}$ and $\omega \in \Omega$. We say that an edge $e \in \mathbb{L}^d$ is **pivotal** for the pair (A, ω) if $1_A(\omega) \neq 1_A(\omega_e)$, where ω_e is the unique configuration which agrees with ω except at the edge e.

Theorem 5.1.6 (Russo's formula). Let A be an increasing event depending only on finitely many edges of $\mathbb{L}^d$. Then

$$\frac{d}{dp} \mathrm{P}_p[A] = \mathrm{E}_p[N(A)] ,$$

where $N(A)$ is the number of edges which are pivotal for A.

Proof. (i) We define a new probability space.
The family of probability spaces $(\Omega, \mathcal{A}, \mathrm{P}_p)$, can be embedded in one probability space

$$([0,1]^{\mathbb{L}^d}, \mathcal{B}([0,1]^{\mathbb{L}^d}), \mathrm{P}) ,$$

where P is the product measure $dx^{\mathbb{L}^d}$. Given a configuration $\eta \in [0,1]^{\mathbb{L}^d}$ and $p \in [0,1]$, we get a configuration in Ω by defining $\eta_p(e) = 1$ if $\eta(e) < p$ and $\eta_p = 0$ else. More generally, given $\mathbf{p} \in [0,1]^{\mathbb{L}^d}$, we get configurations $\eta_{\mathbf{p}}(e) = 1$ if $\eta(e) < \mathbf{p}(e)$ and $\eta_{\mathbf{p}} = 0$ else. Like this, we can define configurations with a large class of probability measures $\mathrm{P}_{\mathbf{p}} = \prod_{e \in \mathbb{L}^d}(p(e), 1-p(e))$ with **one** probability space and we have

$$\mathrm{P}_{\mathbf{p}}[A] = \mathrm{P}[\eta_{\mathbf{p}} \in A] .$$

(ii) Derivative with respect to one $p(f)$.
Assume $\mathbf{p}$ and $\mathbf{p}'$ differ only at an edge f such that $p(f) \leq p'(f)$. Then $\{\eta_{\mathbf{p}} \in A\} \subset \{\eta_{\mathbf{p}'} \in A\}$ so that

$$\begin{aligned} \mathrm{P}_{\mathbf{p}'}[A] - \mathrm{P}_{\mathbf{p}}[A] &= \mathrm{P}[\eta_{\mathbf{p}'} \in A] - \mathrm{P}[\eta_{\mathbf{p}} \in A] \\ &= \mathrm{P}[\eta_{\mathbf{p}'} \in A; \eta_{\mathbf{p}} \notin A] \\ &= (p'(f) - p(f))\mathrm{P}_p[f \text{ pivotal for } A] . \end{aligned}$$

Divide both sides by $(p'(f) - p(f))$ and let $p'(f) \to p(f)$. This gives

$$\frac{\partial}{\partial p(f)} P_{\mathbf{p}}[A] = P_{\mathbf{p}}[f \text{ pivotal for } A] .$$

(iii) The claim, if A depends on finitely many edges. If A depends on finitely many edges, then $P_{\mathbf{p}}[A]$ is a function of a finite set $\{p(f_i)\}_{i=1}^m$ of edge probabilities. The chain rule gives then

$$\begin{aligned} \frac{d}{dp} P_p[A] &= \sum_{i=1}^m \frac{\partial}{\partial p(f_i)} \mathrm{P}_{\mathbf{p}}[A]|_{\mathbf{p}=(p,p,p,\dots,p)} \\ &= \sum_{i=1}^m \mathrm{P}_{\mathbf{p}}[f_i \text{ pivotal for } A] \\ &= \mathrm{E}_p[N(A)] . \end{aligned}$$

(iv) The general claim.
In general, define for every finite set $F \subset E$

$$\mathbf{p}_F(e) = p + 1_{\{e \in F\}}\delta$$

where $0 \leq p \leq p + \delta \leq 1$. Since A is increasing, we have

$$P_{p+\delta}[A] \geq P_{\mathbf{p}_F}[A]$$

and therefore

$$\frac{1}{\delta}(P_{p+\delta}[A] - P_p[A]) \geq \frac{1}{\delta}(P_{\mathbf{p}_F}[A] - \mathrm{P}_p[A]) \to \sum_{e \in F} P_p[\mathrm{e} \text{ pivotal for } A]$$

as $\delta \to 0$. The claim is obtained by making F larger and larger filling out E. □

Example. Let $F = \{e_1, e_2, \dots, e_m\} \subset E$ be a finite set in of edges.

$$A = \{\text{the number of open edges in } F \text{ is} \geq k\}\ .$$

An edge $e \in F$ is pivotal for A if and only if $A \setminus \{e\}$ has exactly $k-1$ open edges. We have

$$\mathrm{P}_p[e \text{ is pivotal}] = \left(\begin{array}{c} m-1 \\ k-1 \end{array} \right) p^{k-1}(1-p)^{m-k}$$

so that by Russo's formula

$$\frac{d}{dp}\mathrm{P}_p[A] = \sum_{e \in F} \mathrm{P}_p[e \text{ is pivotal}] = m \left(\begin{array}{c} m-1 \\ k-1 \end{array} \right) p^{k-1}(1-p)^{m-k}\ .$$

Since we know $\mathrm{P}_0[A] = 0$, we obtain by integration

$$\mathrm{P}_p[A] = \sum_{l=k}^{m} \left(\begin{array}{c} m \\ l \end{array} \right) p^l (1-p)^{m-1}\ .$$

Remark. If A does no more depend on finitely many edges, then $\mathrm{P}_p[A]$ need no more be differentiable for all values of p.

Definition. The **mean size of the open cluster** is $\chi(p) = \mathrm{E}_p[|C|]$.

Theorem 5.1.7 (Uniqueness). For $p < p_c$, the mean size of the open cluster is finite $\chi(p) < \infty$.

The proof of this theorem is quite involved and we will not give the full argument. Let $S(n,x) = \{y \in \mathbb{Z}^d \mid |x-y| = \sum_{i=1}^d |x_i| \leq n\}$ be the ball of radius n around x in $\mathbb{Z}^d$ and let A_n be the event that there exists an open path joining the origin with some vertex in $\delta S(n,0)$.

Lemma 5.1.8. (Exponential decay of radius of the open cluster) If $p < p_c$, there exists ψ_p such that $\mathrm{P}_p[A_n] < e^{-n\psi_p}$.

Proof. Clearly, $|S(n,0)| \leq C_d \cdot (n+1)^d$ with some constant C_d. Let $M = \max\{n \mid A_n \text{ occurs }\}$. By definition of p_c, if $p < p_c$, then $\mathrm{P}_p[M < \infty] = 1$. We get

$$\begin{aligned}
\mathrm{E}_p[|C|] &\leq \sum_n \mathrm{E}_p[|C| \mid M = n] \cdot \mathrm{P}_p[M = n] \\
&\leq \sum_n |S(n,0)| \mathrm{P}_p[A_n] \\
&\leq \sum_n C_d (n+1)^d e^{-n\psi_p} < \infty\ .
\end{aligned}$$

□

Proof. We are concerned with the probabilities $g_p(n) = \mathrm{P}_p[A_n]$. Sine A_n are increasing events, Russo's formula gives

$$g_p'(n) = \mathrm{E}_p[N(A_n)]\,,$$

where $N(A_n)$ is the number of pivotal edges in A_n. We have

$$\begin{aligned}
g_p'(n) &= \sum_e \mathrm{P}_p[e \text{ pivotal for } A] \\
&= \sum_e \frac{1}{p}\mathrm{P}_p[e \text{ open and pivotal for } A] \\
&= \sum_e \frac{1}{p}\mathrm{P}_p[A \cap \{e \text{ pivotal for } A\}] \\
&= \sum_e \frac{1}{p}\mathrm{P}_p[A \cap \{e \text{ pivotal for } A\}|A] \cdot \mathrm{P}_p[A] \\
&= \sum_e \frac{1}{p}\mathrm{E}_p[N(A) \mid A] \cdot \mathrm{P}_p[A] \\
&= \sum_e \frac{1}{p}\mathrm{E}_p[N(A) \mid A] \cdot g_p(n)
\end{aligned}$$

so that

$$\frac{g_p'(n)}{g_p(n)} = \frac{1}{p}\mathrm{E}_p[N(A_n) \mid A_n]\,.$$

By integrating up from α to β, we get

$$\begin{aligned}
g_\alpha(n) = g_\beta(n)\exp(-\int_\alpha^\beta \frac{1}{p}\mathrm{E}_p[N(A_n) \mid A_n]\, dp) \\
\leq g_\beta(n)\exp(-\int_\alpha^\beta \mathrm{E}_p[N(A_n) \mid A_n]\, dp) \\
\leq \exp(-\int_\alpha^\beta \mathrm{E}_p[N(A_n) \mid A_n]\, dp)\,.
\end{aligned}$$

One needs to show then that $\mathrm{E}_p[N(A_n)\ |A_n]$ grows roughly linearly when $p < p_c$. This is quite technical and we skip it. □

Definition. The **number of open clusters per vertex** is defined as

$$\kappa(p) = \mathrm{E}_p[|C|^{-1}] = \sum_{n=1}^{\infty} \frac{1}{n}\mathrm{P}_p[|C| = n]\,.$$

Let B_n the box with side length $2n$ and center at the origin and let K_n be the number of open clusters in B_n. The following proposition explains the name of κ.

Proposition 5.1.9. In $\mathcal{L}^1(\Omega, \mathcal{A}, \mathrm{P}_p)$ we have

$$K_n/|B_n| \to \kappa(p) .$$

Proof. Let $C_n(x)$ be the connected component of the open cluster in B_n which contains $x \in \mathbb{Z}^d$. Define $\Gamma(x) = |C(x)|^{-1}$.

(i) $\sum_{x\in B_n} \Gamma_n(x) = K_n$.
Proof. If Σ is an open cluster of B_n, then each vertex $x \in \Sigma$ contributes $|\Sigma|^{-1}$ to the left hand side. Thus, each open cluster contributes 1 to the left hand side.

(ii) $\frac{K_n}{|B_n|} \geq \frac{1}{|B_n|} \sum_{x\in B_n} \Gamma(x)$ where $\Gamma(x) = |C(x)|^{-1}$.
Proof. Follows from (i) and the trivial fact $\Gamma(x) \leq \Gamma_n(x)$.

(iii) $\frac{1}{|B_n|} \sum_{x\in B_n} \Gamma(x) \to \mathrm{E}_p[\Gamma(0)] = \kappa(p)$.
Proof. $\Gamma(x)$ are bounded random variables which have a distribution which is invariant under the ergodic group of translations in $\mathbb{Z}^d$. The claim follows from the ergodic theorem.

(iv) $\liminf_{n\to\infty} \frac{K_n}{|B_n|} \geq \kappa(p)$ almost everywhere.
Proof. Follows from (ii) and (iii).

(v) $\sum_{x\in B(n)} \Gamma_n(x) \leq \sum_{x\in B(n)} \Gamma(x) + \sum_{x\sim \delta B_n} \Gamma_n(x)$, where $x \sim Y$ means that x is in the same cluster as one of the elements $y \in Y \subset \mathbb{Z}^d$.

(vi) $\frac{1}{|B_n|} \sum_{x\in B_n} \Gamma_n(x) \leq \frac{1}{|B_n|} \sum_{x\in B_n} \Gamma(x) + \frac{|\delta B_n|}{|B_n|}$. □

Remark. It is known that function $\kappa(p)$ is continuously differentiable on $[0, 1]$. It is even known that κ and the mean size of the open cluster $\chi(p)$ are real analytic functions on the interval $[0, p_c)$. There would be much more to say in percolation theory. We mention:
The uniqueness of the infinite open cluster:
For $p > p_c$ and if $\theta(p_c) > 0$ also for $p = p_c$, there exists a unique infinite open cluster.
Regularity of some functions $\theta(p)$
For $p > p_c$, the functions $\theta(p), \chi^f(p), \kappa(p)$ are differentiable. In general, $\theta(p)$ is continuous from the right.
The critical probability in two dimensions is $1/2$.

5.2 Random Jacobi matrices

Definition. A **Jacobi matrix** with IID potential $V_\omega(n)$ is a bounded self-adjoint operator on the Hilbert space

$$l^2(\mathbb{Z}) = \{(\dots, x_{-1}, x_0, x_1, x_2 \dots) \mid \sum_{k=-\infty}^{\infty} x_k^2 = 1 \}$$

of the form

$$L_\omega u(n) = \sum_{|m-n|=1} u(m) + V_\omega(n) u(n) = (\Delta + V_\omega)(u)(n) ,$$

where $V_\omega(n)$ are IID random variables in $\mathcal{L}^\infty$. These operators are called **discrete random Schrödinger operators**. We are interested in properties of L which hold for almost all $\omega \in \Omega$. In this section, we mostly write the elements ω of the probability space $(\Omega, \mathcal{A}, \mathrm{P})$ as a lower index.

Definition. A bounded linear operator L has **pure point spectrum**, if there exists a countable set of eigenvalues λ_i with eigenfunctions ϕ_i such that $L\phi_i = \lambda_i \phi_i$ and ϕ_i span the Hilbert space $l^2(\mathbb{Z})$. A random operator has **pure point spectrum** if L_ω has pure point spectrum for almost all $\omega \in \Omega$.

Our goal is to prove the following theorem:

Theorem 5.2.1 (Fröhlich-Spencer). Let $V(n)$ are IID random variables with uniform distribution on $[0,1]$. There exists λ_0 such that for $\lambda > \lambda_0$, the operator $L_\omega = \Delta + \lambda \cdot V_\omega$ has pure point spectrum for almost all ω.

We will give a recent elegant proof of Aizenman-Molchanov following [94].

Definition. Given $\mathrm{E} \in \mathbb{C} \setminus \mathbb{R}$, define the **Green function**

$$G_\omega(m, n, E) = [(L_\omega - E)^{-1}]_{mn} .$$

Let $\mu = \mu_\omega$ be the **spectral measure** of the vector e_0. This measure is defined as the functional $C(\mathbb{R}) \to \mathbb{R}, f \mapsto f(L_\omega)_{00}$ by $f(L_\omega)_{00} = \mathrm{E}[f(L)_{00}]$. Define the function

$$F(z) = \int_{\mathbb{R}} \frac{d\mu(y)}{y - z}$$

It is a function on the complex plane and called the **Borel transform** of the measure μ. An important role will play its derivative

$$F'(z) = \int_{\mathbb{R}} \frac{d\mu(\lambda)}{(y-z)^2} .$$

Definition. Given any Jacobi matrix L, let L_α be the operator $L + \alpha P_0$, where P_0 is the projection onto the one-dimensional space spanned by δ_0. One calls L_α a **rank-one perturbation** of L.

Theorem 5.2.2 (Integral formula of Javrjan-Kotani). The average over all specral measures $d\mu_\alpha$ is the Lebesgue measure:

$$\int_{\mathbb{R}} d\mu_\alpha \, d\alpha = dE \, .$$

Proof. The second resolvent formula gives

$$(L_\alpha - z)^{-1} - (L - z)^{-1} = -\alpha (L_\alpha - z)^{-1} P_0 (L - z)^{-1} \, .$$

Looking at 00 entry of this matrix identity, we obtain

$$F_\alpha(z) - F(z) = -\alpha F_\alpha(z) F(z)$$

which gives, when solved for F_α, the **Aronzajn-Krein formula**

$$F_\alpha(z) = \frac{F(z)}{1 + \alpha F(z)} \, .$$

We have to show that for any continuous function $f : \mathbb{C} \to \mathbb{C}$

$$\int_{\mathbb{R}} \int_{\mathbb{R}} f(x) \, d\mu_\alpha(x) \, d\alpha = \int f(x) \, dE(x)$$

and it is enough to verify this for the dense set of functions

$$\{ f_z(x) = (x - z)^{-1} - (x + i)^{-1} \, | z \in \mathbb{C} \setminus \mathbb{R} \} \, .$$

Contour integration in the upper half plane gives $\int_{\mathbb{R}} f_z(x) \; dx = 0$ for $\text{Im}(z) < 0$ and $2\pi i$ for $\text{Im}(z) > 0$. On the other hand

$$\int f_z(x) d\mu_\alpha(x) = F_\alpha(z) - F_\alpha(-i)$$

which is by the Aronzajn-Krain formula equal to

$$h_z(\alpha) := \frac{1}{\alpha + F(z)^{-1}} - \frac{1}{\alpha + F(-i)^{-1}} \, .$$

Now, if $\pm\text{Im}(z) > 0$, then $\pm\text{Im}F(z) > 0$ so that $\pm\text{Im}F(z)^{-1} < 0$. This means that $h_z(\alpha)$ has either two poles in the lower half plane if $\text{Im}(z) < 0$ or one in each half plane if $\text{Im}(z) > 0$. Contour integration in the upper half plane (now with α) implies that $\int_{\mathbb{R}} h_z(\alpha) \; d\alpha = 0$ for $\text{Im}(z) < 0$ and $2\pi i$ for $\text{Im}(z) > 0$. □

In theorem (2.12.2), we have seen that any Borel measure μ on the real line has a unique Lebesgue decomposition $d\mu = d\mu_{ac} + d\mu_{sing} = d\mu_{ac} + d\mu_{sc} + d\mu_{pp}$. The function F is related to this decomposition in the following way:

Proposition 5.2.3. (Facts about Borel transform) For $\epsilon \to 0$, the measures $\pi^{-1}\mathrm{Im}F(E+i\epsilon)\, dE$ converges weakly to μ.
$d\mu_{sing}(\{E \mid \mathrm{Im}F(E+i0) = \infty\}) = 1$,
$d\mu(\{E_0\}) = \lim_{\epsilon\to 0} \mathrm{Im}F(E_0 + i\epsilon)\epsilon$,
$d\mu_{ac}(E) = \pi^{-1}\mathrm{Im}F(E+i0)\, dE$.

Definition. Define for $\alpha \neq 0$ the sets

$$\begin{aligned} S_\alpha &= \{x \in \mathbb{R} \mid F(x+i0) = -\alpha^{-1},\ F'(x) = \infty\} \\ P_\alpha &= \{x \in \mathbb{R} \mid F(x+i0) = -\alpha^{-1},\ F'(x) < \infty\} \\ L &= \{x \in \mathbb{R} \mid \mathrm{Im}F(x+i0) \neq 0\} \end{aligned}$$

Lemma 5.2.4. (Aronzajn-Donoghue) The set P_α is the set of eigenvalues of L_α. One has $(d\mu_\alpha)_{sc}(S_\alpha) = 1$ and $(d\mu_\alpha)_{ac}(L) = 1$. The sets P_α, S_α, L are mutually disjoint.

Proof. If $F(E+i0) = -1/\alpha$, then

$$\lim_{\epsilon\to 0} \epsilon\, \mathrm{Im}F_\alpha(E+i\epsilon) = (\alpha^2 F'(E))^{-2}$$

since $F(E+i\epsilon) = -1/\alpha + i\epsilon F'(x) + o(\epsilon)$ if $F'(E) < \infty$ and $\epsilon^{-1}\mathrm{Im}(1+\alpha F) \to \infty$ if $F'(E) = \infty$ which means $\epsilon|1+\alpha F|^{-1} \to 0$ and since $F \to -1/\alpha$, one gets $\epsilon|F/(1+\alpha F)| \to 0$.

The theorem of de la Vallée Poussin (see [88]) states that the set

$$\{E \mid |F_\alpha(E+i0)| = \infty\}$$

has full $(d\mu_\alpha)_{sing}$ measure. Because $F_\alpha = F/(1+\alpha F)$, we know that $|F_\alpha(E+i0)| = \infty$ is equivalent to $F(E+i0) = -1/\alpha$. □

The following criterion of Simon-Wolff [96] will be important. In the case of IID potentials with absolutely continuous distribution, a spectral averaging argument will then lead to pure point spectrum also for $\alpha = 0$.

Theorem 5.2.5 (Simon-Wolff criterion). For any interval $[a,b] \subset \mathbb{R}$, the random operator L has pure point spectrum if

$$F'(E) < \infty$$

for almost almost all E $\in [a,b]$.

Proof. By hypothesis, the Lebesgue measure of $S = \{E \mid F'(E) = \infty\}$ is zero. This means by the integral formula that $d\mu_\alpha(S) = 0$ for almost all α. The Aronzajn-Donoghue lemma (5.2.4) implies

$$\mu_\alpha(S_\alpha \cap [a,b]) = \mu_\alpha(L \cap [a,b]) = 0$$

so that μ_α has only point spectrum. □

Lemma 5.2.6. (Formula of Simon-Wolff) For each E $\in \mathbb{R}$, the sum $\sum_{n\in\mathbb{Z}} |(L - E - i\epsilon)^{-1}_{0n}|^2$ increases monotonically as $\epsilon \searrow 0$ and converges point wise to $F'(E)$.

Proof. For $\epsilon > 0$, we have

$$\begin{aligned}\sum_{n\in\mathbb{Z}} |(L - E - i\epsilon)^{-1}_{0n}|^2 &= ||(L - E - i\epsilon)^{-1}\delta_0||^2 \\ &= |[(L - E - i\epsilon)^{-1}(L - E + i\epsilon)^{-1}]_{00}| \\ &= \int_{\mathbb{R}} \frac{d\mu(x)}{(x-E)^2 + \epsilon^2}\end{aligned}$$

from which the monotonicity and the limit follow. □

Lemma 5.2.7. There exists a constant C, such that for all $\alpha, \beta \in \mathbb{C}$

$$\int_0^1 |x-\alpha|^{1/2}|x-\beta|^{-1/2}\, dx \geq C \int_0^1 |x-\beta|^{-1/2}\, dx\,.$$

Proof. We can assume without loss of generality that $\alpha \in [0,1]$, because replacing a general $\alpha \in \mathbb{C}$ with the nearest point in $[0,1]$ only decreases the left hand side. Because the symmetry $\alpha \mapsto 1-\alpha$ leaves the claim invariant, we can also assume that $\alpha \in [0,1/2]$. But then

$$\int_0^1 |x-\alpha|^{1/2}|x-\beta|^{-1/2}\,dx \geq (\frac{1}{4})^{1/2}\int_{3/4}^1 |x-\beta|^{-1/2}\,dx .$$

The function

$$h(\beta) = \frac{\int_{3/4}^1 |x-\beta|^{-1/2}\,dx}{\int_0^1 |x-\alpha|^{1/2}|x-\beta|^{-1/2}\,dx}$$

is non-zero, continuous and satisfies $h(\infty) = 1/4$. Therefore

$$C := \inf_{\beta\in\mathbb{C}} h(\beta) > 0 .$$

□

The next lemma is an estimate for the free Laplacian.

Lemma 5.2.8. Let $f, g \in l^\infty(\mathbb{Z})$ be nonnegative and let $0 < a < (2d)^{-1}$.

$$(1-a\Delta)f \leq g \Rightarrow f \leq (1-a\Delta)^{-1}g .$$

$$[(1-a\Delta)^{-1}]_{ij} \leq (2da)^{|j-i|}(1-2da)^{-1} .$$

Proof. Since $||\Delta|| < 2d$, we can write $(1-a\Delta)^{-1} = \sum_{m=0}^\infty (a\Delta)^m$ which is preserving positivity. Since $[(a\Delta)^m]_{ij} = 0$ for $m < |i-j|$ we have

$$[(a\Delta)^m]_{ij} = \sum_{m=|i-j|}^\infty [(a\Delta)^m]_{ij} \leq \sum_{m=|i-j|}^\infty (2da)^m .$$

□

We come now to the proof of theorem (5.2.1):

Proof. In order to prove theorem (5.2.1), we have by Simon-Wolff only to show that $F'(E) < \infty$ for almost all E. This will be achieved by proving $\mathrm{E}[F'(E)^{1/4}] < \infty$. By the formula of Simon-Wolff, we have therefore to show that

$$\sup_{z\in\mathbb{C}} \mathrm{E}[(\sum_n |G(n,0,z)|^2)^{1/4}] < \infty .$$

Since

$$(\sum_n |G(n,0,z)|^2)^{1/4} \leq \sum_n |G(n,0,z)|^{1/2} ,$$

we have only to control the later the term. Define $g_z(n) = G(n,0,z)$ and $k_z(n) = \mathrm{E}[|g_z(n)|^{1/2}]$. The aim is now to give an estimate for

$$\sum_{n\in\mathbb{Z}} k_z(n)$$

which holds uniformly for $\mathrm{Im}(z) \neq 0$.

(i)

$$\mathrm{E}[|\lambda V(n) - z|^{1/2}|g_z(n)|^{1/2}] \leq \delta_{n,0} + \sum_{|j|=1} k_z(n+j) \; .$$

Proof. $(L-z)g_z(n) = \delta_{n0}$ means

$$(\lambda V(n) - z)g_z(n) = \delta_{n0} - \sum_{|j|=1} g_z(n+j) \; .$$

Jensen's inequality gives

$$\mathrm{E}[|\lambda V(n) - z|^{1/2}|g_z(n)|^{1/2}] \leq \delta_{n0} + \sum_{|j|=1} k_z(n+j) \; .$$

(ii)

$$\mathrm{E}[|\lambda V(n) - z|^{1/2}|g_z(n)|^{1/2}] \geq C\lambda^{1/2}k(n) \; .$$

Proof. We can write $g_z(n) = A/(\lambda V(n) + B)$, where A, B are functions of $\{V(l)\}_{l\neq n}$. The independent random variables $V(k)$ can be realized over the probability space $\Omega = [0,1]^{\mathbb{Z}} = \prod_{k\in\mathbb{Z}} \Omega(k)$. We average now $|\lambda V(n) - z|^{1/2}|g_z(n)|^{1/2}$ over $\Omega(n)$ and use an elementary integral estimate:

$$\begin{aligned}
\int_{\Omega(n)} \frac{|\lambda v - z|^{1/2}|A|^{1/2}}{|\lambda v + B|^{1/2}} \, dv &= |A|^{1/2} \int_0^1 |v - z\lambda^{-1}||v + B\lambda^{-1}|^{-1/2} \, dv \\
&\geq C|A|^{1/2} \int_0^1 |v + B\lambda^{-1}|^{-1/2} \, dv \\
&= C\lambda^{1/2} \int_0^1 |A/(\lambda v + B)|^{1/2} \\
&= \mathrm{E}[g_z(n)^{1/2}] = k_z(n) \; .
\end{aligned}$$

(iii)

$$k_z(n) \leq (C\lambda^{1/2})^{-1} \left(\sum_{|j|=1} k_z(n+j) + \delta_{n0} \right) \; .$$

Proof. Follows directly from (i) and (ii).

(iv)

$$(1 - C\lambda^{1/2}\Delta)k \leq \delta_{n0} \; .$$

Proof. Rewriting (iii).

(v) Define $\alpha = C\lambda^{1/2}$.

$$k_z(n) \leq \alpha^{-1}(2d/\alpha)^{|n|}(1-2d/\alpha)^{-1} .$$

Proof. For $\mathrm{Im}(z) \neq 0$, we have $k_z \in l^\infty(\mathbb{Z})$. From lemma (5.2.8) and (iv), we have

$$k(n) \leq \alpha^{-1}[(1-\Delta/\alpha)^{-1}]_{0n} \leq \alpha^{-1}(\frac{2}{\alpha})^{|n|}(1-\frac{2}{\alpha})^{-1} .$$

(vi) For $\lambda > 4C^{-2}$, we get a uniform bound for $\sum_n k_z(n)$.
Proof. Since $C\lambda^{1/2} < 1/2$, we get the estimate from (v).
(vii) Pure point spectrum.
Proof. By Simon-Wolff, we have pure point spectrum for L_α for almost all α. Because the set of random operators of L_α and L_0 coincide on a set of measure $\geq 1-2\alpha$, we get also pure point spectrum of L_ω for almost all ω. □

5.3 Estimation theory

Estimation theory is a branch of mathematical statistics. The aim is to estimate continuous or discrete parameters for models in an optimal way. This leads to extremization problems. We start with some terminology.

Definition. A collection $(\Omega, \mathcal{A}, \mathrm{P}_\theta)$ of probability spaces is called a **statistical model**. If X is a random variable, its **expectation** with respect to the measure P_θ is denoted by $\mathrm{E}_\theta[X]$, its **variance** is $\mathrm{Var}_\theta[X] = \mathrm{E}_\theta[(X-\mathrm{E}_\theta[X])^2]$. If X is continuous, then its probability density function is denoted by f_θ. In that case one has of course $\mathrm{E}_\theta[X] = \int_\Omega f_\theta(x)\, dx$. The parameters θ are taken from a **parameter space** Θ, which is assumed to be a subset of $\mathbb{R}$ or $\mathbb{R}^k$.

Definition. A probability distribution $\mu = p(\theta)\, d\theta$ on $(\Theta, \mathcal{B})$ is called an **a priori distribution** on $\Theta \subset \mathbb{R}$. It allows to define the **global expectation** $\mathrm{E}[X] = \int_\Theta \mathrm{E}_\theta[X]\, d\mu(\theta)$.

Definition. Given n independent and identically distributed random variables $X_1, \dots, X_n$ on the probability space $(\Omega, \mathcal{A}, \mathrm{P}_\theta)$, we want to estimate a **quantity** $g(\theta)$ using an **estimator** $T(\omega) = t(X_1(\omega), \dots, X_n(\omega))$.

Example. If the quantity $g(\theta) = \mathrm{E}_\theta[X_i]$ is the expectation of the random variables, we can look at the estimator $T(\omega) = \frac{1}{n}\sum_{j=1}^n X_i(\omega)$, the arithmetic mean. The arithmetic mean is natural because for any data $x_1, \dots, x_n$, the function $f(x) = \sum_{i=1}^n (x_i - x)^2$ is minimized by the arithmetic mean of the data.

Example. We can also take the estimator $T(\omega)$ which is the **median** of $X_1(\omega), \dots, X_n(\omega)$. The median is a natural quantity because the function $f(x) = \sum_{i=1}^n |x_i - x|$ is minimized by the median. Proof. $|a-x| + |b-x| =$

$|b-a| + C(x)$, where $C(x)$ is zero if $a \leq x \leq b$ and $C(x) = x - b$ if $x > b$ and $D(x) = a - x$ if $x < a$. If $n = 2m+1$ is odd, we have $f(x) = \sum_{j=1}^{m} |x_i - x_{n+1-i}| + \sum_{x_j > x_m} C(x_j) + \sum_{x_j < x_m} D(x_j)$ which is minimized for $x = x_m$. If $n = 2m$, we have $f(x) = \sum_{j=1}^{m} |x_i - x_{n+1-i}| + \sum_{x_j > x_{m+1}} C(x_j) + \sum_{x_j < x_{m-1}} D(x_j)$ which is minimized for $x \in [x_m, x_{m+1}]$.

Example. Define the **bias** of an estimator T as

$$B(\theta) = B_\theta[T] = \mathrm{E}_\theta[T] - g(\theta) .$$

The bias is also called the **systematic error.** If the bias is zero, the estimator is called **unbiased.** With an a priori distribution on Θ, one can define the **global error** $B(T) = \int_\Theta B(\theta)\, d\mu(\theta)$.

Proposition 5.3.1. A linear estimator $T(\omega) = \sum_{j=1}^{n} \alpha_i X_i(\omega)$ with $\sum_i \alpha_i = 1$ is unbiased for the estimator $g(\theta) = \mathrm{E}_\theta[X_i]$.

Proof. $\mathrm{E}_\theta[T] = \sum_{j=1}^{n} \alpha_i \mathrm{E}_\theta[X_i] = \mathrm{E}_\theta[X_i]$. □

Proposition 5.3.2. For $g(\theta) = \mathrm{Var}_\theta[X_i]$ and fixed mean m, the estimator $T = \frac{1}{n}\sum_{j=1}^{n}(X_i - m)^2$ is unbiased. If the mean is unknown, the estimator $T = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ with $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is unbiased.

Proof. a) $\mathrm{E}_\theta[T] = \frac{1}{n}\sum_{j=1}^{n}(X_i - m)^2 = \mathrm{Var}_\theta[T] = g(\theta)$.

b) For $T = \frac{1}{n}\sum_i (X_i - \overline{X}_i)^2$, we get

$$\begin{aligned}
\mathrm{E}_\theta[T] &= \mathrm{E}_\theta[X_i^2] - \mathrm{E}_\theta[\frac{1}{n^2}\sum_{i,j} X_i X_j] \\
&= \mathrm{E}_\theta[X_i^2] - \frac{1}{n}\mathrm{E}_\theta[X_i^2] - \frac{n(n-1)}{n^2}\mathrm{E}_\theta[X_i]^2 \\
&= (1 - \frac{1}{n})\mathrm{E}_\theta[X_i^2] - \frac{n-1}{n}\mathrm{E}_\theta[X_i]^2 \\
&= \frac{n-1}{n}\mathrm{Var}_\theta[X_i] .
\end{aligned}$$

Therefore $n/(n-1)T$ is the correct unbiased estimate. □

Remark. Part b) is the reason, why statisticians often take the average of $\frac{n}{(n-1)}(x_i - \overline{x})^2$ as an estimate for the variance of n data points x_i with mean $\overline{m}$ if the actual mean value m is not known.

Definition. The expectation of the quadratic estimation error

$$\mathrm{Err}_\theta[T] = \mathrm{E}_\theta[(T - g(\theta))^2]$$

is called the **risk function** or the **mean square error** of the estimator T. It measures the estimator performance. We have

$$\mathrm{Err}_\theta[T] = \mathrm{Var}_\theta[T] + B_\theta[T] \ ,$$

where $B_\theta[T]$ is the bias.

Example. If T is unbiased, then $\mathrm{Err}_\theta[T] = \mathrm{Var}_\theta[T]$.

Example. The arithmetic mean is the "best linear unbiased estimator". Proof. With $T = \sum_i \alpha_i X_i$, where $\sum_i \alpha_i = 1$, the risk function is

$$\mathrm{Err}_\theta[T] = \mathrm{Var}_\theta[T] = \sum_i \alpha_i^2 \mathrm{Var}_\theta[X_i] \ .$$

It is by Lagrange minimal for $\alpha_i = 1/n$.

Definition. For continuous random variables, the **maximum likelihood function** $t(x_1, \dots, x_n)$ is defined as the maximum of $\theta \mapsto L_\theta(x_1, \dots, x_n) := f_\theta(x_1) \cdot \dots \cdot f_\theta(x_n)$. The **maximum likelihood estimator** is the random variable

$$T(\omega) = t(X_1(\omega), \dots, X_n(\omega)) \ .$$

For discrete random variables, $L_\theta(x_1, \dots, x_n)$ would be replaced by $\mathrm{P}_\theta[X_1 = x_1, \dots, X_n = x_n]$.
One also looks at the **maximum a posteriori** estimator, which is the maximum of

$$\theta \mapsto L_\theta(x_1, \dots, x_n) = f_\theta(x_1) \cdot \dots \cdot f_\theta(x_n) p(\theta) \ ,$$

where $p(\theta)\ d\theta$ was the a priori distribution on Θ.

Definition. The **minimax principle** is the aim to find

$$\min_T \max_\theta R(\theta, T) \ .$$

The **Bayes principle** is the aim to find

$$\min_T \int_\Theta (R(\theta, T)\ d\mu(\theta) \ .$$

Example. Assume $f_\theta(x) = \frac{1}{2} e^{-|x-\theta|}$. The maximum likelihood function

$$L_\theta(x_1, \dots, x_n) = \frac{1}{2^n} e^{-\sum_j |x_i - \theta|}$$

is maximal when $\sum_j |x_i - \theta|$ is minimal which means that $t(x_1, \dots, x_n)$ is the **median** of the data $x_1, \dots, x_n$.

Example. Assume $f_\theta(x) = \theta^x e^{-\theta}/x!$ is the probability density of the Poisson distribution. The maximal likelihood function

$$l_\theta(x_1, \dots, x_n) = \frac{e^{\sum_i \log(\theta) x_i - n\theta}}{x_1! \cdots x_n!}$$

is maximal for $\theta = \sum_{i=1}^n x_i/n$.

Example. The maximum likelihood estimator for $\theta = (m, \sigma^2)$ for Gaussian distributed random variables $f_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-m)^2}{2\sigma^2}}$ has the maximum likelihood function maximized for

$$t(x_1, \dots, x_n) = (\frac{1}{n} \sum_i x_i, \frac{1}{n} \sum_i (x_i - \overline{x})^2) .$$

Definition. Define the **Fisher information** of a random variable X with density f_θ as

$$I(\theta) = \int (\frac{f'_\theta(x)}{f_\theta(x)})^2 f_\theta(x) \, dx .$$

If θ is a vector, one defines the **Fisher information matrix**

$$I_{ij}(\theta) = \int \frac{f'_{\theta_i} f'_{\theta_j}}{f_\theta^2} f_\theta \, dx .$$

Lemma 5.3.3. $I(\theta) = \mathrm{Var}_\theta[\frac{f'_\theta}{f_\theta}]$.

Proof. $\mathrm{E}[\frac{f'_\theta}{f_\theta}] = \int_\Omega f'_\theta dx = 0$ so that

$$\mathrm{Var}_\theta[\frac{f'_\theta}{f_\theta}] = \mathrm{E}_\theta[(\frac{f'_\theta}{f_\theta})^2] .$$

□

Lemma 5.3.4. $I(\theta) = -\mathrm{E}_\theta[(\log(f_\theta)'']$.

Proof. Integration by parts gives:

$$\mathrm{E}[\log(f_\theta)''] = \int \log(f_\theta)'' f_\theta \, dx = -\int \log(f_\theta)' f'_\theta \, dx = -\int (f'_\theta/f_\theta)^2 f_\theta \, dx .$$

□

Definition. The **score function** for a continuous random variable is defined as the logarithmic derivative $\rho_\theta = f_\theta'/f_\theta$. One has $I(\theta) = \mathrm{E}_\theta[\rho_\theta^2] = \mathrm{Var}_\theta[\rho_\theta]$.

Example. If X is a Gaussian random variable, the score function $\rho_\theta = f'(\theta)/f(\theta) = -(x-m)/(\sigma^2)$ is linear and has variance 1. The Fisher information I is $1/\sigma^2$. We see that $\mathrm{Var}[X] = 1/I$. This is a special case $n = 1, T = X, \theta = m$ of the following bound:

Theorem 5.3.5 (Rao-Cramer inequality).

$$\mathrm{Var}_\theta[T] \geq \frac{(1+B'(\theta))^2}{nI(\theta)} .$$

In the unbiased case, one has

$$\mathrm{Err}_\theta[T] \geq \frac{1}{nI(\theta)}$$

Proof. 1) $\theta + B(\theta) = \mathrm{E}_\theta[T] = \int t(x_1,\dots,x_n)L_\theta(x_1,\dots,x_n)\,dx_1\cdots dx_n$.
2)

$$\begin{aligned}
1+B'(\theta) &= \int t(x_1,\dots,x_n)L_\theta'(x_1,\dots,x_n)\,dx_1\cdots dx_n \\
&= \int t(x_1,\dots,x_n)\frac{L_\theta'(x_1,\dots,x_n)}{L_\theta(x_1,\dots,x_n)}\,dx_1\cdots dx_n \\
&= \mathrm{E}_\theta[T\frac{L_\theta'}{L_\theta}]
\end{aligned}$$

3) $1 = \int L_\theta(x_1,\dots,x_n)\,dx_1\cdots dx_n$ implies

$$0 = \int L_\theta'(x_1,\dots,x_n)/L_\theta(x_1,\dots,x_n) = \mathrm{E}[L_\theta'/L_\theta] .$$

4) Using 3) and 2)

$$\begin{aligned}
\mathrm{Cov}[T, L_\theta'/L_\theta] &= \mathrm{E}_\theta[TL_\theta'/L_\theta] - 0 \\
&= 1+B'(\theta) .
\end{aligned}$$

5)

$$\begin{aligned}
(1+B'(\theta))^2 &= \mathrm{Cov}^2[T, \frac{L_\theta'}{L_\theta}] \\
&\leq \mathrm{Var}_\theta[T]\mathrm{Var}_\theta[\frac{L_\theta'}{L_\theta}] \\
&= \mathrm{Var}_\theta[T]\sum_{i=1}^n \mathrm{E}_\theta[(\frac{f_\theta'(x_i)}{f_\theta(x_i)})^2] \\
&= \mathrm{Var}_\theta[T]\,nI(\theta) ,
\end{aligned}$$

where we used 4), the lemma and

$$L_\theta'/L_\theta = \sum_{i=1}^n f_\theta'(x_i)/f_\theta(x_i) .$$

□

Definition. Closely related to the Fisher information is the already defined **Shannon entropy** of a random variable X:

$$S(\theta) = -\int f_\theta \log(f_\theta) \, dx ,$$

as well as the **power entropy**

$$N(\theta) = \frac{1}{2\pi e} e^{2S(\theta)} .$$

Theorem 5.3.6 (Information Inequalities). If X, Y are independent random variables then the following inequalities hold:
a) **Fisher information inequality**: $I_{X+Y}^{-1} \geq I_X^{-1} + I_Y^{-1}$.
b) **Power entropy inequality**: $N_{X+Y} \geq N_X + N_Y$.
c) **Uncertainty property**: $I_X N_X \geq 1$.
In all cases, equality holds if and only if the random variables are Gaussian.

Proof. a) $I_{X+Y} \leq c^2 I_X + (1-c)^2 I_Y$ is proven using the Jensen inequality (2.5.1). Take then $c = I_Y/(I_X + I_Y)$.
b) and c) are exercises. □

Theorem 5.3.7 (Rao-Cramer bound). A random variable X with mean m and variance σ^2 satisfies: $I_X \geq 1/\sigma^2$. Equality holds if and only if X is the Normal distribution.

Proof. This is a special case of Rao-Cramer inequality, where θ is fixed, $n = 1$. The bias is automatically zero. A direct computation giving also uniqueness: $\mathrm{E}[(aX+b)\rho(X)] = \int (ax+b) f'(x) \, dx = -a \int f(x) \, dx = -a$ implies

$$\begin{aligned} 0 &\leq \mathrm{E}[(\rho(X) + (X-m)/\sigma^2)^2] \\ &= \mathrm{E}[(\rho(X)^2] + 2\mathrm{E}[(X-m)\rho(X)]/\sigma^2 + \mathrm{E}[(X-m)^2/\sigma^4] \\ &\leq I_X - 2/\sigma^2 + 1\sigma_2 . \end{aligned}$$

Equality holds if and only if ρ_X is linear, that is if X is normal. □

We see that the normal distribution has the smallest Fisher information among all distributions with the same variance σ^2.

5.4 Vlasov dynamics

Vlasov dynamics generalizes Hamiltonian n-body particle dynamics. It deals with the evolution of the law P^t of a discrete random vector X^t. If P^t is a discrete measure located on finitely many points, then it is the usual dynamics of n bodies which attract or repel each other. In general, the stochastic process X^t describes the evolution of densities or the evolution of surfaces. It is an important feature of Vlasov theory that while the random variables X^t stay smooth, their laws P^t can develop singularities. This can be useful to model shocks. Due to the overlap of this section with geometry and dynamics, the notation slightly changes in this section. We write X^t for the stochastic process for example and not X_t as before.

Definition. Let $\Omega = M$ be a $2p$-dimensional Euclidean space or torus with a probability measure m and let N be an Euclidean space of dimension $2q$. Given a potential $V : \mathbb{R}^q \to \mathbb{R}$, the **Vlasov flow** $X^t = (f^t, g^t) : M \to N$ is defined by the differential equation

$$\dot{f} = g, \dot{g} = -\int_M \nabla V(f(\omega) - f(\eta))\, dm(\eta) .$$

These equations are called the **Hamiltonian equations** of the Vlasov flow. We can interpret X^t as a vector-valued stochastic process on the probability space $(M, \mathcal{A}, m)$. The probability space $(M, \mathcal{A}, m)$ labels the particles which move on the target space N.

Example. If $p = 0$ and M is a finite set $\Omega = \{\omega_1, \dots, \omega_n\}$, then X^t describes the evolution of n particles $(f_i, g_i) = X(\omega_i)$. Vlasov dynamics is therefore a generalization of n-body dynamics. For example, if

$$V(x_1, \dots .x_n) = \sum_i \frac{x_i^2}{2} ,$$

then $\nabla V(x) = x$ and the Vlasov Hamiltonian system

$$\dot{f} = g, \dot{g}(\omega) = -\int_M f(\omega) - f(\eta)\, dm(\eta)$$

is equivalent to the n-body evolution

$$\begin{aligned} \dot{f}_i &= g_i \\ \dot{g}_i &= -\sum_{j=1}^n (f_i - f_j) . \end{aligned}$$

In a center of mass coordinate system where $\sum_{i=1}^n f_i(x) = 0$, this simplifies to a system of coupled harmonic oscillators

$$\frac{d^2}{dt^2} f_i(x) = -f_i(x) .$$

Example. If $N = M = \mathbb{R}^2$ and m is a measure, then the process X^t describes a volume-preserving deformation of the plane M. In other words, X^t is a one-parameter family of volume-preserving diffeomorphisms in the plane.

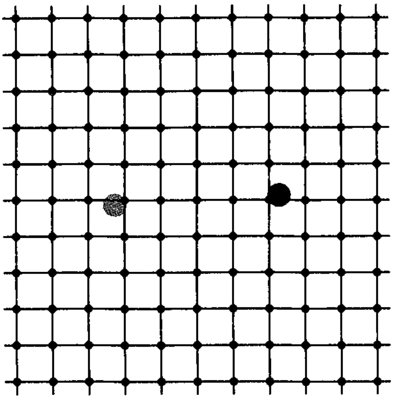


Figure. *An example with $M = N = \mathbb{R}^2$, where the measure m is located on 2 points. The Vlasov evolution describes a deformation of the plane. The situation is shown at time $t = 0$. The coordinates (x, y) describe the position and the speed of the particles.*

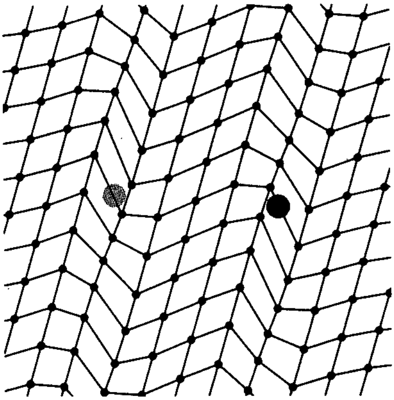


Figure. *The situation at time $t = 0.1$. The two particles have evolved in the phase space N. Each point moves as "test particle" in the force field of the 2 particles. Even so the 2 body problem is integrable, its periodic motion acts like a "mixer" for the complicated evolution of the test particles.*

Example. Let $M = N = \mathbb{R}^2$ and assume that the measure m has its support on a smooth closed curve C. The process X^t is again a volume-preserving deformation of the plane. It describes the evolution of a continuum of particles on the curve. Dynamically, it can for example describe the evolution of a curve where each part of the curve interacts with each other part. The picture sequence below shows the evolution of a particle gas with support on a closed curve in phase space. The interaction potential is $V(x) = e^{-x}$. Because the curve at time t is the image of the diffeomorphism X^t, it will never have self intersections. The curvature of the curve is expected to grow exponentially at many points. The deformation transformation $X^t = (f^t, g^t)$ satisfies the differential equation

$$\begin{aligned} \frac{d}{dt}f &= g \\ \frac{d}{dt}g &= \int_M e^{-(f(\omega)-f(\eta))} \, dm(\eta) \, . \end{aligned}$$

If $r(s), s \in [0, 1]$ is the parameterization of the curve C so that $m(r[a, b]) =$

$(b-a)$, then the equations are

$$\begin{aligned} \frac{d}{dt} f^t(x) &= g^t(x) \\ \frac{d}{dt} g^t(x) &= \int_0^1 e^{-(f^t(x)-f^t(r(s)))} \, ds \, . \end{aligned}$$

The evolved curve C^t at time t is parameterized by $s \to (f^t(r(s)), g^t(r(s)))$.

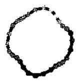


Figure. *The support of the measure* P^0 *on* $N = \mathbb{R}^2$.

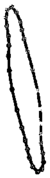


Figure. *The support of the measure* $\mathrm{P}^{0.4}$ *on* $N = \mathbb{R}^2$.

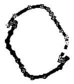


Figure. *The support of the measure* $\mathrm{P}^{1.2}$ *on* $N = \mathbb{R}^2$.

Example. If X^t is a stochastic process on $(\Omega = M, \mathcal{A}, m)$ with takes values in N, then P^t is a probability measure on N defined by $\mathrm{P}^t[A] = m(X^{-1}A)$. It is called the **push-forward measure** or **law** of the random vector X. The measure P^t is a measure in the phase space N. The Vlasov evolution defines a family of probability spaces $(N, \mathcal{B}, \mathrm{P}^t)$. The spatial **particle density** ρ is the law of the random variable $x(x,y) = x$.

Example. Assume the measure P^0 is located on a curve $\vec{r}(s) = (s, \sin(s))$ and assume that there is no particle interaction at all: $V = 0$. Then P^t is supported on a curve $(s + \sin(s), \sin(s))$. While the **spatial particle density** has initially a smooth density $\sqrt{1 + \cos(s)^2}$, it becomes discontinuous after some time.

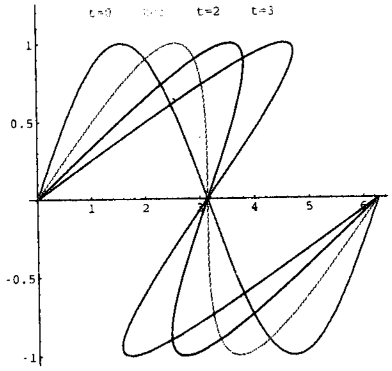


Figure. *Already for the free evolution of particles on a curve in phase space, the spatial particle density can become non-smooth after some time.*

Example. In the case of the quadratic potential $V(x) = x^2/2$ assume m has a density $\rho(x,y) = e^{-x^2-2y^2}$, then P^t has the density $\rho^t(x,y) = f(x\cos(t) + y\sin(t), -x\sin(t) + y\cos(t))$. To get from this density in the phase space, the spatial density of particles, we have to do integrate y out and do a conditional expectation.

Lemma 5.4.1. (Maxwell) If $X^t = (f^t, g^t)$ is a solution of the Vlasov Hamiltonian flow, then the law $\mathrm{P}^t = (X^t)^* m$ satisfies the Vlasov equation

$$\dot{\mathrm{P}}^t(x,y) + y \cdot \nabla_x \mathrm{P}^t(x,y) - W(x) \cdot \nabla_y \mathrm{P}^t(x,y) = 0$$

with $W(x) = \int_M \nabla_x V(x - x') \cdot \mathrm{P}^t(x', y')) \, dy' dx'$.

Proof. We have $\int \nabla V(f(\omega) - f(\eta)) \, dm(\eta) = W(f(\omega))$. Given a smooth function h on N of compact support, we calculate

$$L = \int_N h(x,y) \frac{d}{dt} P^t(x,y) \, dxdy$$

as follows:

$$\begin{aligned} L &= \frac{d}{dt}\int_N h(x,y)P^t(x,y)\,dxdy \\ &= \frac{d}{dt}\int_M h(f(\omega,t),g(\omega,t))\,dm(\omega) \\ &= \int_M \nabla_x h(f(\omega,t),g(\omega,t))g(\omega,t)\,dm(\omega) \\ &\quad -\int_M \nabla_y h(f(\omega,t),g(\omega,t))\int_M \nabla V(f(\omega)-f(\eta))\,dm(\eta)\,dm(\omega) \\ &= \int_N \nabla_x h(x,y)yP^t(x,y)\,dxdy - \int_N P^t(x,y)\nabla_y h(x,y) \\ &\quad \int_N \nabla V(x-x')P^t(x',y')\,dx'dy'dxdy \\ &= -\int_N h(x,y)\nabla_x P^t(x,y)y\,dxdy \\ &\quad +\int_N h(x,y)W(x)\cdot\nabla_y P^t(x,y)\,dxdy \end{aligned}$$

□

Remark. The Vlasov equation is an example of an **integro-differential equation.** The right hand side is an integral. In a short hand notation, the Vlasov equation is

$$\dot{\mathrm{P}} + y\cdot P_x - W(x)\cdot \mathrm{P}_y = 0\,,$$

where $W = \nabla_x V \star \mathrm{P}$ is the convolution of the force $\nabla_x V$ with P.

Example. $V(x)=0$. Particles move freely. The Vlasov equation becomes the **transport equation** $\dot{P}(x,y,t) + y\cdot\nabla_x P^t(x,y) = 0$ which is in one dimensions a partial differential equation $u_t + yu_x = 0$. It has solutions $u(t,x,y) = u(u,x+ty)$. Restricting this function to $y=x$ gives the **Burgers equation** $u_t + xu_x = 0$.

Example. For a quadratic potential $V(x)=x^2$, the Hamilton equations are

$$\ddot{f}(\omega) = -(f(\omega) - \int_M f(\eta)\,dm(\eta))\,.$$

In center-of-mass-coordinates $\tilde{f} = f - E[f]$, the system is a decoupled system of a continuum of oscillators $\dot{f} = g, \dot{g} = -f$ with solutions

$$f(t) = f(0)\cos(t) + g(0)\sin(t),\ g(t) = -f(0)\sin(t) + g(0)\cos(t)\,.$$

The evolution for the density P is the partial differential equation

$$\frac{d}{dt}\mathrm{P}^t(x,y) + y\cdot\nabla_x\mathrm{P}^t(x,y) - x\cdot\nabla_y\mathrm{P}^t(x,y) = 0$$

written in short hand as $u_t + y\cdot u_x - x\cdot u_y = 0$, which has the explicit solution $P^t(x,y) = P^0(\cos(t)x + \sin(t)y, -\sin(t)x + \cos(t)y)$. It is an example of a **Hamilton-Jacobi equation.**

Example. On any Riemannian manifold with Laplace-Beltrami operator Δ, there are natural potentials: the **Poisson equation** $\Delta\phi = \rho$ is solved by $\phi = V \star \rho$, where $\star$ is the convolution. This defines **Newton potentials** on the manifold. Here are some examples:

- $N = \mathbb{R}$: $V(x) = \frac{|x|}{2}$.
- $N = \mathbb{T}$: $V(x) = \frac{|x(2\pi - x)|}{4\pi}$
- $N = \mathbf{S}^2$. $V(x) = \log(1 - x \cdot x)$.
- $N = \mathbb{R}^2$ $V(x) = \frac{1}{2\pi} \log|x|$.
- $N = \mathbb{R}^3$ $V(x) = \frac{1}{4\pi}\frac{1}{|x|}$.
- $N = \mathbb{R}^4$ $V(x) = \frac{1}{8\pi}\frac{1}{|x|^2}$.

For example, for $N = \mathbb{R}$, the Laplacian $\Delta f = f''$ is the second derivative. It is diagonal in Fourier space: $\hat{\Delta}\hat{f}(k) = -k^2\hat{f}$, where $k \in \mathbb{R}$. From $D\hat{elta}\hat{f}(k) = -k^2\hat{f} = \hat{\rho}(k)$ we get $\hat{f}(k) = -(1/k^2)\hat{\rho}(k)$, so that $f = V \star \rho$, where V is the function which has the Fourier transform $\hat{V}(k) = -1/k^2$. But $V(x) = |x|/2$ has this Fourier transform:

$$\int_{-\infty}^{\infty} \frac{|x|}{2} e^{-ikx}\, dx = -\frac{1}{k^2} \, .$$

Also for $N = \mathbb{T}$, the Laplacian $\Delta f = f''$ is diagonal in Fourier space. It is the 2π-periodic function $V(x) = x(2\pi - x)/(4\pi)$, which has the Fourier series $\hat{V}(k) = -1/k^2$.

For general $N = \mathbb{R}^n$, see for example [58]

Remark. The function $G_y(x) = V(x - y)$ is also called the **Green function** of the Laplacian. Because Newton potentials V are not smooth, establishing global existence for the Vlasov dynamics is not easy but it has been done in many cases [30]. The potential $|x|$ models galaxy motion and appears in plasma dynamics [90, 65, 82].

Lemma 5.4.2. (Gronwall) If a function u satisfies $u'(t) \le |g(t)|u(t)$ for all $0 \le t \le T$, then $u(t) \le u(0)\exp(\int_0^t |g(s)|\, ds)$ for $0 \le t \le T$.

Proof. Integrating the assumption gives $u(t) \le u(0) + \int_0^t g(s)u(s)\, ds$. The function $h(t)$ satisfying the differential equation $h'(t) = |g(t)|u(t)$ satisfies $h'(t) \le |g(t)|h(t)$. This leads to $h(t) \le h(0)\exp(\int_0^t |g(s)|\, ds)$ so that $u(t) \le u(0)\exp(\int_0^t |g(s)|\, ds)$. This proof for real valued functions [20] generalizes to the case, where $u^t(x)$ evolves in a function space. One just can apply the same proof for any fixed x. □

Theorem 5.4.3 (Batt-Neunzert-Brown-Hepp-Dobrushin). If $\nabla_x V$ is bounded and globally Lipshitz continuous, then the Hamiltonian Vlasov flow has a unique global solution X^t and consequently, the Vlasov equation has a unique and global solution P^t in the space of measures. If V and P^0 are smooth, then P^t is piecewise smooth.

Proof. The Hamiltonian differential equation for $X = (f, g)$ evolves on on the complete metric space of all continuous maps from M to N. The distance is $d(X,Y) = \sup_{\omega \in M} d(X(\omega), Y(\omega))$, where d is the distance in N.

We have to show that the differential equation $\dot{f} = g$ and $\dot{g} = G(f) = -\int_M \nabla_x V(f(\omega) - f(\eta))\, dm(\eta)$ in $C(M,N)$ has a unique solution: because of Lipshitz continuity

$$||G(f) - G(f')||_\infty \leq 2||D(\nabla_x V)||_\infty \cdot ||f - f'||_\infty$$

the standard Piccard existence theorem for differential equations assures local existence of solutions.

The Gronwall's lemma assures that $||X(\omega)||$ can not grow faster than exponentially. This gives the global existence. □

Remark. If m is a point measure supported on finitely many points, then one could also invoke the global existence theorem for differential equations. For smooth potentials, the dynamics depends continuously on the measure m. One could approximate a smooth measure m by point measures.

Definition. The evolution of DX^t at a point $\omega \in M$ is called the **linearized Vlasov flow**. It is the differential equation

$$D\ddot{f}(\omega) = -\int_M \nabla^2 V(f(\omega) - f(\eta))\, dm(\eta) Df(\omega) =: B(f^t) Df(\omega)$$

and we can write it as a first order differential equation

$$\begin{aligned} \frac{d}{dt} DX &= \frac{d}{dt} \begin{bmatrix} f \\ g \end{bmatrix} \\ &= \begin{bmatrix} 0 & 1 \\ \int_M -\nabla^2 V(f(\omega) - f(\eta))\, dm(\eta) & 0 \end{bmatrix} \begin{bmatrix} f \\ g \end{bmatrix} \\ &= A(f^t) \begin{bmatrix} f \\ g \end{bmatrix} . \end{aligned}$$

Remark. The rank of the matrix $DX^t(\omega)$ stays constant. $Df^t(\omega)$ is a linear combination of $Df^0(\omega)$ and $Dg^0(\omega)$. Critical points of f^t can only appear for ω, where $Df^0(\omega), D\dot{f}^0(\omega)$ are linearly dependent. More generally $Y_k(t) = \{\omega \in M \mid DX^t(\omega) \text{ has rank } 2q - k = dim(N) - k\}$ is time independent. The set Y_q contains $\{\omega \mid D(f)(\omega) = \lambda D(g)(\omega), \lambda \in \mathbf{R} \cup \{\infty\}\}$.

Definition. The random variable

$$\lambda(\omega) = \limsup_{t\to\infty} \frac{1}{t} \log(||D(X^t(\omega))||) \in [0,\infty]$$

is called the maximal **Lyapunov exponent** of the $SL(2q,\mathbf{R})$-cocycle $A^t = A(f^t)$ along an orbit $X^t = (f^t, g^t)$ of the Vlasov flow. The Lyapunov exponent could be infinite. Differentiation of $D\ddot{f} = B(f^t)f^t$ at a critical point ω^t gives $D^2\ddot{f}^t(\omega^t) = B(f^t)D^2f^t(\omega^t)$. The eigenvalues λ_j of the Hessian D^2f satisfy $\ddot{\lambda}_j = B(f^t)\lambda_j$.

Definition. Time independent solutions of the Vlasov equation are called **equilibrium measures** or **stationary solutions.**

Definition. One can construct some of them with a **Maxwellian ansatz**

$$P(x,y) = C\exp(-\beta(\frac{y^2}{2} + \int V(x-x')Q(x')\,dx)) = S(y)Q(x)\,,$$

The constant C is chosen such that $\int_{\mathbf{R}^d} S(y)\,dy = 1$. These measures are called **Bernstein-Green-Kruskal** (BGK) modes.

Proposition 5.4.4. If $Q : N \mapsto \mathbf{R}$ satisfies the integral equation

$$Q(x) = \exp(-\int_{\mathbf{R}^d} \beta V(x-x')Q(x'))\,dx' = \exp(-\beta V \star Q(x))$$

then the Maxwellian distribution $P(x,y) = S(y)Q(x)$ is an equilibrium solution of the Vlasov equation to the potential V.

Proof.

$$\begin{aligned} y\nabla_x P &= yS(y)Q_x(x) \\ &= yS(y)(-\beta Q(x)\int_{\mathbf{R}^d} \nabla_x V(x-x')Q(x')\,dx') \end{aligned}$$

and

$$\begin{aligned} &\int_N \nabla_x V(x-x')\nabla_y P(x,y)P(x',y')\,dx'dy' \\ =\ & Q(x)(-\beta S(y)y)\int \nabla_x V(x-x')Q(x')\,dx' \end{aligned}$$

gives $y\nabla_x P(x,y) = \int_N \nabla_x V(x-x')\nabla_y P(x,y)P(x',y')\,dx'\,dy'$. □

5.5 Multidimensional distributions

Random variables which are vector-valued can be treated in an analogous way as random variables. One often adds the term "multivariate" to indicate that one has multiple dimensions.

Definition. A **random vector** is a vector-valued random variable. It is in $\mathcal{L}^p$ if each coordinate is in $\mathcal{L}^p$. The **expectation** $\mathrm{E}[X]$ of a random vector $X = (X_1, \dots, X_d)$ is the vector $(\mathrm{E}[X_1], \dots, \mathrm{E}[X_d])$, the **variance** is the vector $(\mathrm{Var}[X_1], \dots, \mathrm{Var}[X_d])$.

Example. The random vector $X = (x^3, y^4, z^5)$ on the unit cube $\Omega = [0,1]^3$ with Lebesgue measure P has the expectation $\mathrm{E}[X] = (1/4, 1/5, 1/6)$.

Definition. Assume $X = (X_1, \dots, X_d)$ is a random vector in $\mathcal{L}^\infty$. The **law of the random vector** X is a measure μ on $\mathbb{R}^d$ with compact support. After some scaling and translation we can assume that μ be a bounded Borel measure on the unit cube $I^d = [0,1]^d$.

Definition. The **multi-dimensional distribution function** of a random vector $X = (X_1, \dots, X_d)$ is defined as

$$F_X(t) = F_{(X_1,..,X_d)}(t_1, .., t_d) = \mathrm{P}[X_1 \leq t_1, \dots, X_d \leq t_d] \; .$$

For a continuous random variable, there is a density $f_X(t)$ satisfying

$$F_X(t) = \int_{-\infty}^{t_1} \dots \int_{-\infty}^{t_d} f(s_1, .., s_d) \, ds_1 \cdots ds_d \; .$$

The multi-dimensional distribution function is also called **multivariate distribution function.**

Definition. We use in this section the **multi-index notation** $x^n = \prod_{i=1}^d x_i^{n_i}$. Denote by $\mu_n = \int_{I^d} x^n \; d\mu$ the **n'th moment** of μ. If X is a random vector, with law μ, call $\mu_n(X)$ the n'th moment of X. It is equal to $\mathrm{E}[X^n] = \mathrm{E}[X_1^{n_1} X_2^{n_2} \cdots X_d^{n_d}]$. We call the map $n \in \mathbb{N}^d \mapsto \mu_n$ the **moment configuration** or, if $d = 1$, the **moment sequence**. We will tacitly assume $\mu_n = 0$, if at least one coordinate n_i in $n = (n_1, \dots, n_d)$ is negative.
If X is a continuous random vector, the moments satisfy

$$\mu_n(X) = \int_{\mathbb{R}^d} x^n f(x) \, dx$$

which is a short hand notation for

$$\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} x_1^{n_d} \cdots x_n^{n_d} f(x_1, .., x_d) \, dx_1 \cdots dx_d \; .$$

Example. The $n = (7,3,4)$'th moment of the random vector $X = (x^3, y^4, z^5)$ is

$$\mathrm{E}[X_1^{n_1} X_2^{n_2} X_3^{n_3}] = \mathrm{E}[x^{21} y^{12} z^{20}] = \frac{1}{22}\frac{1}{13}\frac{1}{20} .$$

The random vector X is continuous and has the probability density

$$f(x,y,z) = (\frac{x^{-2/3}}{3})(\frac{y^{-3/4}}{4})(\frac{z^{-4/5}}{5}) .$$

Remark. As in one dimension, one can define a **multidimensional moment generating function**

$$M_X(t) = \mathrm{E}[e^{t \cdot X}] = \mathrm{E}[e^{t_1 X_1} e^{t_2 X_2} \dots e^{t_d X_d}]$$

which contains all the information about the moments because of the **multidimensional moment formula**

$$\mathrm{E}[X^n]_j = \int_{\mathbb{R}^d} x_j^n \, d\mu = \frac{d^n M_X}{dt_j^n}(t)|_{t=0} .$$

where the n'th derivative is defined as

$$\frac{d}{dt^n} f(x) = \frac{\partial^{n_1}}{\partial x_1^{n_1}} \frac{\partial^{n_2}}{\partial x_2^{n_2}} \cdots \frac{\partial^{n_d}}{\partial x_d^{n_d}} f(x_1, \dots, x_d) .$$

Example. The random variable $X = (x, \sqrt{y}, z^{1/3})$ has the moment generating function

$$\begin{aligned} M(s,t,u) &= \int_0^1 \int_0^1 \int_0^1 e^{sx + t\sqrt{y} + uz^{1/3}} \, dxdydz \\ &= \frac{(e^s - 1)}{s} \frac{2 + 2e^t(t-1)}{t^2} \frac{-6 + 3e^u(2 - 2u + u^2)}{u^3} . \end{aligned}$$

Because the components X_1, X_2, X_3 in this example were independent random variables, the moment generating function is of the form

$$M(s)M(t)M(u) ,$$

where the factors are the one-dimensional moments of the one-dimensional random variables X_1, X_2 and X_3.

Definition. Let e_i be the standard basis in $\mathbb{Z}^d$. Define the **partial difference** $(\Delta_i a)_n = a_{n-e_i} - a_n$ on configurations and write $\Delta^k = \prod_i \Delta_i^{k_i}$. Unlike the usual convention, we take a particular sign convention for Δ. This allows us to avoid many negative signs in this section. By induction in $\sum_{i=1}^d n_i$, one proves the relation

$$(\Delta^k \mu)_n = \int_{I^d} x^{n-k} (1-x)^k \, d\mu \qquad (5.1)$$

using $x^{n-e_i-k}(1-x)^k - x^{n-k}(1-x)^k = x^{n-e_i-k}(1-x)^{k+e_i}$. To improve readability, we also use notation like $\frac{k}{n} = \prod_{i=1}^n \frac{k_i}{n_i}$ or $\left(\begin{array}{c} n \\ k \end{array}\right) = \prod_{i=1}^d \left(\begin{array}{c} n_i \\ k_i \end{array}\right)$ or $\sum_{k=0}^n = \sum_{k_1=0}^{n_1} \dots \sum_{k_d=0}^{n_d}$. We mean $n \to \infty$ in the sense that $n_i \to \infty$ for all $i = 1 \dots d$.

Definition. Given a continuous function $f : I^d \to \mathbb{R}$. For $n \in \mathbb{N}^d, n_i > 0$ we define the higher dimensional *Bernstein polynomials*

$$B_n(f)(x) = \sum_{k=0}^{n} f(\frac{n_1}{k_1}, \dots, \frac{n_d}{k_d}) \left(\begin{array}{c} n \\ k \end{array} \right) x^k (1-x)^{n-k} .$$

Lemma 5.5.1. (Multidimensional Bernstein) In the uniform topology in $C(I^d)$, we have $B_n(f) \to f$ if $n \to \infty$.

Proof. By the Weierstrass theorem, multi-dimensional polynomials are dense in $C(I^d)$ as they separate points in $C(I^d)$. It is therefore enough to prove the claim for $f(x) = x^m = \prod_{i=1}^d x_i^{m_i}$. Because $B_n(y^m)(x)$ is the product of one dimensional Bernstein polynomials

$$B_n(y^m)(x) = \prod_{i=1}^{d} B_{n_i}(y_i^{m_i})(x_i) ,$$

the claim follows from the result corollary (2.6.2) in one dimensions. □

Remark. Hildebrandt and Schoenberg refer for the proof of lemma (5.5.1) to Bernstein's proof in one dimension. While a higher dimensional adaptation of the probabilistic proof could be done involving a stochastic process in $\mathbb{Z}^d$ with drift x_i in the i'th direction, the factorization argument is more elegant.

Theorem 5.5.2 (Hausdorff,Hildebrandt-Schoenberg). There is a bijection between signed bounded Borel measures μ on $[0,1]^d$ and configurations μ_n for which there exists a constant C such that

$$\sum_{k=0}^{n} | \left(\begin{array}{c} n \\ k \end{array} \right) (\Delta^k \mu)_n | \leq C, \; \forall n \in \mathbb{N}^d . \tag{5.2}$$

A configuration μ_n belongs to a positive measure if and only if additionally to (5.2) one has $(\Delta^k \mu)_n \geq 0$ for all $k, n \in \mathbb{N}^d$.

Proof. (i) Because by lemma (5.5.1), polynomials are dense in $C(I^d)$, there exists a unique solution to the moment problem. We show now existence of a measure μ under condition (5.2). For a measures μ, define for $n \in \mathbb{N}^d$

the atomic measures $\mu^{(n)}$ on I^d which have weights $\left(\begin{array}{c} n \\ k \end{array}\right) (\Delta^k\mu)_n$ on the $\prod_{i=1}^d (n_i+1)$ points $(\frac{n_1-k_1}{n_1}, \dots, \frac{n_d-k_d}{n_d}) \in I^d$ with $0 \leq k_i \leq n_i$. Because

$$\begin{aligned}
\int_{I^d} x^m d\mu^{(n)}(x) &= \sum_{k=0}^n \left(\begin{array}{c} n \\ k \end{array}\right) (\frac{n-k}{n})^m (\Delta^k\mu)_n \\
&= \int_{I^d} \sum_{k=0}^n \left(\begin{array}{c} n \\ k \end{array}\right) (\frac{n-k}{n})^m x^{n-k}(1-x)^k \, d\mu(x) \\
&= \int_{I^d} \sum_{k=0}^n \left(\begin{array}{c} n \\ k \end{array}\right) (\frac{k}{n})^m x^k (1-x)^{n-k} \, d\mu(x) \\
&= \int_{I^d} B_n(y^m)(x) \, d\mu(x) \to \int_0^1 x^m \, d\mu(x) \; ,
\end{aligned}$$

we know that any signed measure μ which is an accumulation point of $\mu^{(n)}$, where $n_i \to \infty$ solves the moment problem. The condition (5.2) implies that the variation of the measures $\mu^{(n)}$ is bounded. By Alaoglu's theorem, there exists an accumulation point μ.

(ii) The left hand side of (5.2) is the variation $||\mu^{(n)}||$ of the measure $\mu^{(n)}$. Because by (i) $\mu^{(n)} \to \mu$, and μ has finite variation, there exists a constant C such that $||\mu^{(n)}|| \leq C$ for all n. This establishes (5.2).

(iii) We see that if $(\Delta^k\mu)_n \geq 0$ for all k, then the measures $\mu^{(n)}$ are all positive and therefore also the measure μ.

(iv) If μ is a positive measure, then by (5.1)

$$\left(\begin{array}{c} n \\ k \end{array}\right) (\Delta^k\mu)_n = \left(\begin{array}{c} n \\ k \end{array}\right) \int_{I^d} x^{n-k}(1-x)^k \, d\mu(x) \geq 0 \, .$$

□

Remark. Hildebrandt and Schoenberg noted in 1933, that this result gives a **constructive proof** of the Riesz representation theorem stating that the dual of $C(I^d)$ is the space of Borel measures $M(I^d)$.

Definition. Let $\delta(x)$ denote the **Dirac point measure** located on $x \in I^d$. It satisfies $\int_{I^d} \delta(x) \, dy = x$.

We extract from the proof of theorem (5.5.2) the construction:

Corollary 5.5.3. An explicit finite constructive approximations of a given measure μ on I^d is given for $n \in \mathbb{N}^d$ by the atomic measures

$$\mu^{(n)} = \sum_{0 \leq k_i \leq n_i} \left(\begin{array}{c} n \\ k \end{array}\right) (\Delta^k\mu)_n \delta((\frac{n_1-k_1}{n_1}, \dots, \frac{n_d-k_d}{n_d})) \, .$$

Hausdorff established a criterion for absolutely continuity of a measure μ with respect to the Lebesgue measure on $[0, 1]$ [72]. This can be generalized to find a criterion for comparing two arbitrary measures and works in d dimensions.

Definition. As usual, we call a measure μ on I^d *uniformly absolutely continuous* with respect to ν, if it satisfies $\mu = f\, d\nu$ with $f \in L^\infty(I^d)$.

Corollary 5.5.4. A positive probability measure μ is uniformly absolutely continuous with respect to a second probability measure ν if and only if there exists a constant C such that $(\Delta^k\mu)_n \leq C \cdot (\Delta^k\nu)_n$ for all $k, n \in \mathbb{N}^d$.

Proof. If $\mu = f\nu$ with $f \in L^\infty(I^d)$, we get using (5.1)

$$\begin{aligned}
(\Delta^k\mu)_n &= \int_{I^d} x^{n-k}(1-x)^k \, d\mu(x) \\
&= \int_{I^d} x^{n-k}(1-x)^k \, f d\nu(x) \\
&\leq ||f||_\infty \int_{I^d} x^{n-k}(1-x)^k \, d\nu(x) \\
&= ||f||_\infty (\Delta^k\nu)_n .
\end{aligned}$$

On the other hand, if $(\Delta^k\mu)_n \leq C(\Delta^k\nu)_n$ then $\rho_n = C(\Delta^k\nu)_n - (\Delta^k\mu)_n$ defines by theorem (5.5.2) a positive measure ρ on I^d. Since $\rho = C\nu - \mu$, we have for any Borel set $A \subset I^d$ $\rho(A) \geq 0$. This gives $\mu(A) \leq C\nu(A)$ and implies that μ is absolutely continuous with respect to ν with a function f satisfying $f(x) \leq C$ almost everywhere. □

This leads to a higher dimensional generalization of Hausdorff's result which allows to characterize the continuity of a multidimensional random vector from its moments:

Corollary 5.5.5. A Borel probability measure μ on I^d is uniformly absolutely continuous with respect to Lebesgue measure on I^d if and only if $|\Delta^k\mu_n| \leq \binom{n}{k} \prod_{i=1}^d (n_i + 1)$ for all k and n.

Proof. Use corollary (5.5.4) and $\int_{I^d} x^n \, dx = \prod_i \binom{n_i}{k_i} \prod_i (n_i + 1)$. □

There is also a characterization of Hausdorff of L^p measures on $I^1 = [0, 1]$ for $p > 2$. This has an obvious generalization to d dimensions:

Proposition 5.5.6. Given a bounded positive probability measure $\mu \in M(I^d)$ and assume $1 < p < \infty$. Then $\mu \in L^p(I^d)$ if and only if there exists a constant C such that for all k, n

$$(n+1)^{p-1} \sum_{k=0}^{n} (\Delta^k(\mu)_n \left(\begin{array}{c} n \\ k \end{array} \right))^p \leq C . \tag{5.3}$$

Proof. (i) Let $\mu^{(n)}$ be the measures of corollary (5.5.3). We construct first from the atomic measures $\mu^{(n)}$ absolutely continuous measures $\tilde{\mu}^{(n)} = g^{(n)} dx$ on I^d given by a function g which takes the constant value

$$(|\Delta^k(\mu)_n| \left(\begin{array}{c} n \\ k \end{array} \right))^p \prod_{i=1}^{d} (n_i + 1)^p$$

on a cube of side lengths $1/(n_i + 1)$ centered at the point $(n-k)/n \in I^d$. Because the cube has Lebesgue volume $(n+1)^{-1} = \prod_{i=1}^{d} (n_i+1)^{-1}$, it has the same measure with respect to both $\tilde{\mu}^{(n)}$ and $g^{(n)} dx$. We have therefore also $g^{(n)} dx \to \mu$ weakly.

(ii) Assume $\mu = f dx$ with $f \in L^p$. Because $g^{(n)} dx \to f dx$ in the weak topology for measures, we have $g^{(n)} \to f$ weakly in L^p. But then, there exists a constant C such that $||g^{(n)}||_p \leq C$ and this is equivalent to (5.3).

(iii) On the other hand, assumption (5.3) means that $||g^{(n)}||_p \leq C$, where $g^{(n)}$ was constructed in (i). Since the unit-ball in the reflexive Banach space $L^p(I^d)$ is weakly compact for $p \in (0, 1)$, a subsequence of $g^{(n)}$ converges to a function $g \in L^p$. This implies that a subsequence of $g^{(n)} dx$ converges as a measure to $g dx$ which is in L^p and which is equal to μ by the uniqueness of the moment problem (Weierstrass). □

5.6 Poisson processes

Definition. A **Poisson process** (S, P, Π, N) over a probability space $(\Omega, \mathcal{F}, Q)$ is given by a complete metric space S, a non-atomic finite Borel measure P on S and a function $\omega \mapsto \Pi(\omega) \subset S$ from Ω to the set of finite subsets of S such that for every measurable set $B \subset S$, the map

$$\omega \to N_B(\omega) = \frac{P[S]}{|\Pi(\omega)|} |\Pi(\omega) \cap B|$$

is a Poisson distributed random variable with parameter P$[B]$. For any finite partition $\{B_i\}_{i=1}^n$ of S, the set of random variables $\{N_{B_i}\}_{i=1}^n$ have to be independent. The measure P is called the **mean measure** of the process. Here $|A|$ denotes the cardinality of a finite set A. It is understood that $N_B(\omega) = 0$ if $\omega \in S^0 = \{0\}$.

Example. We have encountered the one-dimensional Poisson process in the last chapter as a martingale. We started with IID Poisson distributed random variables X_k which are "waiting times" and defined $N_t(\omega) = \sum_{k=1}^{\infty} 1_{S_k(\omega) \leq t}$. Lets translate this into the current framework. The set S is $[0,t]$ with Lebesgue measure P as mean measure. The set $\Pi(\omega)$ is the discrete point set $\Pi(\omega) = \{S_n(\omega) \mid n = 1,2,3,\dots\} \cap S$. For every Borel set B in S, we have

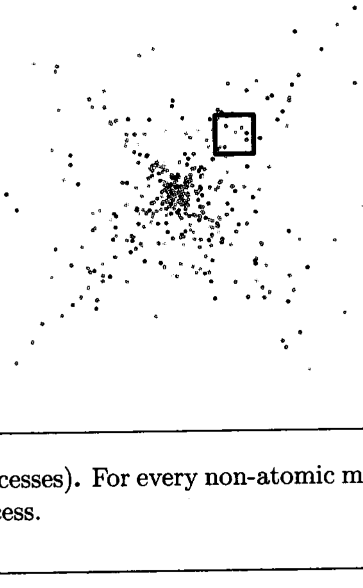
$$N_B(\omega) = t\frac{|\Pi(\omega) \cap B|}{|\Pi(\omega)|} .$$

Remark. The Poisson process is an example of a **point process**, because we can see it as assigning a random point set $\Pi(\omega)$ on S which has density P on S. If S is part of the Euclidean space and the mean measure P is continuous $\mathrm{P} = f dx$, then the interpretation is that $f(x)$ is the average density of points at x.

,

Figure. *A Poisson process in $\mathbb{R}^2$ with mean density*

$$\mathrm{P} = \frac{e^{-x^2-y^2}}{2\pi}\ dxdy .$$



Theorem 5.6.1 (Existence of Poisson processes). For every non-atomic measure P on S, there exists a Poisson process.

Proof. Define $\Omega = \bigcup_{d=0}^{\infty} S^d$, where $S^d = S \times \cdots \times S$ is the Cartesian product and $S^0 = \{0\}$. Let $\mathcal{F}$ be the Borel σ-algebra on Ω. The probability measure Q restricted to S^d is the product measure $(\mathrm{P} \times \mathrm{P} \times \cdots \times \mathrm{P}) \cdot Q[N_S = d]$, where $Q[N_S = d] = Q[S^d] = e^{-P[S]}(d!)^{-1} P[S]^d$. Define $\Pi(\omega) = \{\omega_1, \dots, \omega_d\}$ if $\omega \in S^d$ and N_B as above. One readily checks that (S, P, Π, N) is a Poisson process on the probability space $(\Omega, \mathcal{F}, Q)$: For any measurable partition $\{B_j\}_{j=0}^m$ of S, we have

$$Q[N_{B_1} = d_1, \dots, N_{B_m} = d_m \mid N_S = d_0 + \sum_{j=1}^m d_j = d] = \frac{d!}{d_0! \cdots d_m!} \prod_{j=0}^m \frac{P[B_j]^{d_j}}{P[S]^{d_j}}$$

so that the independence of $\{N_{B_j}\}_{j=1}^m$ follows:

$$\begin{aligned}
Q[N_{B_1} = d_1, \dots, N_{B_m} = d_m] &= \sum_{d=d_1+\cdots+d_m}^{\infty} Q[N_S = d]\, Q[N_{B_1} \\
&= d_1, \dots, N_{B_m} = d_m \mid N_S = d] \\
&= \sum_{d=d_1+\cdots+d_m}^{\infty} \frac{e^{-P[S]}}{d!} \frac{d!}{d_0! \cdots d_m!} \prod_{j=0}^{m} P[B_j]^{d_j} \\
&= [\sum_{d_0=0}^{\infty} \frac{e^{-P[B_0]} P[B_0]^{d_0}}{d_0!}] \prod_{j=1}^{m} \frac{e^{-P[B_j]} P[B_j]^{d_j}}{d_j!} \\
&= \prod_{j=1}^{m} \frac{e^{-P[B_j]} P[B_j]^{d_j}}{d_j!} \\
&= \prod_{j=1}^{m} Q[N_{B_j} = d_j] \, .
\end{aligned}$$

This calculation in the case $m = 1$, leaving away the last step shows that N_B is Poisson distributed with parameter $P[B]$. The last step in the calculation is then justified. □

Remark. The random discrete measure $P(\omega)[B] = N_B(\omega)$ is a normalized counting measure on S with support on $\Pi(\omega)$. The expectation of the random measure $P(\omega)$ is the measure $\tilde{P}$ on S defined by $\tilde{P}[B] = \int_\Omega P(\omega)[B]\, dQ(\omega)$. But this measure is just P:

Lemma 5.6.2. $\mathrm{P} = \int_\Omega P(\omega)\, dQ(\omega) = \tilde{P}$.

Proof. Because the Poisson distributed random variable $N_B(\omega) = P(\omega)[B]$ has by assumption the Q-expectation $P[B] = \sum_{k=0}^{\infty} k\ Q[N_B = k] = \int_\Omega P(\omega)[B]\, dQ(\omega)$ one gets $P = \int_\Omega P(\omega)\, dQ(\omega) = \tilde{P}$. □

Remark. The existence of Poisson processes can also be established by assigning to a basis $\{e_i\}$ of the Hilbert space $L^2(S, \mathrm{P})$ some independent Poisson-distributed random variables $Z_i = \phi(e_i)$ and define then a map $\phi(f) = \sum_i a_i \phi(e_i)$ if $f = \sum_i a_i e_i$. The image of this map is a Hilbert space of random variables with dot product $\mathrm{Cov}[\phi(f), \phi(g)] = (f, g)$. Define $N_B = \phi(1_B)$. These random variables have the correct distribution and are uncorrelated for disjoint sets B_j.

Definition. A **point process** is a map Π a probability space $(\Omega, \mathcal{F}, Q)$ to the set of finite subsets of a probability space $(S, \mathcal{B}, \mathrm{P})$ such that $N_B(\omega) := |\omega \cap B|$ is a random variable for all measurable sets $B \in \mathcal{B}$.

Definition. Assume Π is a point process on $(S, \mathcal{B}, \mathrm{P})$. For a function $f : S \to \mathbb{R}^+$ in $L^1(S, \mathrm{P})$, define the random variable

$$\Sigma_f(\omega) = \sum_{z \in \Pi(\omega)} f(z) .$$

Example. For a Poisson process and $f = 1_B$, one gets $\Sigma_f(\omega) = N_B(\omega)$.

Definition. The **moment generating function** of Σ_f is defined as for any random variable as

$$M_{\Sigma_f}(t) = \mathrm{E}[e^{t\Sigma_f}] .$$

It is called the **characteristic functional** of the point process.

Example. For a Poisson process and $f = a1_B$, the moment generating function of $\Sigma_f(\omega) = N_B(\omega)$ is $\mathrm{E}[e^{atN_B}] = e^{P[B](1-e^{at})}$. We have computed the moment generating function of a Poisson distributed random variable in the first chapter.

Example. For a Poisson process and $f = \sum_{k=1}^n a_j 1_{B_k}$, where B_k are disjoint sets, we have the characteristic functional

$$\prod_{j=1}^n \mathrm{E}[e^{a_j t N_{B_j}}] = e^{\sum_{j=1}^n P[B_j](1-e^{a_j t})} .$$

Example. For a Poisson process, and $f \in L^1(S, \mathrm{P})$, the moment generating function of Σ_f is

$$M_{\Sigma_f}(t) = \exp(-\int_S (1 - \exp(tf(z)))\, dP(z)) .$$

This is called **Campbell's theorem**. The proof is done by writing $f = f^+ - f^-$, where both f^+ and f^- are nonnegative, then approximating both functions with step functions $f_k^+ = \sum_j a_j^+ 1_{B_j^+} = \sum_j f_{kj}^+$ and $f_k^- = \sum_j a_j^- 1_{B_j^-} \sum_j f_{kj}^-$. Because for Poisson process, the random variables $\Sigma_{f_{kj}^\pm}$ are independent for different j or different sign, the moment generating function of Σ_f is the product of the moment generating functions $\Sigma_{f_{kj}^\pm} = N_{B_j}^\pm$.

The next theorem of Alfréd Rényi (1921-1970) gives a handy tool to check whether a **point process**, a random variable Π with values in the set of finite subsets of S, defines a Poisson process.

Definition. A **k-cube** in an open subset S of $\mathbb{R}^d$ is is a set

$$\prod_{i=1}^d [\frac{n_i}{2^k}, \frac{(n_i+1)}{2^k}) ,$$

Theorem 5.6.3 (Rényi's theorem, 1967). Let P be a non-atomic probability measure on $(S, \mathcal{B})$ and let Π be a point process on $(\Omega, \mathcal{F}, Q)$. Assume for any finite union of k-cubes $B \subset S$, $Q[N_B = 0] = \exp(-P[B])$. Then (S, P, Π, N) is a Poisson process with mean measure P.

Proof. (i) Define $O(B) = \{\omega \in \Omega \mid N_B(\omega) = 0\} \subset \Omega$ for any measurable set B in S. By assumption, $Q[O(B)] = \exp(-P[B])$.

(ii) For m disjoint k-cubes $\{B_j\}_{j=1}^m$, the sets $O(B_j) \subset \Omega$ are independent. Proof:

$$\begin{aligned} Q[\bigcap_{j=1}^m O(B_j)] &= Q[\{N_{\bigcup_{j=1}^m B_j} = 0\}] \\ &= \exp(-P[\bigcup_{j=1}^m B_j]) \\ &= \prod_{j=1}^m \exp(-P[B_j]) \\ &= \prod_{j=1}^m Q[O(B_j)] . \end{aligned}$$

(iii) We count the number of points in an open open subset U of S using k-cubes: define for $k > 0$ the random variable $N_U^k(\omega)$ as the number k-cubes B for which $\omega \in O(B \cap U)$. These random variable $N_U^k(\omega)$ converge to $N_U(\omega)$ for $k \to \infty$, for almost all ω.

(iv) For an open set U, the random variable N_U is Poisson distributed with parameter $P[U]$. Proof: we compute its moment generating function. Because for different k-cubes, the sets $O(B_j) \subset O(U)$ are independent, the moment generating function of $N_U^k = \sum_k 1_{O(B)j)}$ is the product of the moment generating functions of $1_{O(B)j)}$:

$$\begin{aligned} \mathrm{E}[e^{tN_U^k}] &= \prod_{k-\text{cube } B} (Q[O(B)] + e^t(1 - Q[O(B)])) \\ &= \prod_{k-\text{cube } B} (\exp(-P[B]) + e^t(1 - \exp(-P[B]))) . \end{aligned}$$

Each factor of this product is positive and the monotone convergence theorem shows that the moment generating function of N_U is

$$\mathrm{E}[e^{tN_U}] = \lim_{k \to \infty} \prod_{k-\text{cube } B} (\exp(-P[B]) + e^t(1 - \exp(-P[B]))) .$$

which converges to $\exp(P[U](1-e^t))$ for $k \to \infty$ if the measure P is non-atomic.

Because the generating function determines the distribution of N_U, this assures that the random variables N_U are Poisson distributed with parameter $P[U]$.

(v) For any disjoint open sets $U_1, \dots, U_m$, the random variables $\{N_{U_j})\}_{j=1}^m$ are independent. Proof: the random variables $\{N_{U_j}^k)\}_{j=1}^m$ are independent for large enough k, because no k-cube can be in more than one of the sets U_j, The random variables $\{N_{U_j}^k)\}_{j=1}^m$ are then independent for fixed k. Letting $k \to \infty$ shows that the variables N_{U_j} are independent.

(vi) To extend (iv) and (v) from open sets to arbitrary Borel sets, one can use the characterization of a Poisson process by its moment generating function of $f \in L^1(S, \mathrm{P})$. If $f = \sum a_i 1_{U_j}$ for disjoint open sets U_j and real numbers a_j, we have seen that the characteristic functional is the characteristic functional of a Poisson process. For general $f \in L^(S, \mathrm{P})$ the characteristic functional is the one of a Poisson process by approximation and the Lebesgue dominated convergence theorem. Use $f = 1_B$ to verify that N_B is Poisson distributed and $f = \sum a_i 1_{B_j}$ with disjoint Borel sets B_j to see that $\{N_{B_j})\}_{j=1}^m$ are independent. □

5.7 Random maps

Definition. Let $(\Omega, \mathcal{A}, \mathrm{P})$ be a probability space and M be a manifold with Borel σ-algebra $\mathcal{B}$. A **random diffeomorphism** on M is a measurable map from $M \times \Omega \to M$ so that $x \mapsto f(x,\omega)$ is a diffeomorphism for all $\omega \in \Omega$. Given a $\mathcal{P}$ measure preserving transformation T on Ω, it defines a **cocycle**

$$S(x,\omega) = (f(x,\omega), T(\omega))$$

which is a map on $M \times \Omega$.

Example. If M is the circle and $f(x,c) = x + c\sin(x)$ is a circle diffeomorphism, we can iterate this map and assume, the parameter c is given by IID random variables which change in each iteration. We can model this by taking $(\Omega, \mathcal{A}, \mathrm{P}) = ([0,1]^{\mathbb{N}}, \mathcal{B}^{\mathbb{N}}, \nu^{\mathbb{N}})$ where ν is a measure on $[0,1]$ and take the shift $T(x_n) = x_{n+1}$ and to define

$$S(x,\omega) = (f(x,\omega_0), T(\omega)) .$$

Iterating this **random logistic map** is done by taking IID random variables c_n with law ν and then iterate

$$x_0, x_1 = f(x_0, c_0), x_2 = f(x_1, c_1) \dots .$$

Example. If $(\Omega, \mathcal{A}, \mathrm{P}, T)$ is an ergodic dynamical system, and $A : \Omega \to SL(d, \mathbb{R})$ is measurable map with values in the special linear group $SL(d, \mathbb{R})$ of all $d \times d$ matrices with determinant 1. With $M = \mathbb{R}^d$, the random diffeomorphism $f(x, v) = A(x)v$ is called a **matrix cocycle**. One often uses the notation

$$A^n(x) = A(T^{n-1}(x)) \cdot A(T^{n-2}(x)) \cdots A(T(x)) \cdot A(x)$$

for the n'th iterate of this random map.

Example. If M is a finite set $\{1, .., n\}$ and $P = P_{ij}$ is a Markov transition matrix, a matrix with entries $P_{ij} \geq 0$ and for which the sum of the column elements is 1 in each column. A random map for which $f(x_i, \omega) = x_j$ with probability P_{ij} is called a **finite Markov chain**.

Random diffeomorphisms are examples of Markov chains as covered in Section (3.14) of the chapter on discrete stochastic processes:

Lemma 5.7.1. a) Any random map defines transition probability functions $\mathcal{P} : M \times \mathcal{B} \to [0, 1]$:

$$\mathcal{P}(x, B) = \mathrm{P}[f(x, \omega) \in B] \, .$$

b) If $\mathcal{A}_n$ is a filtration of σ-algebras and $X_n(\omega) = T^n(\omega)$ is $\mathcal{A}_n$ adapted, then $\mathcal{P}$ is a discrete Markov process.

Proof. a) We have to check that for all x, the measure $\mathcal{P}(x, \cdot)$ is a probability measure on M. This is easily be done by checking all the axioms. We further have to verify that for all $B \in \mathcal{B}$, the map $x \to \mathcal{P}(x, B)$ is $\mathcal{B}$-measurable. This is the case because f is a diffeomorphism and so continuous and especially measurable.
b) is the definition of a discrete Markov process. □

Example. If $\Omega = (\Delta^{\mathbb{N}}, \mathcal{F}^{\mathbb{N}}, \nu^{\mathbb{N}})$ and $T(x)$ is the shift, then the random map defines a discrete Markov process.

Definition. In case, we get IID Δ-valued random variables $X_n = T^n(x)_0$. A random map $f(x, \omega)$ defines so a **IID diffeomorphism-valued random variables** $f_1(x)(\omega) = f(x, X_1(\omega)), f_2(x) = f(x, X_2(\omega))$. We will call a random diffeomorphism in this case an **IID random diffeomorphism**. If the transition probability measures are continuous, then the random diffeomorphism is called a **continuous IID random diffeomorphism**. If $f(x, \omega)$ depends smoothly on ω and the transition probability measures are smooth, then the random diffeomorphism is called a **smooth IID random diffeomorphism**. It is important to note that "continuous" and "smooth" in this definition is only with respect to the transition probabilities that Δ must have at least dimension $d \geq 1$. With respect to M, we have already assumed smoothness from the beginning.

Definition. A measure μ on M is called a **stationary measure** for the random diffeomorphism if the measure $\mu \times \mathrm{P}$ is invariant under the map S.

Remark. If the random diffeomorphism defines a Markov process, the stationary measure μ is a stationary measure of the Markov process.

Example. If every diffeomorphism $x \to f(x, \omega)$ from $\omega \in \Omega$ preserves a measure μ, then μ is a automatically a stationary measure.

Example. Let $M = \mathbb{T}^2 = \mathbb{R}^2/\mathbb{Z}^2$ denote the two-dimensional torus. It is a group with addition modulo 1 in each coordinate. Given an IID random map:

$$f_n(x) = \begin{cases} x + \alpha & \text{with probability } 1/2 \\ x + \beta & \text{with probability } 1/2 \end{cases} .$$

Each map either rotates the point by the vector $\alpha = (\alpha_1, \alpha_2)$ or by the vector $\beta = (\beta_1, \beta_2)$. The Lebesgue measure on $\mathbb{T}^2$ is invariant because it is invariant for each of the two transformations. If α and β are both rational vectors, then there are infinitely many ergodic invariant measures. For example, if $\alpha = (3/7, 2/7), \beta = (1/11, 5/11)$ then the 77 rectangles $[i/7, (i+1)/7] \times [j/11, (j+1)/11]$ are permuted by both transformations.

Definition. A stationary measure μ of a random diffeomorphism is called **ergodic**, if $\mu \times \mathrm{P}$ is an ergodic invariant measure for the map S on $(M \times \Omega, \mu \times \mathrm{P})$.

Remark. If μ is a stationary invariant measure, one has

$$\mu(A) = \int_M P(x, A) \, d\mu$$

for every Borel set $A \in \mathcal{A}$. We have earlier written this as a fixed point equation for the Markov operator $\mathcal{P}$ acting on measures: $\mathcal{P}\mu = \mu$. In the context of random maps, the Markov operator is also called a **transfer operator**.

Remark. Ergodicity especially means that the transformation T on the "base probability space" $(\Omega, \mathcal{A}, \mathrm{P})$ is ergodic.

Definition. The **support** of a measure μ is the complement of the open set of points x for which there is a neighborhood U with $\mu(U) = 0$. It is by definition a closed set.

The previous example 2) shows that there can be infinitely many ergodic invariant measures of a random diffeomorphism. But for smooth IID random diffeomorphisms, one has only finitely many, if the manifold is compact:

Theorem 5.7.2 (Finitely many ergodic stationary measures (Doob)). If M is compact, a smooth IID random diffeomorphism has finitely many ergodic stationary measures μ_i. Their supports are mutually disjoint and separated by open sets.

Proof. (i) Let μ_1 and μ_2 be two ergodic invariant measures. Denote by Σ_1 and Σ_2 their support. Assume Σ_1 and Σ_2 are not disjoint. Then there exist points $x_i \in \Sigma_i$ and open sets U_i of x_i so that the transition probability $P(x_1, U_2)$ is positive. This uses the assumption that the transition probabilities have smooth densities. But then $\mu_2(U \times \Omega) = 0$ and $\mu_2(S(U \times \Omega)) > 0$ violating the measure preserving property of S.

(ii) Assume there are infinitely many ergodic invariant measures, there exist at least countably many. We can enumerate them as $\mu_1, \mu_2, ...$ Denote by Σ_i their supports. Choose a point y_i in Σ_i. The sequence of points has an accumulation point $y \in M$ by compactness of M. This implies that an arbitrary ϵ-neighborhood U of y intersects with infinitely many Σ_i. Again, the smoothness assumption of the transition probabilities $P(y, \cdot)$ contradicts with the S invariance of the measures μ_i having supports Σ_i. □

Remark. If μ_1, μ_2 are stationary probability measures, then $\lambda\mu_1 + (1-\lambda)\mu_2$ is an other stationary probability measure. This theorem implies that the set of stationary probability measures forms a closed convex simplex with finitely many corners. It is an example of a **Choquet simplex**.

5.8 Circular random variables

Definition. A measurable function from a probability space $(\Omega, \mathcal{A}, \mathrm{P})$ to the circle $(\mathbb{T}, \mathcal{B})$ with Borel σ-algebra $\mathcal{B}$ is is called a **circle-valued random variable**. It is an example of a **directional random variable**. We can realize the circle as $\mathbb{T} = [-\pi, \pi)$ or $\mathbb{T} = [0, 2\pi) = \mathbb{R}/(2\pi\mathbb{Z})$.

Example. If $(\Omega, \mathcal{A}, \mathrm{P}) = (\mathbb{R}, \mathcal{A}, e^{-x^2/2}/\sqrt{2\pi}dx$, then $X(x) = x \bmod 2\pi$ is a circle-valued random variable. In general, for any real-valued random variable Y, the random variable $X(x) = X \bmod 2\pi$ is a circle-valued random variable.

Example. For a positive integer k, the first significant digit is $X(k) = 2\pi \log_{10}(k) \bmod 1$. It is a circle-valued random variable on every finite probability space $(\Omega = \{1, \ldots, n\}, \mathcal{A}, \mathrm{P}[\{k\}] = 1/n)$.

Example. A dice takes values in $0, 1, 2, 3, 4, 5$ (count $6 = 0$). We roll it two times, but instead of adding up the results X and Y, we add them up modulo 6. For example, if $X = 4$ and $Y = 3$, then $X + Y = 1$. Note that $\mathrm{E}[X + Y] = \mathrm{E}[X] \neq \mathrm{E}[X] + \mathrm{E}[Y]$. Even if X is an unfair dice and if Y is fair, then $X + Y$ is a fair dice.

Definition. The **law** of a circular random variable X is the **push-forward measure** $\mu = X^*\mathrm{P}$ on the circle $\mathbb{T}$. If the law is absolutely continuous, it has a probability density function f_X on the circle and $\mu = f_X(x)dx$. As on the real line the Lebesgue decomposition theorem (2.12.2) assures that every measure on the circle can be decomposed $\mu = \mu_{pp} + \mu_{ac} + \mu_{sc}$, where μ_{pp} is (pp), μ_{sc} is (sc) and μ_{ac} is (ac).

Example. The law of the wrapped normal distribution in the first example is a measure on the circle with a smooth density

$$f_X(x) = \sum_{k=-\infty}^{\infty} e^{-(x+2\pi k)^2/2}/\sqrt{2\pi} .$$

It is an example of a **wrapped normal distribution**.

Example. The law of the **first significant digit** random variable $X_n(k) = 2\pi \log_{10}(k) \mod 1$ defined on $\{1, \dots, n\ \}$ is a discrete measure, supported on $\{k2\pi/10 | 0 \leq k < 10\ \}$. It is an example of a **lattice distribution**.

Definition. The **entropy** of a circle-valued random variable X with probability density function f_X is defined as $H(f) = -\int_0^{2\pi} f(x) \log(f(x))\ dx$. The **relative entropy** for two densities is defined as

$$H(f|g) = \int_0^{2\pi} f(x) \log(f(x)/g(x))\ dx .$$

The Gibbs inequality lemma (2.15.1) assures that $H(f|g) \geq 0$ and that $H(f|g) = 0$, if $f = g$ almost everywhere.

Definition. The **mean direction** m and **resultant length** ρ of a circular random variable taking values in $\{|z| = 1\} \subset \mathbb{C}$ are defined as

$$\rho e^{im} = \mathrm{E}[e^{iX}] .$$

One can write $\rho = \mathrm{E}[\cos(X - m)]$. The **circular variance** is defined as $V = 1 - \rho = \mathrm{E}[1 - \cos(X - m)] = \mathrm{E}[(X - m)^2/2 - (X - m)^4/4! \dots]$. The later expansion shows the relation with the variance in the case of real-valued random variables. The circular variance is a number in $[0, 1]$. If $\rho = 0$, there is no distinguished mean direction. We define $m = 0$ just to have one in that case.

Example. If the distribution of X is located a single point x_0, then $\rho = 1, m = x_0$ and $V = 0$. If the distribution of X is the uniform distribution on the circle, then $\rho = 0, V = 1$. There is no particular mean direction in this case. For the wrapped normal distribution $m = 0, \rho = e^{-\sigma^2/2}, V = 1 - e^{-\sigma^2/2}$.

The following lemma is analogous to theorem (2.5.5):

Theorem 5.8.1 (Chebychev inequality on the circle). If X is a circular random variable with circular mean m and variance V, then

$$\mathrm{P}[|\sin((X-m)/2)| \geq \epsilon] \leq \frac{V}{2\epsilon^2} \cdot$$

Proof. We can assume without loss of generality that $m = 0$, otherwise replace X with $X - m$ which does not change the variance. We take $\mathbb{T} = [-\pi, \pi)$. We use the trigonometric identity $1 - \cos(x) = 2\sin^2(x/2)$, to get

$$\begin{aligned} V &= \mathrm{E}[1 - \cos(X)] = 2\mathrm{E}[\sin^2(\frac{X}{2})] \\ &\geq 2\mathrm{E}[1_{|\sin(\frac{X}{2})| \geq \epsilon} \sin(\frac{X}{2})] \\ &\geq 2\epsilon^2 \mathrm{P}[|\sin(\frac{X}{2})| \geq \epsilon] \,. \end{aligned}$$

□

Example. Let X be the random variable which has a discrete distribution with a law supported on the two points $x = x_0 = 0$ and $x = x_\pm = \pm 2\arcsin(\epsilon)$ and $\mathrm{P}[X = x_0] = 1 - V/(2\epsilon^2)$ and $\mathrm{P}[X = x_\pm] = V/(4\epsilon^2)$. This distribution has the circular mean m and the variance V. The equality

$$\mathrm{P}[|\sin(X/2)| \geq \epsilon] = 2V/(4\epsilon^2) = V/(2\epsilon^2) \,.$$

shows that the Chebychev inequality on the circle is "sharp": one can not improve it without further assumptions on the distribution.

Definition. A sequence of circle-valued random variables X_n converges **weakly** to a circle-valued random variable X if the law of X_n converges weakly to the law of X. As with real valued random variables weak convergence is also called **convergence by law**.

Example. The sequence X_n of significant digit random variables X_n converges weakly to a random variable with lattice distribution $\mathrm{P}[X = k] = \log_{10}(k+1) - \log_{10}(k)$ supported on $\{k2\pi/10 \mid 0 \leq k < 10\}$. It is called the **distribution of the first significant digit**. The interpretation is that if you take a large random number, then the probability that the first digit is 1 is $\log(2)$, the probability that the first digit is 6 is $\log(7/6)$. The law is also called **Benford's law**.

Definition. The **characteristic function** of a circle-valued random variable X is the Fourier transform $\phi_X = \hat{\nu}$ of the law of X. It is a sequence (that is a function on $\mathbb{Z}$) given by

$$\phi_X(n) = \mathrm{E}[e^{inX}] = \int_{\mathbb{T}} e^{inx} \, d\nu_X(x) .$$

Definition. More generally, the **characteristic function** of a $\mathbb{T}^d$-valued random variable (circle-valued random vector) is the Fourier transform of the law of X. It is a function on $\mathbb{Z}^d$ given by

$$\phi_X(n) = \mathrm{E}[e^{in \cdot X}] = \int_{\mathbb{T}^d} e^{in \cdot x} \, d\nu_X(x) .$$

The following lemma is analog to corollary (2.17).

Lemma 5.8.2. A sequence X_n of circle-valued random variables converges in law to a circle-valued random variable X if and only if for every integer k, one has $\phi_{X_n}(k) \to \phi_X(k)$ for $n \to \infty$.

Example. A circle valued random variable with probability density function $f(x) = Ce^{\kappa \cos(x-\alpha)}$ is called the **Mises distribution**. It is also called the **circular normal distribution**. The constant C is $1/(2\pi I_0(\kappa))$, where $I_0(\kappa) = \sum_{n=0}^{\infty} (\kappa/2)^{2n}/(n!^2)$ a modified Bessel function. The parameter κ is called the **concentration parameter**, the parameter α is called the **mean direction**. For $\kappa \to 0$, the Mises distribution approaches the uniform distribution on the circle.

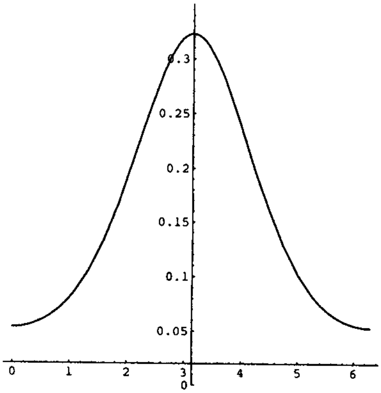


Figure. *The density function of the Mises distribution on $[-\pi, \pi]$.*

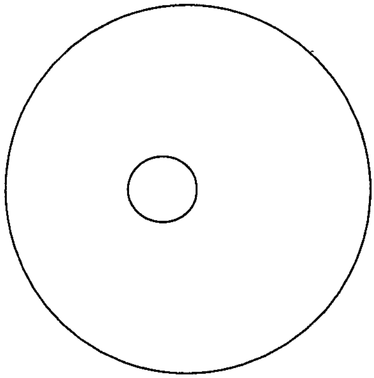


Figure. *The density function of the Mises distribution plotted as a polar graph.*

Proposition 5.8.3. The Mises distribution maximizes the entropy among all circular distributions with fixed mean α and circular variance V.

Proof. If g is the density of the Mises distribution, then $\log(g) = \kappa\cos(x-\alpha) + \log(C)$ and $H(g) = \kappa\rho + 2\pi\log(C)$.
Now compute the relative entropy

$$0 \geq H(f|g) = \int f(x)\log(f(x))dx - \int f(x)\log(g(x))dx\ .$$

This means with the resultant length ρ of f and g:

$$H(f) \geq -\mathrm{E}[\kappa\cos(x-\alpha) + \log(C)] = -\kappa\rho + 2\pi\log(C) = H(g)\ .$$

□

Definition. A circle-valued random variable with probability density function

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \sum_{k=-\infty}^{\infty} e^{-(x-\alpha-2k\pi)^2 2\sigma^2}$$

is the **wrapped normal distribution**. It is obtained by taking the normal distribution and wrapping it around the circle: if X is a normal distribution with mean α and variance σ^2, then X mod 1 is the wrapped normal distribution with those parameters.

Example. A circle-valued random variable with constant density is called a random variable with the **uniform distribution**.

Example. A circle-valued random variable with values in a closed finite subgroup H of the circle is called a **lattice distribution**. For example, the random variable which takes the value 0 with probability 1/2, the value $2\pi/3$ with probability 1/4 and the value $4\pi/3$ with probability 1/4 is an example of a lattice distribution. The group H is the finite cyclic group $\mathbb{Z}_3$.

Remark. Why do we bother with new terminology and not just look at real-valued random variables taking values in $[0, 2\pi)$? The reason to change the language is that there is a natural addition of angles given by rotations. Also, any modeling by vector-valued random variables is kind of arbitrary. An advantage is also that the characteristic function is now a sequence and no more a function.

Distribution	Parameter	characteristic function
point	x_0	$\phi_X(k) = e^{ikx_0}$
uniform		$\phi_X(k) = 0$ for $k \neq 0$ and $\phi_X(0) = 1$
Mises	$\kappa, \alpha = 0$	$I_k(\kappa)/I_0(\kappa)$
wrapped normal	$\sigma, \alpha = 0$	$e^{-k^2\sigma^2/2} = \rho^{k^2}$

The functions $I_k(\kappa)$ are modified Bessel functions of the first kind of k'th order.

Definition. If $X_1, X_2, \dots$ is a sequence of circle-valued random variables, define $S_n = X_1 + \cdots + X_n$.

Theorem 5.8.4 (Central limit theorem for circle-valued random variable). The sum S_n of IID-valued circle-valued random variables X_i which do not have a lattice distribution converges in distribution to the uniform distribution.

Proof. We have $|\phi_X(k)| < 1$ for all $k \neq 0$ because if $\phi_X(k) = 1$ for some $k \neq 0$, then X has a lattice distribution. Because $\phi_{S_n}(k) = \prod_{k=1}^n \phi_{X_i}(k)$, all Fourier coefficients $\phi_{S_n}(k)$ converge to 0 for $n \to \infty$ for $k \neq 0$. □

Remark. The IID property can be weakened. The Fourier coefficients

$$\phi_{X_n}(k) = 1 - a_{nk}$$

should have the property that $\sum_{n=1}^\infty a_{nk}$ diverges, for all k, because then, $\prod_{n=1}^\infty (1 - a_{nk}) \to 0$. If X_i converges in law to a lattice distribution, then there is a subsequence, for which the central limit theorem does not hold.

Remark. Every Fourier mode goes to zero exponentially. If $\phi_X(k) \leq 1 - \delta$ for $\delta > 0$ and all $k \neq 0$, then the convergence in the central limit theorem is exponentially fast.

Remark. Naturally, the usual central limit theorem still applies if one considers a circle-valued random variable as a random variable taking values in $[-\pi, \pi]$ Because the classical central limit theorem shows that $\sum_{i=1}^n X_n/\sqrt{n}$ converges weakly to a normal distribution, $\sum_{i=1}^n X_n/\sqrt{n} \mod 1$ converges to the wrapped normal distribution. Note that such a restatement of the central limit theorem is **not** natural in the context of circular random variables because it assumes the circle to be embedded in a particular way in the real line and also because the operation of dividing by n is not natural on the circle. It uses the field structure of the cover $\mathbb{R}$.

Example. Circle-valued random variables appear as magnetic fields in mathematical physics. Assume the plane is partitioned into squares $[j, j+1) \times [k, k+1)$ called **plaquettes**. We can attach IID random variables $B_{jk} = e^{iX_{jk}}$ on each plaquette. The total magnetic field in a region G is the product of all the magnetic fields B_{jk} in the region:

$$\prod_{(j,k) \in G} B_{jk} = e^{\sum_{j,k \in G} X_{jk}} .$$

The central limit theorem assures that the total magnetic field distribution in a large region is close to a uniform distribution.

Example. Consider standard Brownian motion B_t on the real line and its graph of $\{(t, B_t) \mid t \in \mathbb{R}\}$ in the plane. The circle-valued random variables $X_n = B_n \bmod 1$ gives the distance of the graph at time $t = n$ to the next lattice point below the graph. The distribution of X_n is the wrapped normal distribution with parameter $m = 0$ and $\sigma = n$.

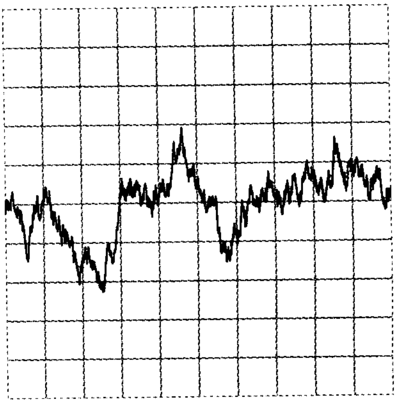


Figure. *The graph of one-dimensional Brownian motion with a grid. The stochastic process produces a circle-valued random variable* $X_n = B_n \bmod 1$.

If X, Y are real-valued IID random variables, then $X+Y$ is not independent of X. Indeed $X + Y$ and Y are positively correlated because

$$\mathrm{Cov}[X+Y, Y] = \mathrm{Cov}[X, Y] + \mathrm{Cov}[Y, Y] = \mathrm{Cov}[Y, Y] = \mathrm{Var}[Y] > 0 .$$

The situation changes for circle-valued random variables. The sum of two independent random variables can be independent to the first random variable. Adding a random variable with uniform distribution immediately renders the sum uniform:

Theorem 5.8.5 (Stability of the uniform distribution). If X, Y are circle-valued random variables. Assume that Y has the uniform distribution and that X, Y are independent, then $X + Y$ is independent of X and has the uniform distribution.

Proof. We have to show that the event $A = \{X + Y \in [c, d]\}$ is independent of the event $B = \{X \in [a, b]\}$. To do so we calculate $\mathrm{P}[A \cap B] = \int_a^b \int_{c-x}^{d-x} f_X(x) f_Y(y)\, dydx$. Because Y has the uniform distribution, we get after a substitution $u = y - x$,

$$\int_a^b \int_{c-x}^{d-x} f_X(x) f_Y(y)\, dydx = \int_a^b \int_c^d f_X(x) f_Y(u)\, dudx = \mathrm{P}[A]\mathrm{P}[B] .$$

By looking at the characteristic function $\phi_{X+Y} = \phi_X \phi_Y = \phi_X$, we see that $X + Y$ has the uniform distribution. □

The interpretation of this lemma is that adding a uniform random noise to a given uniform distribution makes it uniform.

On the n-dimensional torus $\mathbb{T}^d$, the uniform distribution plays the role of the normal distribution as the following central limit theorem shows:

Theorem 5.8.6 (Central limit theorem for circular random vectors). *The sum S_n of IID-valued circle-valued random vectors X converges in distribution to the uniform distribution on a closed subgroup H of G.*

Proof. Again $|\phi_X(k)| \leq 1$. Let Λ denote the set of k such that $\phi_X(k) = 1$.

(i) Λ is a lattice. If $\int e^{ikX(x)}\, dx = 1$ then $X(x)k = 1$ for all x. If λ, λ_2 are in Λ, then $\lambda_1 + \lambda_2 \in \Lambda$.

(ii) The random variable takes values in a group H which is the dual group of $\mathbb{Z}^d/H$.

(iii) Because $\phi_{S_n}(k) = \prod_{k=1}^n \phi_{X_i}(k)$, all Fourier coefficients $\phi_{S_n}(k)$ which are not 1 converge to 0.

(iv) $\phi_{S_n}(k) \to 1_\Lambda$, which is the characteristic function of the uniform distribution on H. □

Example. If $G = \mathbb{T}^2$ and $\Lambda = \{\dots, (-1,0), (1,0), (2,0), \dots\}$, then the random variable X takes values in $H = \{(0,y) \mid y \in \mathbb{T}^1\}$, a one dimensional circle and there is no smaller subgroup. The limiting distribution is the uniform distribution on that circle.

Remark. If X is a random variable with an absolutely continuous distribution on $\mathbb{T}^d$, then the distribution of S_n converges to the uniform distribution on $\mathbb{T}^d$.

Exercice. Let Y be a real-valued random variable which has standard normal distribution. Then $X(x) = Y(x) \mod 1$ is a circle-valued random variable. If Y_i are IID normal distributed random variables, then $S_n = Y_1 + \cdots + Y_n \mod 1$ are circle-valued random variable. What is $\text{Cov}[S_n, S_m]$?

The central limit theorem applies to all compact Abelian groups. Here is the setup:

Definition. A **topological group** G is a group with a topology so that addition on this group is a continuous map from $G \times G \to G$ and such that the inverse $x \to x^{-1}$ from G to G is continuous. If the group acts transitively as transformations on a space H, the space H is called a **homogeneous space**. In this case, H can be identified with G/G_x, where G_x is the isotopy subgroup of G consisting of all elements which fix a point x.

Example. Any finite group G with the discrete topology $d(x, y) = 1$ if $x \neq y$ and $d(x, y) = 0$ if $x = y$ is a topological group.

Example. The real line $\mathbb{R}$ with addition or more generally, the Euclidean space $\mathbb{R}^d$ with addition are topological groups when the usual Euclidean distance is the topology.

Example. The circle $\mathbb{T}$ with addition or more generally, the torus $\mathbb{T}^d$ with addition is a topological group with addition. It is an example of a **compact Abelian topological group**.

Example. The **general linear group** $G = Gl(n, \mathbb{R})$ with matrix multiplication is a topological group if the topology is the topology inherited as a subset of the Euclidean space $\mathbb{R}^{n^2}$ of $n \times n$ matrices. Also subgroup of $Gl(n, \mathbb{R})$, like the **special linear group** $SL(n, \mathbb{R})$ of matrices with determinant 1 or the **rotation group** $SO(n, R)$ of orthogonal matrices are topological groups. The rotation group has the sphere S^n as a homogeneous space.

Definition. A measurable function from a probability space $(\Omega, \mathcal{A}, \mathrm{P})$ to a topological group $(G, \mathcal{B})$ with Borel σ-algebra $\mathcal{B}$ is is called a **G-valued random variable**.

Definition. The **law** of a spherical random variable X is the **push-forward measure** $\mu = X^*\mathrm{P}$ on G.

Example. If $(G, \mathcal{A}, \mathrm{P})$ is a the probability space by taking a compact topological group G with a group invariant distance d, a Borel σ-algebra $\mathcal{A}$ and the Haar measure P, then $X(x) = x$ is a group valued random variable. The law of X is called the **uniform distribution** on G.

Definition. A measurable function from a probability space $(\Omega, \mathcal{A}, \mathrm{P})$ to the group $(G, \mathcal{B})$ is called a G-valued random variable. A measurable function to a homogeneous space is called H-valued random variable. Especially, if H is the d-dimensional sphere $(S^d, \mathcal{B})$ with Borel probability measure, then X is called a **spherical random variable**. It is used to describe **spherical data**.

5.9 Lattice points near Brownian paths

The following law of large numbers deals with sums S_n of n random variables, where the law of random variables depends on n.

Theorem 5.9.1 (Law of large numbers for random variables with shrinking support). If X_i are IID random variables with uniform distribution on $[0,1]$. Then for any $0 \leq \delta < 1$, and $A_n = [0, 1/n^\delta]$, we have

$$\lim_{n\to\infty} \frac{1}{n^{1-\delta}} \sum_{k=1}^{n} 1_{A_n}(X_k) \to 1$$

in probability. For $\delta < 1/2$, we have almost everywhere convergence.

Proof. For fixed n, the random variables $Z_k(x) = 1_{[0,1/n^\delta]}(X_k)$ are independent, identically distributed random variables with mean $\mathrm{E}[Z_k] = p = 1/n^\delta$ and variance $p(1-p)$. The sum $S_n = \sum_{k=1}^n X_k$ has a binomial distribution with mean $np = n^{1-\delta}$ and variance $\mathrm{Var}[S_n] = np(1-p) = n^{1-\delta}(1-p)$. Note that if n changes, then the random variables in the sum S_n change too, so that we can not invoke the law of large numbers directly. But the tools for the proof of the law of large numbers still work.

For fixed $\epsilon > 0$ and n, the set

$$B_n = \{x \in [0,1] \mid |\frac{S_n(x)}{n^{1-\delta}} - 1| > \epsilon\}$$

has by the Chebychev inequality (2.5.5), the measure

$$\mathrm{P}[B_n] \leq \mathrm{Var}[\frac{S_n}{n^{1-\delta}}]/\epsilon^2 = \frac{\mathrm{Var}[S_n]}{n^{2-2\delta}\epsilon^2} = \frac{1-p}{\epsilon^2 n^{1-\delta}} \leq \frac{1}{\epsilon^2 n^{1-\delta}} .$$

This proves convergence in probability and the weak law version for all $\delta < 1$ follows.

In order to apply the Borel-Cantelli lemma (2.2.2), we need to take a subsequence so that $\sum_{k=1}^\infty \mathrm{P}[B_{n_k}]$ converges. Like this, we establish complete convergence which implies almost everywhere convergence.

Take $\kappa = 2$ with $\kappa(1-\delta) > 1$ and define $n_k = k^\kappa = k^2$. The event $B = \limsup_k B_{n_k}$ has measure zero. This is the event that we are in infinitely many of the sets B_{n_k}. Consequently, for large enough k, we are in none of the sets B_{n_k}: if $x \in B$, then

$$|\frac{S_{n_k}(x)}{n_k^{1-\delta}} - 1| \leq \epsilon$$

for large enough k. Therefore,

$$|\frac{S_{n_k+l}(x)}{n_k^{1-\delta}} - 1| \leq |\frac{S_{n_k}(x)}{n_k^{1-\delta}} - 1| + \frac{S_l(T_k^n(x))}{n_k^{1-\delta}} .$$

Because for $n_k = k^2$ we have $n_{k+1} - n_k = 2k + 1$ and

$$\frac{S_l(T_k^n(x))}{n_k^{1-\delta}} \leq \frac{2k+1}{k^{2(1-\delta)}} .$$

For $\delta < 1/2$, this goes to zero assuring that we have not only convergence of the sum along a subsequence S_{n_k} but for S_n (compare lemma (2.11.2)). We know now $|\frac{S_n(x)}{n^{1-\delta}} - 1| \to 0$ almost everywhere for $n \to \infty$. □

Remark. If we sum up independent random variables $Z_k = n^\delta 1_{[0,1/n^\delta]}(X_k)$ where X_k are IID random variables, the moments $\mathrm{E}[Z_k^m] = n^{(m-1)\delta}$ become infinite for $m \geq 2$. The laws of large numbers do not apply because $\mathrm{E}[Z_k^2]$ depends on n and diverges for $n \to \infty$. We also change the random variables, when taking larger sums. For example, the assumption $\sup_n \frac{1}{n} \sum_{i=1}^n \mathrm{Var}[X_i] < \infty$ does not apply.

Remark. We could not conclude the proof in the same way as in theorem (2.9.3) because $U_n = \sum_{k=1}^n Z_k$ is not monotonically increasing. For $\delta \in [1/2, 1)$ we have only proven a weak law of large numbers. It seems however that a strong law should work for all $\delta < 1$.

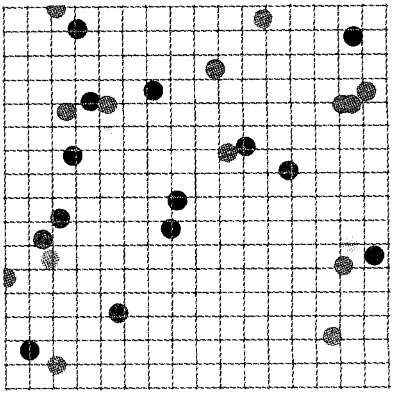
Here is an application of this theorem in **random geometry**.

Corollary 5.9.2. Assume we place randomly n discs of radius $r = 1/n^{1/2-\delta/2}$ onto the plane. Their total area without overlap is $\pi n r^2 = \pi n^\delta$. If S_n is the number of lattice points hit by the discs, then for $\delta < 1/2$

$$\frac{S_n}{n^\delta} \to \pi .$$

almost surely.

Figure. *Throwing randomly discs onto the plane and counting the number of lattice points which are hit. The size of the discs depends on the number of discs on the plane. If $\delta = 1/3$ and if $n = 1'000'000$, then we have discs of radius $1/10000$ and we expect S_n, the number of lattice point hits, to be 100π.*



Remark. Similarly as with the Buffon needle problem mentioned in the introduction, we can get a limit. But unlike the Buffon needle problem, where we keep the setup the same, independent of the number of experiments. We adapt the experiment depending on the number of tries. If we make a large number of experiments, we take a small radius of the disk. The case $\delta = 0$ is the trivial case, where the radius of the disc stays the same.

The proof of theorem (5.9.1) shows that the assumption of independence can be weakened. It is enough to have asymptotically exponentially decorrelated random variables.

Definition. A measure preserving transformation T of $[0,1]$ has **decay of correlations** for a random variable X satisfying $\mathrm{E}[X] = 0$, if

$$\mathrm{Cov}[X, X(T^n)] \to 0$$

for $n \to \infty$. If

$$\mathrm{Cov}[X, X(T^n)] \leq e^{-Cn}$$

for some constant $C > 0$, then X has **exponential decay of correlations.**

Lemma 5.9.3. If B_t is standard Brownian motion. Then the random variables $X_n = B_n \bmod 1$ have exponential decay of correlations.

Proof. B_n has the standard normal distribution with mean 0 and standard deviation $\sigma = n$. The random variable X_n is a circle-valued random variable with wrapped normal distribution with parameter $\sigma = n$. Its characteristic function is $\phi_X(k) = e^{-k^2\sigma^2/2}$. We have $X_{n+m} = X_n + Y_m \bmod 1$, where X_n and Y_m are independent circle-valued random variables. Let $g_n = \sum_{k=0}^{\infty} e^{-k^2n^2/2}\cos(kx) = 1 - \epsilon(x) \geq 1 - e^{-Cn^2}$ be the density of X_n which is also the density of Y_n. We want to know the correlation between X_{n+m} and X_n:

$$\int_0^1 \int_0^1 f(x)f(x+y)g(x)g(y)\; dy\; dx\;.$$

With $u = x + y$, this is equal to

$$\begin{aligned} & \int_0^1 \int_0^1 f(x)f(u)g(x)g(u-x)\; dudx \\ = & \int_0^1 \int_0^1 f(x)f(u)(1-\epsilon(x))(1-\epsilon(u-x))\; dudx \\ \leq & \; C_1 |f|_\infty^2 e^{-Cn^2}\;. \end{aligned}$$

□

Proposition 5.9.4. If $T : [0,1] \to [0,1]$ is a measure-preserving transformation which has exponential decay of correlations for X_j. Then for any $\delta \in [0,1/2)$, and $A_n = [0,1/n^\delta]$, we have

$$\lim_{n\to\infty} \frac{1}{n^{1-\delta}} \sum_{k=1}^{n} 1_{A_n}(T^k(x)) \to 1 \ .$$

Proof. The same proof works. The decorrelation assumption implies that there exists a constant C such that

$$\sum_{i \neq j \leq n} \mathrm{Cov}[X_i, X_j] \leq C \ .$$

Therefore,

$$\mathrm{Var}[S_n] = n\mathrm{Var}[X_n] + \sum_{i\neq j\leq n} \mathrm{Cov}[X_i, X_j] \leq C_1 |f|_\infty^2 \sum_{i,j\leq n} e^{-C(i-j)^2} \ .$$

The sum converges and so $\mathrm{Var}[S_n] = n\mathrm{Var}[X_i] + C$. □

Remark. The assumption that the probability space Ω is the interval $[0,1]$ is not crucial. Many probability spaces $(\Omega, \mathcal{A}, \mathrm{P})$ where Ω is a compact metric space with Borel σ-algebra $\mathcal{A}$ and $\mathrm{P}[\{x\}] = 0$ for all $x \in \Omega$ is measure theoretically isomorphic to $([0,1], \mathcal{B}, dx)$, where $\mathcal{B}$ is the Borel σ-algebra on $[0,1]$ (see [13] proposition (2.17). The same remark also shows that the assumption $A_n = [0,1/n^\delta]$ is not essential. One can take any nested sequence of sets $A_n \in \mathcal{A}$ with $\mathrm{P}[A_n] = 1/n^\delta$, and $A_{n+1} \subset A_n$.

Figure. *We can apply this proposition to a lattice point problem near the graphs of one-dimensional Brownian motion, where we have a probability space of paths and where we can make a statement about almost every path in that space. This is a result in the geometry of numbers for connected sets with fractal boundary.*

Corollary 5.9.5. Assume B_t is standard Brownian motion. For any $0 \leq \delta < 1/2$, there exists a constant C, such that any $1/n^{1+\delta}$ neighborhood of the graph of B over $[0,1]$ contains at least $C/n^{1-\delta}$ lattice points, if the lattice has a minimal spacing distance of $1/n$.

Proof. $B_{t+1/n}$ mod $1/n$ is not independent of B_t but the Poincaré return map T from time $t = k/n$ to time $(k+1)/n$ is a Markov process from $[0,1/n]$ to $[0,1/n]$ with transition probabilities. The random variables X_i have exponential decay of correlations as we have seen in lemma (5.9.3). □

Remark. A similar result can be shown for other dynamical systems with strong recurrence properties. It holds for example for irrational rotations with $T(x) = x + \alpha$ mod 1 with Diophantine α, while it does not hold for Liouville α. For any irrational α, we have $f_n = \frac{1}{n^{1-\delta}} \sum_{k=1}^{n} 1_{A_n}(T^k(x))$ near 1 for arbitrary large $n = q_l$, where p_l/q_l is the periodic approximation of δ. However, if the q_l are sufficiently far apart, there are arbitrary large n, where f_n is bounded away from 1 and where f_n do not converge to 1.

The theorem we have proved above belongs to the research area of **geometry of numbers**. Mixed with probability theory it is a result in the **random geometry of numbers**.

A prototype of many results in the geometry of numbers is Minkowski's theorem:

Theorem 5.9.6 (Minkowski theorem). A convex set M which is invariant under the map $T(x) = -x$ and with area > 4 contains a lattice point different from the origin.

Proof. One can translate all points of the set M back to the square $\Omega = [-1,1] \times [-1,1]$. Because the area is > 4, there are two different points $(x,y), (a,b)$ which have the same identification in the square Ω. But if $(x,y) = (u+2k, v+2l)$ then $(x-u, y-v) = (2k, 2l)$. By point symmetry also $(a,b) = (-u,-v)$ is in the set M. By convexity $((x+a)/2, (y+b)/2) = (k,l)$ is in M. This is the lattice point we were looking for. □

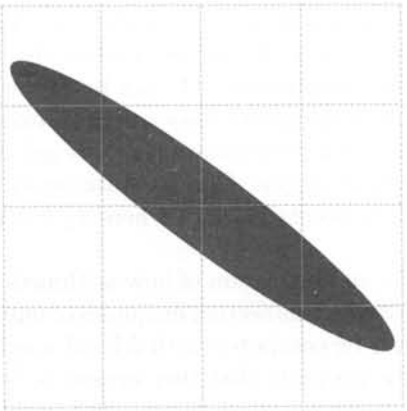


Figure. *A convex, symmetric set M. For illustration purposes, the area has been chosen smaller than 4 in this picture. The theorem of Minkowski assumes, it is larger than 4.*

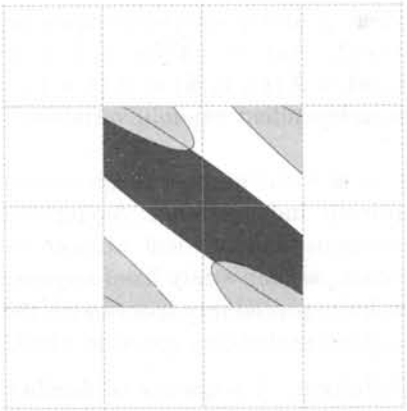


Figure. *Translate all points back to the square $[-1,1] \times [-1,1]$ of area 4. One obtains overlapping points. The symmetry and convexity allows to conclude the existence of a lattice point in M.*

There are also open questions:

- The **Gauss circle problem** asks to estimate the number of $1/n$-lattice points $g(n) = \pi n^2 + E(n)$ enclosed in the unit disk. One believes that an estimate $E(n) \leq Cn^\theta$ holds for every $\theta > 1/2$. The smallest θ for which one knows the is $\theta = 46/73$.
- For a smooth curve of length 1 which is not a line, we have a similar result as for the random walk but we need $\delta < 1/3$. Is there a result for $\delta < 1$?
- If we look at Brownian motion in $\mathbb{R}^d$. How many $1/n$ lattice points are there in a Wiener sausage, in a $1/n^{1+\delta}$ neighborhood of the path?

5.10 Arithmetic random variables

Because large numbers are virtually infinite - we have no possibility to inspect all of of the numbers from $\Omega_n = \{1, \dots n = 10^{100}\}$ for example - functions like $X_n = k^2 + 5 \bmod n$ are accessible on a small subset only. The function X_n behaves as random variable on an infinite probability space. If we could find the events $U_n = \{X_n = 0\}$ easily, then factorization would be easy as its factors can be determined from in U_n. A finite but large probability space Ω_n can be explored statistically and the question is how much information we can draw from a small number of data. It is unknown how much information can we get from a large integer n with finitely many computations. Can we statistically recover the factors of n from $O(\log(n))$ data points (k_j, x_j), where $x_j = n \bmod k_j$ for example?

As an illustration of how arithmetic complexity meets randomness, we consider in this section examples of number theoretical random variables, which can be computed with a fixed number of arithmetic operations. Both have the property that they appear to be "random" for large n. These functions belong to a class of random variables

$$X(k) = p(k, n) \bmod q(k, n) ,$$

where p and q are polynomials in two variables. For these functions, the sets $X^{-1}(a) = \{X(k) = a\}$ are in general difficult to compute and $Y_0(k) = X(k), Y_1(k) = X(k+1), \dots, Y_l(k) = X(k+l)$ behave very much as independent random variables.

To deal with "number theoretical randomness", we use the notion of **asymptotically independence**. Asymptotically independent random variables approximate independent random variables in the limit $n \to \infty$. With this notion, we can study fixed sequences or deterministic arithmetic functions on finite probability spaces with the language of probability, even so there is no fixed probability space on which the sequences form a stochastic process.

Definition. A **sequence of number theoretical random variables** is a collection of integer valued random variables X_n defined on finite probability spaces $(\Omega_n, \mathcal{A}_n, \mathrm{P}_n)$ for which $\Omega_n \subset \Omega_{n+1}$ and $\mathcal{A}_n$ is the set of all subsets of Ω_n. An example is a sequence X_n of integer valued functions defined on $\Omega_n = \{0, \dots, n-1\}$. If there exists a constant C such that X_n on $\{0, \dots, n\}$ is computable with a total of less than C additions, multiplications, comparisons, greatest common divisor and modular operations, we call X a **sequence of arithmetic random variables**.

Example. For example

$$X_n(x) = \left(((x^5 - 7) \bmod 9)^3 x - x^2\right) \bmod n$$

defines a sequence of arithmetic random variables on $\Omega_n = \{0, \dots, n-1\}$.

Example. If x_n is a fixed integer sequence, then $X_n(k) = x_k$ on $\Omega_n = \{0, \dots, n-1\}$ is a sequence of number theoretical random variables. For example, the digits x_n of the decimal sequence of π defines a sequence of number theoretical random variables $X_n(k) = x_n$ for $k \leq n$. However, in the case of π, it is not known, whether this sequence is an arithmetic sequence. It would be a surprise, if one could compute x_n with a finite n-independent number of basic operations. Also other deterministic sequences like the decimal expansions of $\pi, \sqrt{2}$ or the **Möbius function** $\mu(n)$ appear "random".

Remark. Unlike for discrete time stochastic processes X_n, where all random variables X_n are defined on a fixed probability space $(\Omega, \mathcal{A}, \mathrm{P})$, an arithmetic sequence of random variables X_n uses different finite probability spaces $(\Omega_n, \mathcal{A}_n, \mathrm{P}_n)$.

Remark. Arithmetic functions are a subset of the complexity class P of functions computable in polynomial time. The class of arithmetic sequences of random variables is expected to be much smaller than the class of sequences of all number theoretical random variables. Because computing $\mathbf{gcd}(x, y)$ needs less than $C(x + y)$ basic operations, we have included it too in the definition of arithmetic random variable.

Definition. If $\lim_{n\to\infty} \mathrm{E}[X_n]$ exists, then it is called the **asymptotic expectation** of a sequence of arithmetic random variables. If $\lim_{n\to\infty} \mathrm{Var}[X_n]$ exists, it is called the **asymptotic variance**. If the law of X_n converges, the limiting law is called the **asymptotic law**.

Example. On the probability space $\Omega_n = [1, \dots, n] \times [1, \dots, n]$, consider the arithmetic random variables $X_d = 1_{S_d}$, where $S_d = \{(n, m), \mathrm{gcd}(n, m) = d\}$.

Proposition 5.10.1. The asymptotic expectation $\mathrm{P}_n[S_1] = \mathrm{E}_n[X_1]$ is $6/\pi^2$. In other words, the probability that two random integers are relatively prime is $6/\pi^2$.

Proof. Because there is a bijection ϕ between S_1 on $[1, \dots, n]^2$ and S_d on $[1, \dots, dn]^2$ realized by $\phi(j, k) \to (dj, dk)$, we have $|S_1|/n^2 = |S_d|/(d^2n^2)$. This shows that $\mathrm{E}_n[X_1]/\mathrm{E}_n[X_d] \to d^2$ has a limit $1/d^2$ for $n \to \infty$. To know $\mathrm{P}[S_1]$, we note that the sets S_d form a partition of $\mathbb{N}^2$ and also when restricted to Ω_n. Because $\mathrm{P}[S_d] = P[S_1]/d^2$, one has

$$\mathrm{P}[S_1] \cdot (\frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} + \dots) = \mathrm{P}[S_1]\frac{\pi^2}{6} = 1 ,$$

so that $\mathrm{P}[S_1] = 6/\pi^2$. □

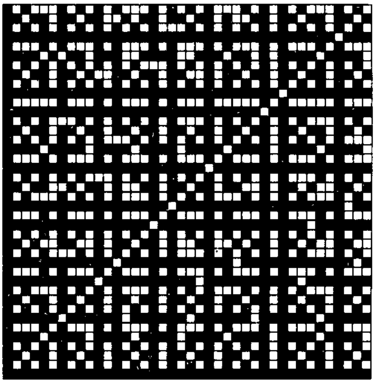


Figure. *The probability that two random integers are relatively prime is* $6/\pi^2$. *A cell* (j,k) *in the finite probability space* $[1,\dots,n] \times [1,\dots,n]$ *is painted black if* $\gcd(j,k) = 1$. *The probability that* $\gcd(j,k) = 1$ *is* $6/\pi^2 = 0.607927\dots$ *in the limit* $n \to \infty$. *So, if you pick two large numbers* (j,k) *at random, the change to have no common divisor is slightly larger than to have a common divisor.*

Exercice. Show that the asymptotic expectation of the arithmetic random variable $X_n(x,y) = \gcd(x,y)$ on $[1,\dots,n]^2$ is infinite.

Example. A large class of arithmetic random variables is defined by

$$X_n(k) = p(n,k) \bmod q(n,k)$$

on $\Omega_n = \{0,\dots,n-1\}$ where p and q are not simultaneously linear polynomials. We will look more closely at the following two examples:

1) $X_n(k) = n^2 + c \bmod k$
2) $X_n(k) = k^2 + c \bmod n$

Definition. Two sequences X_n, Y_n of arithmetic random variables, (where X_n, Y_n are defined on the same probability spaces Ω_n), are called **uncorrelated** if $\text{Cov}[X_n, Y_n] = 0$. The are called **asymptotically uncorrelated**, if their asymptotic correlation is zero:

$$\text{Cov}[X_n, Y_n] \to 0$$

for $n \to \infty$.

Definition. Two sequences X, Y of arithmetic random variables are called **independent** if for every n, the random variables X_n, Y_n are independent. Two sequences X, Y of arithmetic random variables with values in $[0,n]$ are called **asymptotically independent**, if for all I, J, we have

$$\text{P}[\frac{X_n}{n} \in I,\ \frac{Y_n}{n} \in J] - \text{P}[\frac{X_n}{n} \in I]\ \text{P}[\frac{Y_n}{n} \in J] \to 0$$

for $n \to \infty$.

Remark. If there exist two uncorrelated sequences of arithmetic random variables U, V such that $||U_n - X_n||_{L^2(\Omega_n)} \to 0$ and $||V_n - Y_n||_{L^2(\Omega_n)} \to 0$, then X, Y are asymptotically uncorrelated. If the same is true for independent sequences U, V of arithmetic random variables, then X, Y are asymptotically independent.

Remark. If two random variables are asymptotically independent, they are asymptotically uncorrelated.

Example. Two arithmetic random variables $X_n(k) = k \bmod n$ and $Y_n(k) = ak + b \bmod n$ are not asymptotic independent. Lets look at the distribution of the random vector (X_n, Y_n) in an example:

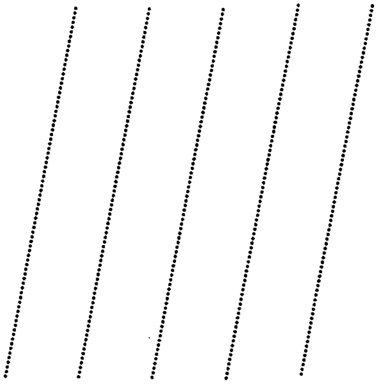


Figure. *The figure shows the points $(X_n(k), Y_n(k))$ for $X_n(k) = k, Y_n(k) = 5k + 3$ modulo n in the case $n = 2000$. There is a clear correlation between the two random variables.*

Exercice. Find the correlation of $X_n(k) = k \bmod n$ and $Y_n(k) = 5k + 3 \bmod n$.

Having asymptotic correlations between sequences of arithmetic random variables is rather exceptional. Most of the time, we observe asymptotic independence. Here are some examples:

Example. Consider the two arithmetic variables $X_n(k) = k$ and

$$Y_n(k) = ck^{-1} \bmod p(n) ,$$

where c is a constant and $p(n)$ is the n'th prime number. The random variables X_n and Y_n are asymptotically independent. Proof: by a lemma of Merel [67, 23], the number of solutions of $(x, y) \in I \times J$ of $xy = c \bmod p$ is

$$\frac{|I||J|}{p} + O(p^{1/2} \log^2(p)) .$$

This means that the probability that $X_n/n \in I_n, Y_n/n \in J_n$ is $|I_n| \cdot |J_n|$.

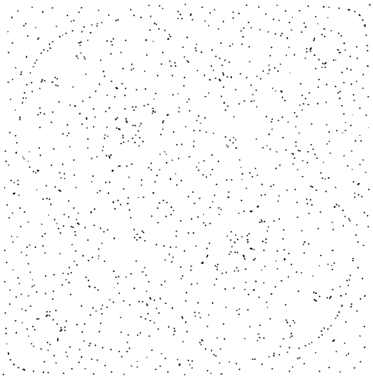


Figure. *Illustration of the lemma of Merel. The picture shows the points* $\{(k, 1/k) \bmod p\}$, *where* p *is the 200'th prime number* $p(200) = 1223$.

Nonlinear polynomial arithmetic random variables lead in general to asymptotic independence. Lets start with an experiment:

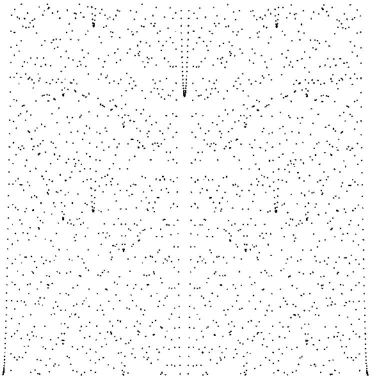


Figure. *We see the points* $(X_n(k), Y_n(k))$ *for* $X_n(k) = k, Y_n(k) = k^2 + 3$ *in the case* $n = 2001$. *Even so there are narrow regions in which some correlations are visible, these regions become smaller and smaller for* $n \to \infty$. *Indeed, we will show that* X_n, Y_n *are asymptotically independent random variables.*

The random variable $X_n(k) = (n^2 + c) \bmod k$ on $\{1, \dots, n\}$ is equivalent to $X_n(k) = n \bmod k$ on $\{0, \dots, [\sqrt{n - c}]\}$, where $[x]$ is the integer part of x. After the rescaling the sequence of random variables is easier to analyze.

To study the distribution of the arithmetic random variable X_n, we can also rescale the image, so that the range in the interval $[0, 1]$. The random variable $Y_n = X_n(x \cdot |\Omega_n|)$ can be extended from the discrete set $\{k/|\Omega_n|)\}$ to the interval $[0, 1]$. Therefore, instead of $n^2 + c \bmod k$, we look at

$$X_n(k) = \frac{n \bmod k}{k} = \frac{n}{k} - [\frac{n}{k}]$$

on $\Omega_{m(n)} = \{1, \dots, m(n)\}$, where $m(n) = \sqrt{n - c}$.

Elements in the set $X^{-1}(0)$ are the integer factors of n. Because factoring is a well studied NP problem, the multi-valued function X^{-1} is probably hard to compute in general because if we could compute it fast, we could factor integers fast.

Proposition 5.10.2. The rescaled arithmetic random variables

$$X_n(k) = \frac{n \bmod k}{k} = \frac{n}{k} - [\frac{n}{k}]$$

converge in law to the uniform distribution on $[0,1]$.

Proof. The functions $f_n^r(k) = n/(k+r) - [n/(k+r)]$ are piecewise continuous circle maps on $[0,1]$. When rescaling the argument $[0,\dots,n]$, the slope of the graph becomes larger and larger for $n \to \infty$. We can use lemma (5.10.3) below. □

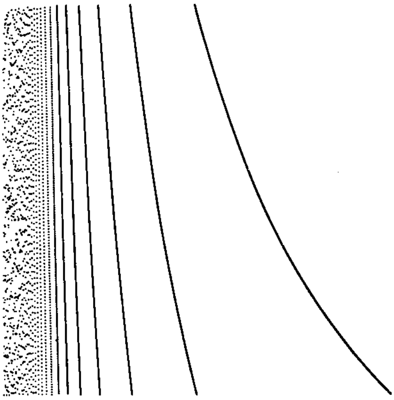


Figure. *Data points*

$$(k, \frac{n \bmod k}{k})$$

for $n = 10'000$ and $1 \leq k \leq n$. For smaller values of k, the data points appear random. The points are located on the graph of the circle map

$$f_n(t) = \frac{n}{t} - [\frac{n}{t}] .$$

To show the asymptotic independence of X_n with any of its translations, we restrict the random vectors to $[1, 1/n^a]$ with $a < 1$.

Lemma 5.10.3. Let f_n be a sequence of smooth maps from $[0,1]$ to the circle $\mathbb{T}^1 = \mathbb{R}/\mathbb{Z}$ for which $(f_n^{-1})''(x) \to 0$ uniformly on $[0,1]$, then the law μ_n of the random variables $X_n(x) = (x, f_n(x))$ converges weakly to the Lebesgue measure $\mu = dxdy$ on $[0,1] \times \mathbb{T}^1$.

Proof. Fix an interval $[a,b]$ in $[0,1]$. Because $\mu_n([a,b] \times \mathbb{T}^1)$ is the Lebesgue measure of $\{(x,y) \,|X_n(x,y) \in [a,b]\}$ which is equal to $b-a$, we only need to compare

$$\mu_n([a,b] \times [c, c+dy])$$

and

$$\mu_n([a,b] \times [d, d+dy])$$

in the limit $n \to \infty$. But $\mu_n([a,b] \times [c, c+dy]) - \mu_n([a,b] \times [c, c+dy])$ is bounded above by

$$|(f_n^{-1})'(c) - (f_n^{-1})'(d)| \leq |(f_n^{-1})''(x)|$$

which goes to zero by assumption.

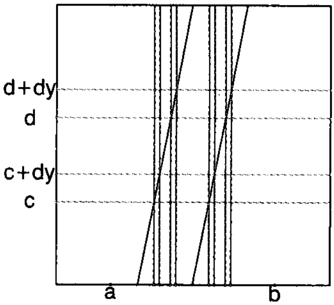


Figure. *Proof of the lemma. The measure μ_n with support on the graph of $f_n(x)$ converges to the Lebesgue measure on the product space $[0,1] \times \mathbb{T}^1$. The condition $f''/f'^2 \to 0$ assures that the distribution in the y direction smooths it out.*

Theorem 5.10.4. Let c be a fixed integer and $X_n(k) = (n^2 + c) \bmod k$ on $\{1, \dots, n\}$ For every integer $r > 0, 0 < a < 1$, the random variables $X(k), Y(k) = X(k+r)$ are asymptotically independent and uncorrelated on $[0, n^a]$.

Proof. We have to show that the discrete measures $\sum_{j=1}^{n^a} \delta(X(k), Y(k))$ converge weakly to the Lebesgue measure on the torus. To do so, we first look at the measure $\mu_n = \int_0^1 \sum_{j=1}^{n^a} \delta(X(k), Y(k))$ which is supported on the curve $t \mapsto (X(t), Y(t))$, where $k \in [0, n^a]$ with $a < 1$ converges weakly to the Lebesgue measure. When rescaled, this curve is the graph of the circle map $f_n(x) = 1/x \bmod 1$ The result follows from lemma (5.10.3). □

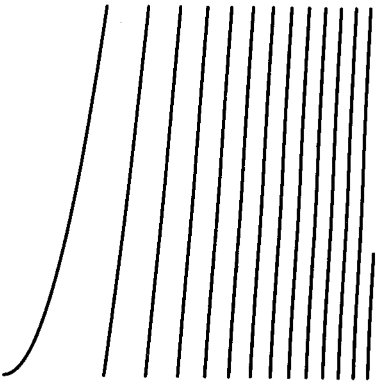
Remark. Similarly, we could show that the random vectors $(X(k), X(k + r_1), X(k+r_2), \dots, X(k+r_l))$ are asymptotically independent.

Remark. Polynomial maps like $T(x) = x^2 + c$ are used as pseudo random number generators for example in the Pollard ρ method for factorization [84]. In that case, one considers the random variables $\{0, \dots, n-1\}$ defined by $X_0(k) = k$, $X_{n+1}(k) = T(X_n(k))$. Already one polynomial map produces randomness asymptotically as $n \to \infty$.

Theorem 5.10.5. If p is a polynomial of degree $d \geq 2$, then the distribution of $Y(k) = p(k) \bmod n$ is asymptotically uniform. The random variables $X(k) = k$ and $Y(k) = p(k) \bmod n$ are asymptotically independent and uncorrelated.

Proof. The map can be extended to a map on the interval $[0, n]$. The graph $(x, T(x))$ in $\{1, \dots, n\} \times \{1, \dots, n\}$ has a large slope on most of the square. Again use lemma (5.10.3) for the circle maps $f_n(x) = p(nx) \bmod n$ on $[0, 1]$. □

Figure. *The slope of the graph of* $p(x) \bmod n$ *becomes larger and larger as* $n \to \infty$. *Choosing an integer* $k \in [0, n]$ *produces essentially a random value* $p(k) \bmod n$. *To prove the asymptotic independence, one has to verify that in the limit, the push forward of the Lebesgue measure on* $[0, n]$ *under the map* $f(x) = (x, p(x)) \bmod n$ *converges in law to the Lebesgue measure on* $[0, n]^2$.



Remark. Also here, we deal with random variables which are difficult to invert: if one could find $Y^{-1}(c)$ in $O(P(\log(n))$ times steps, then factorization would be in the complexity class P of tasks which can be computed in polynomial time. The reason is that taking square roots modulo n is at least as hard as factoring is the following: if we could find two square roots x, y of a number modulo n, then $x^2 = y^2 \bmod n$. This would lead to factor $\gcd(x - y, n)$ of n. This fact which had already been known by Fermat. If factorization was a NP complete problem, then inverting those maps would be hard.

Remark. The **Möbius function** is a function on the positive integers defined as follows: the value of $\mu(n)$ is defined as 0, if n has a factor p^2 with a prime p and is $(-1)^k$, if it contains k distinct prime factors. For example, $\mu(14) = 1$ and $\mu(18) = 0$ and $\mu(30) = -1$. The **Mertens conjecture** claimed hat

$$M(n) = |\mu(1) + \cdots + \mu(n)| \leq C\sqrt{n}$$

for some constant C. It is now believed that $M(n)/\sqrt{n}$ is unbounded but it is hard to explore this numerically, because the $\sqrt{\log\log(n)}$ bound in the law of iterated logarithm is small for the integers n we are able to compute - for example for $n = 10^{100}$, one has $\sqrt{\log\log(n)}$ is less then $8/3$. The fact

$$\frac{M(n)}{n} = \frac{1}{n}\sum_{k=1}^{n}\mu(k) \to 0$$

is known to be equivalent to the **prime number theorem**. It is also known that $\limsup M(n)/\sqrt{n} \geq 1.06$ and $\liminf M(n)/\sqrt{n} \leq -1.009$.
If one restricts the function μ to the finite probability spaces Ω_n of all numbers $\leq n$ which have no repeated prime factors, one obtains a sequence of number theoretical random variables X_n, which take values in $\{-1,1\}$. Is this sequence asymptotically independent? Is the sequence $\mu(n)$ random enough so that the law of the iterated logarithm

$$\limsup_{n\to\infty}\sum_{k=1}^{n}\frac{\mu(k)}{\sqrt{2n\log\log(n)}} \leq 1$$

holds? Nobody knows. The question is probably very hard, because if it were true, one would have

$$M(n) \leq n^{1/2+\epsilon}, \quad \text{for all } \epsilon > 0$$

which is called the **modified Mertens conjecture** . This conjecture is known to be equivalent to the **Riemann hypothesis**, the probably most notorious unsolved problem in mathematics. In any case, the connection with the Möbius functions produces a convenient way to formulate the Riemann hypothesis to non-mathematicians (see for example [14]). Actually, the question about the randomness of $\mu(n)$ appeared in classic probability text books like Fellers. Why would the law of the iterated logarithm for the Möbius function imply the Riemann hypothesis? Here is a sketch of the argument: the Euler product formula - sometimes referred to as "the Golden key" - says

$$\zeta(s) = \sum_{n=1}^{\infty}\frac{1}{n^s} = \prod_{p \text{ prime}}(1-\frac{1}{p^s})^{-1} .$$

The function $\zeta(s)$ in the above formula is called the **Riemann zeta function**. With $M(n) \leq n^{1/2+\epsilon}$, one can conclude from the formula

$$\frac{1}{\zeta(s)} = \sum_{n=1}^{\infty}\frac{\mu(n)}{n^s}$$

that $\zeta(s)$ could be extended analytically from $\text{Re}(s) > 1$ to any of the half planes $\text{Re}(s) > 1/2 + \epsilon$. This would prevent roots of $\zeta(s)$ to be to the right of the axis $\text{Re}(s) = 1/2$. By a result of Riemann, the function $\Lambda(s) = \pi^{-s/2}\Gamma(s/2)\zeta(s)$ is a meromorphic function with a simple pole at $s = 1$ and satisfies the **functional equation** $\Lambda(s) = \Lambda(1-s)$. This would imply that $\zeta(s)$ has also no nontrivial zeros to the left of the axis $\text{Re}(s) = 1/2$ and that the Riemann hypothesis were proven. The upshot is that the Riemann hypothesis could have aspects which are rooted in probability theory.

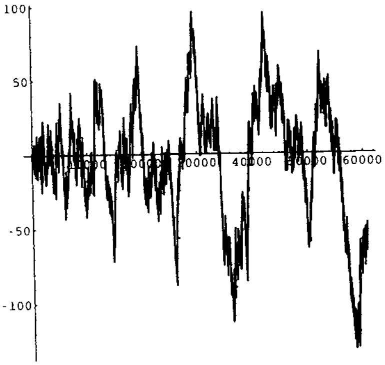


Figure. *The sequence $X_k = \mu(l(k))$, where $l(k)$ is the k nonzero entry in the sequence $\{\mu(1), \mu(2), \mu(3), \dots\}$ produces a "random walk" $S_n = \sum_{k=1}^{n} X_k$. While X_k is a deterministic sequence, the behavior of S_n resembles a typical random walk. If that were true and the law of the iterated logarithm would hold, this would imply the Riemann hypothesis.*

5.11 Symmetric Diophantine Equations

Definition. A **Diophantine equation** is an equation $f(x_1, \dots, x_k) = 0$, where p is a polynomial in k integer variables $x_1, \dots, x_k$ and where the polynomial f has integer coefficients. The Diophantine equation has **degree** m if the polynomial has degree m. The Diophantine equation is **homogeneous**, if every summand in the polynomial has the same degree. A homogeneous Diophantine equation is also called a **form**.

Example. The quadratic equation $x^2 + y^2 - z^2 = 0$ is a homogeneous Diophantine equation of degree 2. It has many solutions. They are called **Pythagorean triples**. One can parameterize them all with two parameters s, t with $x = 2st, y = s^2 - t^2$, $z = s^2 + t^2$, as has been known since antiquity already [15].

Definition. A Diophantine equation of the form

$$p(x_1, \dots, x_k) = p(y_1, \dots, y_k)$$

is called a **symmetric Diophantine equation**. More generally, a Diophantine equation

$$\sum_{i=1}^{k} x_i^m = \sum_{j=1}^{l} x_j^m$$

is called an **Euler Diophantine equation** of type (k, l) and degree m. It is a symmetric Diophantine equation if $k = l$. [28, 35, 15, 4, 5]

Remark. An Euler Diophantine equation is equivalent to a symmetric Diophantine equation if m is odd and $k + l$ is even.

Definition. A solution $(x_1, .., x_k), (y_1, \dots, y_k)$ to a symmetric Diophantine equation $p(x) = p(y)$ is called **nontrivial**, if $\{x_1, \dots, x_k \}$ and $\{y_1, \dots, y_k \}$ are different sets. For example, $5^3 + 7^3 + 3^3 = 3^3 + 7^3 + 5^3$ is a trivial solution of $p(x) = p(y)$ with $p(x, y, z) = x^3 + y^3 + z^3$.
The following theorem was proved in [68]:

Theorem 5.11.1 (Jaroslaw Wroblewski 2002). For $k > m$, the Diophantine equation $x_1^m + \dots + x_k^m = y_1^m + \dots + y_k^m$ has infinitely many nontrivial solutions.

Proof. Let R be a collection of different integer multi-sets in the finite set $[0, \dots, n]^k$. It contains at least $n^k/k!$ elements. The set $S = \{p(x) = x_1^m + \dots + x_k^m \in [0, \sqrt{k}n^{m/2}] \mid x \in R \}$ contains at least $n^k/k!$ numbers. By the **pigeon hole principle**, there are different multi-sets x, y for which $p(x) = p(y)$. This is the case if $n^k/k! > \sqrt{k}n^m$ or $n^{k-m} > k!\sqrt{k}$. □

The proof generalizes to the case, where p is an arbitrary polynomial of degree m with integer coefficients in the variables $x_1, \dots, x_k$.

Theorem 5.11.2. For an arbitrary polynomial p in k variables of degree m, the Diophantine equation $p(x) = p(y)$ has infinitely many nontrivial solutions.

Remark. Already small deviations from the symmetric case leads to local constraints: for example, $2p(x) = 2p(y) + 1$ has no solution for any nonzero polynomial p in k variables because there are no solutions modulo 2.

Remark. It has been realized by Jean-Charles Meyrignac, that the proof also gives nontrivial solutions to simultaneous equations like $p(x) = p(y) = p(z)$ etc. again by the pigeon hole principle: there are some slots, where more than 2 values hit. Hardy and Wright [28] (theorem 412) prove that in the case $k = 2, m = 3$: for every r, there are numbers which are representable as sums of two positive cubes in at least r different ways. No solutions of $x_1^4 + y_1^4 = x_2^4 + y_2^4 = x_3^4 + y_3^4$ were known to those authors [28], nor whether there are infinitely many solutions for general $(k, m) = (2, m)$. Mahler proved that $x^3 + y^3 + z^3 = 1$ has infinitely many solutions. It is believed that $x^3+y^3+z^3+w^3 = n$ has solutions for all n. For $(k, m) = (2, 3)$, multiple solutions lead to so called **taxi-cab** or **Hardy-Ramanujan numbers**.

Remark. For general polynomials, the degree and number of variables alone does not decide about the existence of nontrivial solutions of $p(x_1, \dots, x_k) = p(y_1, \dots, y_k)$. There are symmetric irreducible homogeneous equations with

$k < m/2$ for which one has a nontrivial solution. An example is $p(x, y) = x^5 - 4y^5$ which has the nontrivial solution $p(1, 3) = p(4, 5)$.

Definition. The **law** of a symmetric Diophantine equation $p(x_1, \dots, x_k) = p(x_1, \dots, x_k)$ with domain $\Omega = [0, \dots, n]^k$ is the law of the random variable defined on the finite probability space Ω.

Remark. Wroblewski's theorem holds because the random variable has an average density which is larger than the lattice spacing of the integers. So, there have to be different integers, which match. The continuum analog is that if a random variable X on a domain Ω takes values in $[a, b]$ and $b - a$ is smaller than the area of Ω, then the density f_X is larger than 1 at some point.

Remark. Wroblewski's theorem covers cases like $x^2 + y^2 + z^2 = u^2 + v^2 + w^2$ or $x^3 + y^3 + z^3 + w^3 = a^3 + b^3 + c^3 + d^3$. It is believed that for $k > m/2$, there are infinitely many solutions and no solution for $k < m/2$. [59].

Remark. For homogeneous Diophantine equations, it is enough to find a single nontrivial solution $(x_1, \dots, x_k)$ to obtain infinitely many. The reason is that $(mx_1, \dots, mx_k)$ is a solution too, for any $m \neq 0$.
Here are examples of solutions. Sources are [69, 35, 15]:

k=2,m=4 $(59, 158)^4 = (133, 134)^4$ (Euler, gave algebraic solutions in 1772 and 1778)

k=2,m=5 (open problem ([35]) all sums $\leq 1.02 \cdot 10^{26}$ have been tested)

k=3,m=5 $(3, 54, 62)^5 = (24, 28, 67)^5$ ([59], two parametric solutions by Moessner 1939, Swinnerton-Dyer)

k=3,m=6 $(3, 19, 22)^6 = (10, 15, 23)^6$ ([28],Subba Rao, Bremner and Brudno parametric solutions)

k=3,m=7 open problem?

k=4,m=7 $(10, 14, 123, 149)^7 = (15, 90, 129, 146)^7$ (Ekl)

k=4,m=8 open problem?

k=5,m=7 $(8, 13, 16, 19)^7 = (2, 12, 15, 17, 18)^7$ ([59])

k=5,m=8 $(1, 10, 11, 20, 43)^8 = (5, 28, 32, 35, 41)^8$.

k=5,m=9 $(192, 101, 91, 30, 26)^9 = (180, 175, 116, 17, 12)^9$ (Randy Ekl, 1997)

k=5,m=10 open problem

k=6,m=3 $(3, 19, 22)^6 = (10, 15, 23)^6$ (Subba Rao [59])

k=6,m=10 $(95, 71, 32, 28, 25, 16)^{10} = (92, 85, 34, 34, 23, 5)^{10}$ (Randy Ekl,1997)

k=6,m=11 open problem?

k=7,m=10 $(1, 8, 31, 32, 55, 61, 68)^{10} = (17, 20, 23, 44, 49, 64, 67)^{10}$ ([59])

k=7,m=12 $(99, 77, 74, 73, 73, 54, 30)^{12} = (95, 89, 88, 48, 42, 37, 3)^{12}$ (Greg Childers, 2000)

k=7,m=13 open problem?

k=8,m=11 $(67, 52, 51, 51, 39, 38, 35, 27)^{11} = (66, 60, 47, 36, 32, 30, 16, 7)^{11}$ (Nuutti Kuosa, 1999)

k=20,m=21 $(76, 74, 74, 64, 58, 50, 50, 48, 48, 45, 41, 32, 21, 20, 10, 9, 8, 6, 4, 4)^{21}$
$= (77, 73, 70, 70, 67, 56, 47, 46, 38, 35, 29, 28, 25, 23, 16, 14, 11, 11, 3, 3)^{21}$ (Greg Childers, 2000)

k=22,m=22 $(85, 79, 78, 72, 68, 63, 61, 61, 60, 55, 43, 42, 41, 38, 36, 34, 30, 28, 24, 12, 11, 11)^{22}$
$= (83, 82, 77, 77, 76, 71, 66, 65, 65, 58, 58, 54, 54, 51, 49, 48, 47, 26, 17, 14, 8, 6)^{22}$ (Greg Childers, 2000)

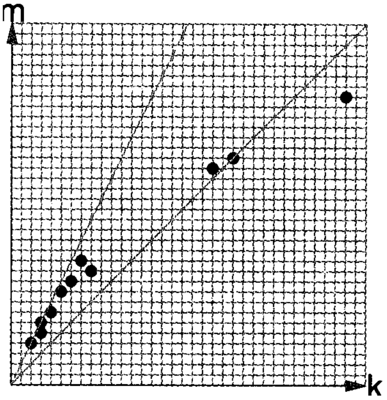


Figure. *Known cases of (k,m) with nontrivial solutions $\vec{x},\vec{y}$ of symmetric Diophantine equations $g(\vec{x}) = g(\vec{y})$ with $g(\vec{x}) = x_1^m + \cdots + x_k^m$. Wroblewski's theorem assures that for $k > m$, there are solutions. The points above the diagonal beat Wroblewski's theorem. The steep line $m = 2k$ is believed to be the threshold for the existence of nontrivial solutions. Above this line, there should be no solutions, below, there should be nontrivial solutions.*

What happens in the case $k = m$? There is no general result known. The problem has a probabilistic flavor because one can look at the distribution of random variables in the limit $n \to \infty$:

Lemma 5.11.3. Given a polynomial $p(x_1, \dots, x_k)$ with integer coefficients of degree k. The random variables

$$X_n(x_1, \dots, x_k) = p(x_1, .., x_k)/n^k$$

on the finite probability spaces $\Omega_n = [0, \dots, n]^k$ converge in law to the random variable $X(x_1, \dots, x_n) = p(x_1, .., x_k)$ on the probability space $([0,1]^k, \mathcal{B}, \mathrm{P})$, where $\mathcal{B}$ is the Borel σ-algebra and P is the Lebesgue measure.

Proof. Let $S_{a,b}(n)$ be the number of points $(x_1, \dots, x_k)$ satisfying

$$p(x_1, \dots, x_k) \in [n^k a, n^k b] .$$

This means

$$\frac{S_{a,b}(n)}{n^k} = F_n(b) - F_n(a) ,$$

where F_n is the distribution function of X_n. The result follows from the fact that $F_n(b) - F_n(a) = S_{a,b}(n)/n^k$ is a Riemann sum approximation of the integral $F(b) - F(a) = \int_{A_{a,b}} 1 \, dx$, where $A_{a,b} = \{x \in [0,1]^k \mid X(x_1, \dots, x_k) \in (a,b) \}$. □

Definition. Lets call the limiting distribution the **distribution** of the symmetric Diophantine equation. By the lemma, it is clearly a piecewise smooth function.

Example. For $k = 1$, we have $F(s) = P[X(x) \leq s] = P[x^m \leq s] = s^{1/m}/n$. The distribution for $k = 2$ for $p(x,y) = x^2 + y^2$ and $p(x,y) = x^2 - y^2$ were plotted in the first part of these notes. The distribution function of $p(x_1, x_2, \dots, x_k)$ is a $k'th$ convolution product $F_k = F \star \cdots \star F$, where $F(s) = O(s^{1/m})$ near $s = 0$. The asymptotic distribution of $p(x,y) = x^2+y^2$ is bounded for all m. The asymptotic distribution of $p(x,y) = x^2 - y^2$ is unbounded near $s = 0$ Proof. We have to understand the laws of the random variables $X(x,y) = x^2+y^2$ on $[0,1]^2$. We can see geometrically that $(\pi/4)s^2 \leq F_X(s) \leq s^2$. The density is bounded. For $Y(x,y) = x^2 - y^2$, we use polar coordinates $F(s) = \{(r,\theta) \mid r^2\cos(2\theta)/2 \leq s\}$. Integration shows that $F(s) = Cs^2 + f(s)$, where $f(s)$ grows logarithmically as $-\log(s)$. For $m > 2$, the area $x^m - y^m \leq s$ is piecewise differentiable and the derivative stays bounded.

Remark. If p is a polynomial of k variables of degree k. If the density $f = F'$ of the asymptotic distribution is unbounded, then then there are solutions to the symmetric Diophantine equation $p(x) = p(y)$.

Corollary 5.11.4. (Generalized Wroblewski) Wroblewski's result extends to polynomials p of degree k for which at least one variable appears in a term of degree smaller than k.

Proof. We can assume without loss of generality that the first variable is the one with a smaller degree m. If the variable x_1 appears only in terms of degree $k-1$ or smaller, then the polynomial p maps the finite space $[0,n]^{k/m} \times [0,n]^{k-1}$ with $n^{k+k/m-1} = n^{k+\epsilon}$ elements into the interval $[\min(p), \max(p)] \subset [-Cn^k, Cn^k]$. Apply the pigeon hole principle. □

Example. Let us illustrate this in the case $p(x,y,z,w) = x^4 + x^3 + z^4 + w^4$. Consider the finite probability space $\Omega_n = [0,n] \times [0,n] \times [0,n^{4/3}] \times [0,n]$ with $n^{4+1/3}$. The polynomial maps Ω_n to the interval $[0, 4n^4]$. The pigeon hole principle shows that there are matches.

Theorem 5.11.5. If the density f_p of the random variable p on a surface $\Omega \subset [0,n]^k$ is larger than $k!$, then there are nontrivial solutions to $p(x) = p(y)$.

In general, we try to find a subsets $\Omega \subset [0,n]^k \subset \mathbb{R}^k$ which contains $n^{k-\beta}$ points which is mapped by X into $[0, n^{m-\alpha}]$. This includes surfaces, subsets or points, where the density of X is large. To decide about this, we definitely have to know the density of X on subsets. This works often because the polynomials p modulo some integer number L do not cover all the conjugacy classes. Much of the research in this part of Diophantine equations is devoted to find such subsets and hopefully parameterize all of the solutions.

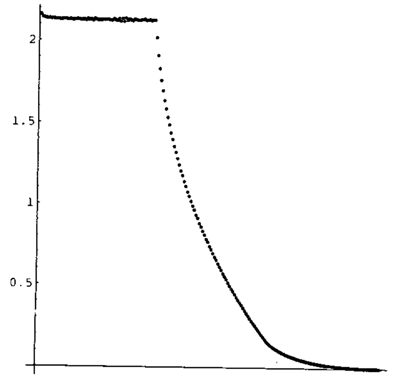


Figure. $X(x, y, z) = x^3 + y^3 + z^3$.

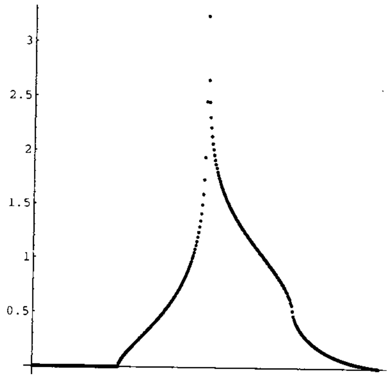


Figure. $X(x, y, z) = x^3 + y^3 - z^3$

Exercice. Show that there are infinitely many integers which can be written in non trivially different ways as $x^4 + y^4 + z^4 - w^2$.

Remark. Here is a heuristic argument for the "rule of thumb" that the **Euler Diophantine equation** $x_1^m + \cdot + x_k^m = x_0^m$ has infinitely many solutions for $k \geq m$ and no solutions if $k < m$.

For given n, the finite probability space $\Omega = \{(x_1, \dots, x_k) \mid 0 \leq x_i < n^{1/m} \}$ contains $n^{k/m}$ different vectors $x = (x_1, \dots, x_k)$. Define the random variable

$$X(x) = (x_1^m + \cdots + x_k^m)^{1/m} .$$

We expect that X takes values $1/n^{k/m} = n^{m/k}$ close to an integer for large n because $Y(x) = X(x) \bmod 1$ is expected to be uniformly distributed on the interval $[0, 1)$ as $n \to \infty$.

How close do two values $Y(x), Y(y)$ have to be, so that $Y(x) = Y(y)$? Assume $Y(x) = Y(y) + \epsilon$. Then

$$X(x)^m = X(y)^m + \epsilon X(y)^{m-1} + O(\epsilon^2)$$

with integers $X(x)^m, X(y)^m$. If $X(y)^{m-1}\epsilon < 1$, then it must be zero so that $Y(x) = Y(y)$. With the expected $\epsilon = n^{m/k}$ and $X(y)^{m-1} \leq Cn^{(m-1)/m}$ we see we should have solutions if $k > m - 1$ and none for $k < m - 1$. Cases like $m = 3, k = 2$, the Fermat Diophantine equation

$$x^3 + y^3 = z^3$$

are tagged as threshold cases by this reasoning.

This argument has still to be made rigorous by showing that the distribution of the points $f(x)$ mod 1 is uniform enough which amounts to understand a dynamical system with multidimensional time. We see nevertheless that probabilistic thinking can help to bring order into the zoo of Diophantine equations. Here are some known solutions, some written in the Lander notation

$$x^m = (x_1, \dots, x_k)^m = x_1^m + \dots + x_k^m .$$

$m = 2k = 2$: $x^2 + y^2 = z^2$ Pythagorean triples like $3^2 + 4^2 = 5^2$ (1900 BC).

$m = 3k = 2$: $x^m + y^m = z^m$ impossible, by Fermat's theorem.

$m = 3, k = 3$: $x^3 + y^3 + u^3 = v^3$ derived from taxicab numbers, like $10^3 + 9^3 = 1^3 + 12^3$ (Viete 1591).

$m = 4, k = 3$: $2682440^4 + 15365639^4 + 18796760^4 = 20615673^4$ (Elkies 1988 [24]) $m = 5, k = 3$: like $x^5 + y^5 + z^5 = w^5$ is open

$m = 4, k = 4$: $30^4 + 120^4 + 272^4 + 315^4 = 353^4$. (R. Norrie 1911 [35])

$m = 5, k = 4$ $27^5 + 84^5 + 110^5 + 133^5 = 144^5$ (Lander Parkin 1967).

$m = 6, k = 5$: $x^6 + y^6 + z^6 + u^6 + v^6 = w^6$ is open.

$m = 6, k = 6$: $(74, 234, 402, 474, 702, 894, 1077)^6 = 1141^6$.

$m = 7, k = 7$: $(525, 439, 430, 413, 266, 258, 127)^7 = 568^7$ (Mark Dodrill, 1999)

$m = 8, k = 8$: $(1324, 1190, 1088, 748, 524, 478, 223, 90)^8 = 1409^8$ (Scott Chase)

$m = 9, k = 12$, $(91, 91, 89, 71, 68, 65, 43, 42, 19, 16, 13, 5)^9 = 103^9$(Jean-Charles Meyrignac,1997)

5.12 Continuity of random variables

Let X be a random variable on a probability space $(\Omega, \mathcal{A}, \mathrm{P})$. How can we see from the characteristic function ϕ_X whether X is continuous or not? If it is continuous, how can we deduce from the characteristic function whether X is absolutely continuous or not? The first question is completely answered by Wieners theorem given below. The decision about singular or absolute continuity is more subtle. There is a necessary condition for absolute continuity:

Theorem 5.12.1 (Riemann Lebesgue-lemma). If $X \in \mathcal{L}^1$, then $\phi_X(n) \to 0$ for $|n| \to \infty$.

Proof. Given $\epsilon > 0$, choose n so large that the n'th Fourier approximation $X_n(x) = \sum_{k=-n}^{n} \phi_X(n) e^{inx}$ satisfies $||X - X_n||_1 < \epsilon$. For $m > n$, we have $\phi_m(X_n) = \mathrm{E}[e^{imX_n}] = 0$ so that

$$|\phi_X(m)| = |\phi_{X-X_n}(m)| \le ||X - X_n||_1 \le \epsilon .$$

□

Remark. The Riemann-Lebesgue lemma can not be reversed. There are random variables X for which $\phi_X(n) \to 0$, but which X is not in $\mathcal{L}^1$. Here is an example of a criterion for the characteristic function which assures that X is absolutely continuous:

Theorem 5.12.2 (Convexity). If $a_n = a_{-n}$ satisfies $a_n \to 0$ for $n \to \infty$ and $a_{n+1} - 2a_n + a_{n-1} \geq 0$, then there exists a random variable $X \in \mathcal{L}^1$ for which $\phi_X(n) = a_n$.

Proof. We follow [48].
(i) $b_n = a_n - a_{n+1}$ decreases monotonically.
Proof: the convexity condition is equivalent to $a_n - a_{n+1} \leq a_{n-1} - a_n$.
(ii) $b_n = a_n - a_{n+1}$ is non-negative for all n.
Proof: b_n decreases monotonically. If some $b_n = c < 0$, then by (i), also $b_m \leq c$ for all m contradicting the assumption that $b_n \to 0$.
(iii) Also nb_n goes to zero.
Proof: Because $\sum_{k=1}^n (a_k - a_{k+1}) = a_1 - a_{n+1}$ is bounded and the summands are positive, we must have $k(a_k - a_{k+1}) \to 0$.
(iv) $\sum_{k=1}^n k(a_{k-1} - 2a_k + a_{k+1}) \to 0$ for $n \to \infty$.
Proof. This sum simplifies to $a_0 - a_{n+1} - n(a_n - a_{n+1}$. By (iiii), it goes to 0 for $n \to \infty$.
(v) The random variable $Y(x) = \sum_{k=1}^\infty k(a_{k-1} - 2a_k + a_{k+1})K_k(x)$ is in $\mathcal{L}^1$, if $K_k(x)$ is the **Féjer kernel** with Fourier coefficients $1 - |j|/(k+1)$.
Proof. The Féjer kernel is a positive summability kernel and satisfies

$$||K_k||_1 = \frac{1}{2\pi}\int_0^{2\pi} K_k(x)\, dx = 1 \, .$$

for all k. The sum converges by (iv).
(vi) The random variables X and Y have the same characteristic functions.
Proof.

$$\begin{aligned} \phi_Y(n) &= \sum_{k=1}^\infty k(a_{k-1} - 2a_k + a_{k+1})\hat{K}_k(n) \\ &= \sum_{k=1}^\infty k(a_{k-1} - 2a_k + a_{k+1})(1 - \frac{|j|}{k+1}) \\ &= \sum_{n+1}^\infty k(a_{k-1} - 2a_k + a_{k+1})(1 - \frac{|j|}{k+1}) = a_n \, . \end{aligned}$$

□

For bounded random variables, the existence of a discrete component of the random variable X is decided by the following theorem. It will follow from corollary (5.12.5) given later on.

Theorem 5.12.3 (Wiener theorem)**.** Given $X \in \mathcal{L}^\infty$ with law μ supported in $[-\pi, \pi]$ and characteristic function $\phi = \phi_X$. Then

$$\lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} |\phi_X(k)|^2 = \sum_{x \in \mathbb{R}} \mathrm{P}[X = x]^2 .$$

Therefore, X is continuous if and only if the Wiener averages $\frac{1}{n} \sum_{k=1}^{n} |\phi_X(k)|^2$ converge to 0.

Lemma 5.12.4. If μ is a measure on the circle $\mathbb{T}$ with Fourier coefficients $\hat{\mu}_k$, then for every $x \in \mathbb{T}$, one has

$$\mu(\{x\}) = \lim_{n \to \infty} \frac{1}{2n+1} \sum_{k=-n}^{n} \hat{\mu}_k e^{ikx} .$$

Proof. We follow [48]. The **Dirichlet kernel**

$$D_n(t) = \sum_{k=-n}^{n} e^{ikt} = \frac{\sin((k+1/2)t)}{\sin(t/2)}$$

satisfies

$$D_n \star f(x) = S_n(f)(x) = \sum_{k=-n}^{n} \hat{f}(k) e^{ikx} .$$

The functions

$$f_n(t) = \frac{1}{2n+1} D_n(t-x) = \frac{1}{2n+1} \sum_{k=-n}^{n} e^{-inx} e^{int}$$

are bounded by 1 and go to zero uniformly outside any neighborhood of $t = x$. From

$$\lim_{\epsilon \to 0} \int_{x-\epsilon}^{x+\epsilon} |d(\mu - \mu(\{x\})\delta_x)| = 0$$

follows

$$\lim_{n \to \infty} \langle f_n, \mu - \mu(\{x\}) \rangle = 0$$

so that

$$\overline{\langle f_n, \mu - \mu(\{x\}) \rangle} = \frac{1}{2n+1} \sum_{k=-n}^{n} \phi(n) e^{inx} - \mu(\{x\}) \to 0 .$$

□

Theorem 5.12.7 (Strichartz). Let μ be a uniformly h-continuous measure on the circle. There exists a constant C such that for all n

$$\frac{1}{n}\sum_{k=1}^{n}|(\hat{\mu})_k|^2 \leq C\cdot h(\frac{1}{n}) .$$

Proof. The computation ([102, 103] for the Fourier transform was adapted to Fourier series in [51]). In the following computation, we abbreviate $d\mu(x)$ with dx:

$$\begin{aligned}
\frac{1}{n}\sum_{k=-n}^{n-1}|\hat{\mu}|_k^2 &\leq_1 e\int_0^1 \sum_{k=-n}^{n-1} \frac{e^{-\frac{(k+\theta)^2}{n^2}}}{n}\, d\theta\; |\hat{\mu}_k|^2 \\
&=_2 e\int_0^1 \sum_{k=-n}^{n-1} \frac{e^{-\frac{(k+\theta)^2}{n^2}}}{n} \int_{\mathbb{T}^2} e^{-i(y-x)k}\, dxdyd\theta \\
&=_3 e\int_{\mathbb{T}^2}\int_0^1 \sum_{k=-n}^{n-1} \frac{e^{-\frac{(k+\theta)^2}{n^2}-i(x-y)k}}{n}\, d\theta dxdy \\
&=_4 e\int_{\mathbb{T}^2}\int_0^1 e^{-\frac{(x-y)^2n^2}{4}+i(x-y)\theta} \\
&\qquad \sum_{k=-n}^{n-1} \frac{e^{-(\frac{k+\theta}{n}+i(x-y)\frac{n}{2})^2}}{n} d\theta dxdy
\end{aligned}$$

and continue

$$\begin{aligned}
\frac{1}{n}\sum_{k=-n}^{n-1}|\hat{\mu}|_k^2 &\leq_5 e\int_{\mathbb{T}^2} e^{-(x-y)^2\frac{n^2}{4}} \Big| \int_0^1 \\
&\qquad \sum_{k=-n}^{n-1} \frac{e^{-(i\frac{k+\theta}{n}+(x-y)\frac{n}{2})^2}}{n}\, d\theta \Big|\, dxdy \\
&=_6 e\int_{\mathbb{T}^2}[\int_{-\infty}^{\infty} \frac{e^{-(\frac{t}{n}+i(x-y)\frac{n}{2})^2}}{n}\, dt]e^{-(x-y)^2\frac{n^2}{4}}\, dxdy \\
&=_7 e\sqrt{\pi}\int_{\mathbb{T}^2}(e^{-(x-y)^2\frac{n^2}{4}})\, dxdy \\
&\leq_8 e\sqrt{\pi}(\int_{\mathbb{T}^2} e^{-(x-y)^2\frac{n^2}{2}}\, dx\, dy)^{1/2} \\
&=_9 e\sqrt{\pi}(\sum_{k=0}^{\infty}\int_{k/n\leq|x-y|\leq(k+1)/n} e^{-(x-y)^2\frac{n^2}{2}}\, dx\, dy)^{1/2} \\
&\leq_{10} e\sqrt{\pi}C_1 h(n^{-1})(\sum_{k=0}^{\infty} e^{-k^2/2})^{1/2} \\
&\leq_{11} Ch(n^{-1}) .
\end{aligned}$$

are tagged as threshold cases by this reasoning.

This argument has still to be made rigorous by showing that the distribution of the points $f(x)$ mod 1 is uniform enough which amounts to understand a dynamical system with multidimensional time. We see nevertheless that probabilistic thinking can help to bring order into the zoo of Diophantine equations. Here are some known solutions, some written in the Lander notation

$$x^m = (x_1, \dots, x_k)^m = x_1^m + \cdots + x_k^m .$$

$m = 2k = 2$: $x^2 + y^2 = z^2$ Pythagorean triples like $3^2 + 4^2 = 5^2$ (1900 BC).

$m = 3k = 2$: $x^m + y^m = z^m$ impossible, by Fermat's theorem.

$m = 3, k = 3$: $x^3 + y^3 + u^3 = v^3$ derived from taxicab numbers, like $10^3 + 9^3 = 1^3 + 12^3$ (Viete 1591).

$m = 4, k = 3$: $2682440^4 + 15365639^4 + 18796760^4 = 20615673^4$ (Elkies 1988 [24]) $m = 5, k = 3$: like $x^5 + y^5 + z^5 = w^5$ is open

$m = 4, k = 4$: $30^4 + 120^4 + 272^4 + 315^4 = 353^4$. (R. Norrie 1911 [35])

$m = 5, k = 4$ $27^5 + 84^5 + 110^5 + 133^5 = 144^5$ (Lander Parkin 1967).

$m = 6, k = 5$: $x^6 + y^6 + z^6 + u^6 + v^6 = w^6$ is open.

$m = 6, k = 6$: $(74, 234, 402, 474, 702, 894, 1077)^6 = 1141^6$.

$m = 7, k = 7$: $(525, 439, 430, 413, 266, 258, 127)^7 = 568^7$ (Mark Dodrill, 1999)

$m = 8, k = 8$: $(1324, 1190, 1088, 748, 524, 478, 223, 90)^8 = 1409^8$ (Scott Chase)

$m = 9, k = 12$, $(91, 91, 89, 71, 68, 65, 43, 42, 19, 16, 13, 5)^9 = 103^9$(Jean-Charles Meyrignac,1997)

5.12 Continuity of random variables

Let X be a random variable on a probability space $(\Omega, \mathcal{A}, \mathrm{P})$. How can we see from the characteristic function ϕ_X whether X is continuous or not? If it is continuous, how can we deduce from the characteristic function whether X is absolutely continuous or not? The first question is completely answered by Wieners theorem given below. The decision about singular or absolute continuity is more subtle. There is a necessary condition for absolute continuity:

Theorem 5.12.1 (Riemann Lebesgue-lemma). If $X \in \mathcal{L}^1$, then $\phi_X(n) \to 0$ for $|n| \to \infty$.

Proof. Given $\epsilon > 0$, choose n so large that the n'th Fourier approximation $X_n(x) = \sum_{k=-n}^{n} \phi_X(n) e^{inx}$ satisfies $||X - X_n||_1 < \epsilon$. For $m > n$, we have $\phi_m(X_n) = \mathrm{E}[e^{imX_n}] = 0$ so that

$$|\phi_X(m)| = |\phi_{X-X_n}(m)| \le ||X - X_n||_1 \le \epsilon .$$

□

Remark. The Riemann-Lebesgue lemma can not be reversed. There are random variables X for which $\phi_X(n) \to 0$, but which X is not in $\mathcal{L}^1$. Here is an example of a criterion for the characteristic function which assures that X is absolutely continuous:

Theorem 5.12.2 (Convexity). If $a_n = a_{-n}$ satisfies $a_n \to 0$ for $n \to \infty$ and $a_{n+1} - 2a_n + a_{n-1} \geq 0$, then there exists a random variable $X \in \mathcal{L}^1$ for which $\phi_X(n) = a_n$.

Proof. We follow [48].
(i) $b_n = a_n - a_{n+1}$ decreases monotonically.
Proof: the convexity condition is equivalent to $a_n - a_{n+1} \leq a_{n-1} - a_n$.
(ii) $b_n = a_n - a_{n+1}$ is non-negative for all n.
Proof: b_n decreases monotonically. If some $b_n = c < 0$, then by (i), also $b_m \leq c$ for all m contradicting the assumption that $b_n \to 0$.
(iii) Also nb_n goes to zero.
Proof: Because $\sum_{k=1}^n (a_k - a_{k+1}) = a_1 - a_{n+1}$ is bounded and the summands are positive, we must have $k(a_k - a_{k+1}) \to 0$.
(iv) $\sum_{k=1}^n k(a_{k-1} - 2a_k + a_{k+1}) \to 0$ for $n \to \infty$.
Proof. This sum simplifies to $a_0 - a_{n+1} - n(a_n - a_{n+1}$. By (iiii), it goes to 0 for $n \to \infty$.
(v) The random variable $Y(x) = \sum_{k=1}^\infty k(a_{k-1} - 2a_k + a_{k+1})K_k(x)$ is in $\mathcal{L}^1$, if $K_k(x)$ is the **Féjer kernel** with Fourier coefficients $1 - |j|/(k+1)$.
Proof. The Féjer kernel is a positive summability kernel and satisfies

$$||K_k||_1 = \frac{1}{2\pi}\int_0^{2\pi} K_k(x)\, dx = 1 \; .$$

for all k. The sum converges by (iv).
(vi) The random variables X and Y have the same characteristic functions.
Proof.

$$\begin{aligned}
\phi_Y(n) &= \sum_{k=1}^\infty k(a_{k-1} - 2a_k + a_{k+1})\hat{K}_k(n) \\
&= \sum_{k=1}^\infty k(a_{k-1} - 2a_k + a_{k+1})(1 - \frac{|j|}{k+1}) \\
&= \sum_{n+1}^\infty k(a_{k-1} - 2a_k + a_{k+1})(1 - \frac{|j|}{k+1}) = a_n \; .
\end{aligned}$$

□

For bounded random variables, the existence of a discrete component of the random variable X is decided by the following theorem. It will follow from corollary (5.12.5) given later on.

Theorem 5.12.3 (Wiener theorem). Given $X \in \mathcal{L}^\infty$ with law μ supported in $[-\pi, \pi]$ and characteristic function $\phi = \phi_X$. Then

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} |\phi_X(k)|^2 = \sum_{x \in \mathbb{R}} \mathrm{P}[X = x]^2 \, .$$

Therefore, X is continuous if and only if the Wiener averages $\frac{1}{n}\sum_{k=1}^{n} |\phi_X(k)|^2$ converge to 0.

Lemma 5.12.4. If μ is a measure on the circle $\mathbb{T}$ with Fourier coefficients $\hat{\mu}_k$, then for every $x \in \mathbb{T}$, one has

$$\mu(\{x\}) = \lim_{n\to\infty} \frac{1}{2n+1} \sum_{k=-n}^{n} \hat{\mu}_k e^{ikx} \, .$$

Proof. We follow [48]. The **Dirichlet kernel**

$$D_n(t) = \sum_{k=-n}^{n} e^{ikt} = \frac{\sin((k+1/2)t)}{\sin(t/2)}$$

satisfies

$$D_n \star f(x) = S_n(f)(x) = \sum_{k=-n}^{n} \hat{f}(k) e^{ikx} \, .$$

The functions

$$f_n(t) = \frac{1}{2n+1} D_n(t-x) = \frac{1}{2n+1} \sum_{k=-n}^{n} e^{-inx} e^{int}$$

are bounded by 1 and go to zero uniformly outside any neighborhood of $t = x$. From

$$\lim_{\epsilon\to 0} \int_{x-\epsilon}^{x+\epsilon} |d(\mu - \mu(\{x\})\delta_x)| = 0$$

follows

$$\lim_{n\to\infty} \langle f_n, \mu - \mu(\{x\}) \rangle = 0$$

so that

$$\overline{\langle f_n, \mu - \mu(\{x\}) \rangle} = \frac{1}{2n+1} \sum_{k=-n}^{n} \phi(n) e^{inx} - \mu(\{x\}) \to 0 \, .$$

□

Definition. If μ and ν are two measures on $(\Omega = \mathbb{T}, \mathcal{A})$, then its **convolution** is defined as

$$\mu \star \nu(A) = \int_{\mathbb{T}} \mu(A - x)\, d\nu(x)$$

for any $A \in \mathcal{A}$. Define for a measure on $[-\pi, \pi]$ also $\mu^*(A) = \mu(-A)$.

Remark. We have $\hat{\mu}^*(n) = \overline{\hat{\mu}(n)}$ and $\mu \hat{\star} \nu(n) = \hat{\mu}(n)\hat{\nu}(n)$. If $\mu = \sum a_j \delta_{x_j}$ is a discrete measure, then $\mu^* = \sum \overline{a_j} \delta_{-x_j}$. Because $\mu \star \mu^* = \sum_j |a_j|^2$, we have in general

$$(\mu \star \mu^*)(\{0\}) = \sum_{x \in \mathbb{T}} |\mu(\{x\})^2| \, .$$

Corollary 5.12.5. (Wiener) $\sum_{x \in \mathbb{T}} |\mu(\{x\})|^2 = \lim_{n \to \infty} \frac{1}{2n+1} \sum_{k=-n}^{n} |\hat{\mu}_n|^2$.

Remark. For bounded random variables, we can rescale the random variable so that their values is in $[-\pi, \pi]$ and so that we can use **Fourier series** instead of **Fourier integrals**. We have also

$$\sum_{x \in \mathbb{R}} |\mu(\{x\})|^2 = \lim_{R \to \infty} \frac{1}{2R} \int_{-R}^{R} |\hat{\mu}(t)|^2 \, dt \, .$$

We turn our attention now to random variables with singular continuous distribution. For these random variables, one does have $\mathrm{P}[X = c] = 0$ for all c. Furthermore, the distribution function F_X of such a random variable X does not have a density. The graph of F_X looks like a **Devil staircase**. Here is a refinement of the notion of continuity for measures.

Definition. Given a function $h : \mathbb{R} \to [0, \infty)$ satisfying $\lim_{x \to 0} h(x) = 0$. A measure μ on the real line or on the circle is called **uniformly h-continuous**, if there exists a constant C such that for all intervals $I = [a, b]$ on $\mathbb{T}$ the inequality

$$\mu(I) \leq Ch(|I|)$$

holds, where $|I| = b - a$ is the length of I. For $h(x) = x^\alpha$ with $0 < \alpha \leq 1$, the measure is called uniformly α-continuous. It is then the derivative of a α-Hölder continuous function.

Remark. If μ is the law of a singular continuous random variable X with distribution function F_X, then F_X is α-Hölder continuous if and only if μ is α-continuous. For general h, one calls F uniformly lip $-$ h continuous [86].

Theorem 5.12.6 (Y. Last). If there exists C, such that $\frac{1}{n} \sum_{k=1}^{n} |\hat{\mu}_k|^2 < C \cdot h(\frac{1}{n})$ for all $n \geq 0$, then μ is uniformly $\sqrt{h}$-continuous.

Proof. We follow [56]. The Dirichlet kernel satisfies

$$\sum_{k=-n}^{n} |\hat{\mu}_k|^2 = \int\int_{\mathbb{T}^2} D_n(y-x)\, d\mu(x)d\mu(y)$$

and the Féjer kernel $K_n(t)$ satisfies

$$\begin{aligned} K_n(t) &= \frac{1}{n+1}\left(\frac{\sin(\frac{n+1}{2}t)}{\sin(t/2)}\right)^2 \\ &= \sum_{k=-n}^{n}(1-\frac{|k|}{n+1})e^{ikt} \\ &= D_n(t) - \sum_{k=-n}^{n}\frac{|k|}{n+1}e^{ikt} . \end{aligned}$$

Therefore

$$\begin{aligned} 0 &\leq \frac{1}{n+1}\sum_{k=-n}^{n}|k||\mu_k|^2 = \int_{\mathbb{T}}\int_{\mathbb{T}}(D_n(y-x)-K_n(y-x))d\mu(x)d\mu(y) \\ &= \sum_{k=-n}^{n}|\hat{\mu}_k|^2 - \int_{\mathbb{T}}\int_{\mathbb{T}}K_n(y-x)d\mu(x)d\mu(y) . \qquad (5.4) \end{aligned}$$

Because $\hat{\mu}_n = \overline{\hat{\mu}}_{-n}$, we can also sum from $-n$ to n, changing only the constant C. If μ is not uniformly $\sqrt{h}$ continuous, there exists a sequence of intervals $|I_k| \to 0$ with $\mu(I_l) \geq l\sqrt{h(|I_l|)}$. A property of the Féjer kernel $K_n(t)$ is that for large enough n, there exists $\delta > 0$ such that $\frac{1}{n}K_n(t) \geq \delta > 0$ if $1 \leq n|t| \leq \pi/2$. Choose n_l, so that $1 \leq n_l \cdot |I_l| \leq \pi/2$. Using estimate (5.4), one gets

$$\begin{aligned} \sum_{k=-n_l}^{n_l}\frac{|\hat{\mu}_k|^2}{n_l} &\geq \int_{\mathbb{T}}\int_{\mathbb{T}}\frac{K_{n_l}(y-x)}{n_l}\, d\mu(x)d\mu(y) \\ &\geq \delta\mu(I_l)^2 \geq \delta l^2 h(|I_l|) \\ &\geq \mathrm{C}\cdot h(\frac{1}{n_l}) . \end{aligned}$$

This contradicts the existence of C such that

$$\frac{1}{n}\sum_{k=-n}^{n}|\hat{\mu}_k|^2 \leq Ch(\frac{1}{n}) .$$

□

Theorem 5.12.7 (Strichartz). Let μ be a uniformly h-continuous measure on the circle. There exists a constant C such that for all n

$$\frac{1}{n}\sum_{k=1}^{n} |(\hat{\mu})_k|^2 \leq C \cdot h(\frac{1}{n}) .$$

Proof. The computation ([102, 103] for the Fourier transform was adapted to Fourier series in [51]). In the following computation, we abbreviate $d\mu(x)$ with dx:

$$\begin{aligned}
\frac{1}{n}\sum_{k=-n}^{n-1} |\hat{\mu}|_k^2 &\leq_1 e\int_0^1 \sum_{k=-n}^{n-1} \frac{e^{-\frac{(k+\theta)^2}{n^2}}}{n}\, d\theta\, |\hat{\mu}_k|^2 \\
&=_2 e\int_0^1 \sum_{k=-n}^{n-1} \frac{e^{-\frac{(k+\theta)^2}{n^2}}}{n} \int_{\mathbb{T}^2} e^{-i(y-x)k}\, dxdyd\theta \\
&=_3 e\int_{\mathbb{T}^2}\int_0^1 \sum_{k=-n}^{n-1} \frac{e^{-\frac{(k+\theta)^2}{n^2} - i(x-y)k}}{n}\, d\theta dxdy \\
&=_4 e\int_{\mathbb{T}^2}\int_0^1 e^{-\frac{(x-y)^2 n^2}{4} + i(x-y)\theta} \\
&\quad \sum_{k=-n}^{n-1} \frac{e^{-(\frac{k+\theta}{n} + i(x-y)\frac{n}{2})^2}}{n} d\theta dxdy
\end{aligned}$$

and continue

$$\begin{aligned}
\frac{1}{n}\sum_{k=-n}^{n-1} |\hat{\mu}|_k^2 &\leq_5 e\int_{\mathbb{T}^2} e^{-(x-y)^2\frac{n^2}{4}} | \int_0^1 \\
&\quad \sum_{k=-n}^{n-1} \frac{e^{-(i\frac{k+\theta}{n} + (x-y)\frac{n}{2})^2}}{n}\, d\theta|\, dxdy \\
&=_6 e\int_{\mathbb{T}^2} [\int_{-\infty}^{\infty} \frac{e^{-(\frac{t}{n} + i(x-y)\frac{n}{2})^2}}{n}\, dt] e^{-(x-y)^2\frac{n^2}{4}}\, dxdy \\
&=_7 e\sqrt{\pi}\int_{\mathbb{T}^2} (e^{-(x-y)^2\frac{n^2}{4}})\, dxdy \\
&\leq_8 e\sqrt{\pi}(\int_{\mathbb{T}^2} e^{-(x-y)^2\frac{n^2}{2}}\, dx\, dy)^{1/2} \\
&=_9 e\sqrt{\pi}(\sum_{k=0}^{\infty}\int_{k/n \leq |x-y| \leq (k+1)/n} e^{-(x-y)^2\frac{n^2}{2}}\, dx\, dy)^{1/2} \\
&\leq_{10} e\sqrt{\pi}C_1 h(n^{-1})(\sum_{k=0}^{\infty} e^{-k^2/2})^{1/2} \\
&\leq_{11} Ch(n^{-1}) .
\end{aligned}$$

Here are some remarks about the steps done in this computation:
(1) is the trivial estimate

$$e \int_0^1 \sum_{k=-n}^{n-1} \frac{e^{-\frac{(k+\theta)^2}{n^2}}}{n} \, d\theta \geq 1$$

(2)

$$\int_{\mathbb{T}^2} e^{-i(y-x)k} \, d\mu(x) d\mu(y) = \int_{\mathbb{T}} e^{-iyk} d\mu(x) \int_{\mathbb{T}} e^{ixk} d\mu(x) = \hat{\mu}_k \overline{\hat{\mu}}_k = |\hat{\mu}_k|^2$$

(3) uses Fubini's theorem.
(4) is a completion of the square.
(5) is the Cauchy-Schwartz inequality,
(6) replaces a sum and the integral $\int_0^1$ by $\int_{-\infty}^{\infty}$,
(7) uses $\int_{-\infty}^{\infty} \frac{e^{-(\frac{t}{n}+i(x-y)\frac{n}{2})^2}}{n} \, dt = \sqrt{\pi}$ because

$$\int_{-\infty}^{\infty} \frac{e^{-(t/n+b)^2}}{n} \, dt = \sqrt{\pi}$$

for all n and complex b,
(8) is Jensen's inequality.
(9) splits the integral over a sum of small intervals of strips of width $1/n$.
(10) uses the assumption that μ is h-continuous.
(11) This step uses that

$$(\sum_{k=0}^{\infty} e^{-k^2/2})^{1/2}$$

is a constant. □

Bibliography

[1] N.I. Akhiezer. *The classical moment problem and some related questions in analysis.* University Mathematical Monographs. Hafner publishing company, New York, 1965.

[2] L. Arnold. *Stochastische Differentialgleichungen.* Oldenbourg Verlag, München, Wien, 1973.

[3] S.K. Berberian. *Measure and Integration.* MacMillan Company, New York, 1965.

[4] A. Choudhry. Symmetric Diophantine systems. *Acta Arithmetica,* 59:291–307, 1991.

[5] A. Choudhry. Symmetric Diophantine systems revisited. *Acta Arithmetica,* 119:329–347, 2005.

[6] F.R.K. Chung. *Spectral graph theory,* volume 92 of *CBMS Regional Conference Series in Mathematics.* AMS.

[7] J.B. Conway. *A course in functional analysis,* volume 96 of *Graduate texts in Mathematics.* Springer-Verlag, Berlin, 1985.

[8] I.P. Cornfeld, S.V.Fomin, and Ya.G.Sinai. *Ergodic Theory,* volume 115 of *Grundlehren der mathematischen Wissenschaften in Einzeldarstellungen.* Springer Verlag, 1982.

[9] D.R. Cox and V. Isham. *Point processes.* Chapman & Hall, London and New York, 1980. Monographs on Applied Probability and Statistics.

[10] R.E. Crandall. The challenge of large numbers. *Scientific American,* (Feb), 1997.

[11] Daniel D. Stroock.

[12] P. Deift. Applications of a commutation formula. *Duke Math. J.,* 45(2):267–310, 1978.

[13] M. Denker, C. Grillenberger, and K. Sigmund. *Ergodic Theory on Compact Spaces.* Lecture Notes in Mathematics 527. Springer, 1976.

[14] John Derbyshire. *Prime obsession.* Plume, New York, 2004. Bernhard Riemann and the greatest unsolved problem in mathematics, Reprint of the 2003 original [J. Henry Press, Washington, DC; MR1968857].

[15] L.E. Dickson. *History of the theory of numbers. Vol.II:Diophantine analysis.* Chelsea Publishing Co., New York, 1966.

[16] S. Dineen. *Probability Theory in Finance, A mathematical Guide to the Black-Scholes Formula,* volume 70 of *Graduate Studies in Mathematics.* American Mathematical Society, 2005.

[17] J. Doob. *Stochastic processes.* Wiley series in probability and mathematical statistics. Wiley, New York, 1953.

[18] J. Doob. *Measure Theory.* Graduate Texts in Mathematics. Springer Verlag, 1994.

[19] P.G. Doyle and J.L. Snell. *Random walks and electric networks,* volume 22 of *Carus Mathematical Monographs.* AMS, Washington, D.C., 1984.

[20] T.P. Dreyer. *Modelling with Ordinary Differential equations.* CRC Press, Boca Raton, 1993.

[21] R. Durrett. *Probability: Theory and Examples.* Duxburry Press, second edition edition, 1996.

[22] M. Eisen. *Introduction to mathematical probability theory.* Prentice-Hall, Inc, 1969.

[23] N. Elkies. An application of Kloosterman sums. http://www.math.harvard.edu/ elkies/M259.02/kloos.pdf, 2003.

[24] Noam Elkies. On $a^4 + b^4 + c^4 = d^4$. *Math. Comput.,* 51:828–838, 1988.

[25] N. Etemadi. An elementary proof of the strong law of large numbers. *Z. Wahrsch. Verw. Gebiete,* 55(1):119–122, 1981.

[26] W. Feller. *An introduction to probability theory and its applications.* John Wiley and Sons, 1968.

[27] D. Freedman. *Markov Chains.* Springer Verlag, New York Heidelberg, Berlin, 1983.

[28] E.M. Wright G.H. Hardy. *An Introduction to the Theory of Numbers.* Oxford University Press, Oxford, fourth edition edition, 1959.

[29] J.E. Littlewood G.H. Hardy and G. Polya. *Inequalities.* Cambridge at the University Press, 1959.

[30] R.T. Glassey. *The Cauchy Problem in Kinetic Theory.* SIAM, Philadelphia, 1996.

[31] J. Glimm and A. Jaffe. *Quantum physics, a functional point of view.* Springer Verlag, New York, second edition, 1987.

[32] G. Grimmet. *Percolation.* Springer Verlag, 1989.

[33] G. Grimmet and D.R. Stirzaker. *Probability and Random Processes, Problems and Solutions.* Clarendon PRess, Oxford.

[34] A. Gut. *Probability: A graduate Course.* Springer texts in statistics. Springer, 2005.

[35] Richard K. Guy. *Unsolved Problems in Number Theory.* Springer, Berlin, 3 edition, 2004.

[36] P. Halmos. *Lectures on ergodic theory.* The mathematical society of Japan, 1956.

[37] Paul R. Halmos. *Measure Theory.* Springer Verlag, New York, 1974.

[38] G.H. Hardy. *Ramanujan.* Cambridge at the University Press, 1940. Twelve Lectures on Subjects by his Life and Work.

[39] W.K. Hayman. *Subharmonic functions I,II,* volume 20 of *London Mathematical Society Monographs.* Academic Press, Inc. Harcourt Brace Jovanovich, Publishers, London, 1989.

[40] H.G.Tucker. *A graduate course in probability.* Probability and Mathematical Statistics. Academic Press, 1967.

[41] O. Gurel-Gurevich I. Benjamini and B. Solomyak. Branching random walk with exponentially decreasing steps and stochastically self-similar measures. arXiv PR/0608271, 2006.

[42] R. Isaac. *The Pleasures of Probability.* Graduate Texts in Mathematics. Springer Verlag, 1995.

[43] K. Itô and H.P. McKean. *Diffusion processes and their sample paths,* volume 125 of *Die Grundlehren der mathematischen Wissenschaften.* Springer-Verlag, Berlin, second printing edition, 1974.

[44] V. Kac and P. Cheung. *Quantum calculus.* Universitext. Springer-Verlag, New York, 2002.

[45] J-P. Kahane and R. Salem. *Ensembles parfaits et séries trigonométriques.* Hermann, 1963.

[46] I. Karatzas and S. Shreve. *Brownian motion and stochastic calculus,* volume 113 of *Graduate Texts in Mathematics.* Springer-Verlag, New York, second edition, 1991.

[47] A.F. Karr. *Probability.* Springer texts in statistics. Springer-Verlag, 1993.

[48] Y. Katznelson. *An introduction to harmonic analysis.* Dover publications, Inc, New York, second corrected edition edition, 1968.

[49] J.F.C. Kingman. *Poisson processes*, volume 3 of *Oxford studies in probability.* Clarendon Press, New York: Oxford University Press, 1993.

[50] O. Knill. A remark on quantum dynamics. *Helvetica Physica Acta*, 71:233–241, 1998.

[51] O. Knill. Singular continuous spectrum and quantitative rates of weakly mixing. *Discrete and continuous dynamical systems*, 4:33–42, 1998.

[52] B. Reznick K.O. Bryant and M. Serbinowska. Almost alternating sums. *Amer. Math. Monthly.*

[53] N. Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung.* Berlin, 1933. English: Foundations of Probability Theory, Chelsea, New York, 1950., 1933.

[54] U. Krengel. *Ergodic Theorems*, volume 6 of *De Gruyter Studies in Mathematics.* Walter de Gruyter, Berlin, 1985.

[55] H. Kunita. *Stochastic flows and stochastic differential equations.* Cambridge University Press, 1990.

[56] Y. Last. Quantum dynamics and decompositions of singular continuous spectra. *J. Func. Anal.*, 142:406–445, 1996.

[57] J. Lewis. An elementary approach to Brownian motion on manifolds. In *Stochastic processes—mathematics and physics (Bielefeld, 1984)*, volume 1158 of *Lecture Notes in Math.*, pages 158–167. Springer, Berlin, 1986.

[58] E.H. Lieb and M. Loss. *Analysis*, volume 14 of *Graduate Studies in Mathematics.* American Mathematical Society, 1996.

[59] J.L. Selfridge L.J. Lander, T.R. Parken. A survey of equal sums of like powers. *Mathematics of Computation*, (99):446–459, 1967.

[60] E. Lukacs and R.G.Laha. *Applications of Characteristic Functions.* Griffin's Statistical Monogrpaphs and Courses.

[61] M.C. Mackey. *Time's arrow: the origins of thermodynamics behavior.* Springer-Verlag, New York, 1992.

[62] N. Madras and G. Slade. *The self-avoiding random walk.* Probability and its applications. Birkh 1993.

[63] L. Malozemov and A. Teplyaev. Pure point spectrum of the Laplacians on fractal graphs. *J. Funct. Anal.*, 129(2):390–405, 1995.

[64] K.V. Mardia. *Statistics of directional data.* Academic press, London and New York, 1972.

[65] A. Matulich and B.N. Miller. Gravity in one dimension: stability of a three particle system. *Colloq. Math.*, 39:191–198, 1986.

[66] H.P. McKean. *Stochastic integrals.* Academic Press, 1969.

[67] L. Merel. Bornes pour la torson des courbes elliptiques sur les corps de nombres. *Inv. Math.*, 124:437–449, 1996.

[68] Jean-Charles Meyrignac. Existence of solutions of (n,n+1,n+1). http://euler.free.fr/theorem.htm.

[69] Jean-Charles Meyrignac. Records of equal sums of like powers. http://euler.free.fr/records.htm.

[70] F. Mosteller. *Fifty Challenging Problems in Probability with solutions.* Dover Publications, inc, New York, 1965.

[71] M. Nagasawa. *Schrödinger equations and diffusion theory*, volume 86 of *Monographs in Mathematics.* Birkhäuser Verlag, Basel, 1993.

[72] I.P. Natanson. *Constructive theory of functions.* Translation series. United states atomic energy commission, 1949. State Publishing House of Technical-Theoretical Literature.

[73] E. Nelson. *Dynamical theories of Brownian motion.* Princeton university press, 1967.

[74] E. Nelson. *Radically elementary probability theory.* Princeton university text, 1987.

[75] T.Lewis N.I. Fisher and B.J. Embleton. *Statistical analysis of spherical data.* Cambridge University Press, 1987.

[76] B. Oksendal. *Stochastic Differential Equations.* Universitext. Springer-Verlag, New York, fourth edition edition, 1995.

[77] K. Petersen. *Ergodic theory.* Cambridge University Press, Cambridge, 1983.

[78] I. Peterson. *The Jungles of Randomness, A mathematical Safari.* John Wiley and Sons, Inc.

[79] S. C. Port and C.J. Stone. *Brownian motion and classical potential theory.* Probability and Mathematical Statistics. Academic Press (Harcourt Brace Jovanovich Publishers),New York, 1978.

[80] T. Ransford. *Potential theory in the complex plane*, volume 28 of *London Mathematical Society Student Texts.* Cambridge University Press, Cambridge, 1995.

[81] M. Reed and B. Simon. *Methods of modern mathematical physics , Volume I.* Academic Press, Orlando, 1980.

[82] C.J. Reidl and B.N. Miller. Gravity in one dimension: The critical population. *Phys. Rev. E*, 48:4250–4256, 1993.

[83] D. Revuz and M.Yor. *Continuous Martingales and Brownian Motion.* Springer Verlag, 1991. Grundlehren der mathmatischen Wissenschaften, 293.

[84] H. Riesel. *Prime numbers and computer methods for factorization*, volume 57 of *Progress in Mathematics.* Birkhäuser Boston Inc., 1985.

[85] P.E. Hart R.O. Duda and D.G. Stork. *Pattern Classification.* John Wiley and Sons, Inc, New York, second edition edition.

[86] C.A. Rogers. *Hausdorff measures.* Cambridge University Press, 1970.

[87] S.M. Ross. *Applied Probability Models with optimization Applications.* Dover Publications, inc, New York, 1970.

[88] W. Rudin. *Real and Complex Anslysis.* McGraw-Hill Series in Higher Mathematics, 1987.

[89] D. Ruelle. *Chance and Chaos.* Princeton Science Library. Princeton University Press, 1991.

[90] G.B. Rybicki. Exact statistical mechanics of a one-dimensional self-gravitating system. In M. Lecar, editor, *Gravitational N-Body Problem*, pages 194–210. D. Reidel Publishing Company, Dordrecht-Holland, 1972.

[91] A. Shiryayev. *Probability*, volume 95 of *Graduate Texts in Mathematics.* Springer-Verlag, New York, 1984.

[92] Hwei P. Shu. *Probability, Random variables and Random Processes.* Schaum's Outlines. McGraw-Hill, 1997.

[93] B. Simon. *Functional Integration and Quantum Physics.* Academic Press, 1979. Pure and applied mathematics.

[94] B. Simon. Spectral analysis of rank one perturbations and applications. In *Mathematical quantum theory. II. Schrödinger operators (Vancouver, BC, 1993)*, volume 8 of *CRM Proc. Lecture Notes*, pages 109–149. AMS, Providence, RI, 1995.

[95] B. Simon. Operators with singular continuous spectrum. VI. Graph Laplacians and Laplace-Beltrami operators. *Proc. Amer. Math. Soc.*, 124(4):1177–1182, 1996.

[96] B. Simon and T. Wolff. Singular continuous spectrum under rank one perturbations and localization for random Hamiltonians. *Commun. Pure Appl. Math.*, 39:75.

[97] Ya. G. Sinai. *Probability Theory, An Introductory Course.* Springer Textbook. Springer Verlag, Berlin, 1992.

[98] J.L. Snell and R. Vanderbei. Three bewitching paradoxes. Probability and Stochastics Series. CRC Press, Boca Raton, 1995.

[99] P.M. Soardi. *Potential theory on infinite networks*, volume 1590 of *Lecture Notes in Mathematics.* Springer-Verlag, Berlin, 1994.

[100] F. Spitzer. *Principles of Random walk.* Graduate texts in mathematics. Springer-Verlag, New York Heidelberg Berlin, 1976.

[101] H. Spohn. *Large scale dynamics of interacting particles.* Texts and monographs in physics. Springer-Verlag, New York, 1991.

[102] R.S. Strichartz. Fourier asymptotics of fractal measures. *J. Func. Anal.*, 89:154–187, 1990.

[103] R.S. Strichartz. Self-similarity in harmonic analysis. *The journal of Fourier analysis and Applications*, 1:1–37, 1994.

[104] D.W. Stroock. *Lectures on Stochastic Analysis and Diffusion Theory.* London Mathematical Society Students Texts. Cambridge University Press, 1987.

[105] D.W. Stroock. *Probability theory, an analytic view.* Cambridge University Press, 1993.

[106] N. van Kampen. *Stochastic processes in physics and chemistry.* North-Holland Personal Library, 1992.

[107] P. Walters. *An introduction to ergodic theory.* Graduate texts in mathematics 79. Springer-Verlag, New York, 1982.

[108] D. Williams. *Probability with Martingales.* Cambridge mathematical Texbooks, 1991.

[109] G.L. Wise and E.B. Hall. *Counterexamples in probability and real analysis.* Oxford University Press, 1993.

Index

Probability Theory and Stochastic Processes with Applications

About the Book-

Chapter 1-2 of this text covers material of a basic probability course. Chapter 3 deals with discrete stochastic processes including Martingale theory. Chapter 4 covers continuous time stochastic processes like Brownian motion and stochastic differential equations. The last chapter "selected topics" got considerably extended in the summer of 2006. In the original course, only localization and percolation problems were included. Now, more topics like estimation theory, Vlasov dynamics, multi-dimensional moment problems, random maps, circle-valued random variables, the geometry of numbers, Diophantine equations and harmonic analysis have been added. No previous knowledge in probability are necessary, but a basic exposure to calculus and linear algebra is required. Some real analysis as well as some background in topology, functional analysis and harmonic analysis can be helpful for the later chapters.

About the Author-

Oliver Knill grew up near Schaffhausen in Switzerland. He got his BA in Mathematics at ETH Zürich with a senior theses in dynamical systems advised by Jürgen Moser. After a postgraduate visit at Technion in Haifa he wrote his PhD theses at the ETH Zürich advised by Oscar Lanford III in the area of dynamical systems and random operator theory. After three years in Pasadena at Caltech as a Taussky-Todd instructor in the group of Barry Simon, and one year at the University of Arizona in Tucson as a Hanno-Rund visiting assistant professor, he was a Swiss National Science foundation fellow at the University of Texas in Austin and worked a year as a research assistant at the Institute for Fusion Studies at UT. Since Fall 2000, he teaches at the Mathematics department at Harvard University.

Oliver Knill started his research in the field of dynamical systems, tackling ergodic and spectral theoretical questions. He worked out and published Moser's lecture notes "Selected chapters in the calculus of variations" which appeared in 2003 in the series Lectures in Mathematics ETH Zürich. More recently he works on applications of dynamical systems to probability theory and elementary number theory as well as inverse problem in analysis, geometry, and computer vision.

Oliver Knill, Department of Mathematics,
One Oxford Street, Harvard University,
Cambridge, MA 02138, USA

Visit Overseas Press on the World Wide Web at : www.overseaspub.com.

ISBN 81-89938-40-1

9 788189 938406

OVERSEAS PRESS (INDIA) PRIVATE LIMITED
7/28, Ansari Road, Daryaganj, New Delhi-110002