An introduction to the theory of Markov processes
mostly for physics students

Christian Maes

1Instituut voor Theoretische Fysica, KU Leuven, Belgium

(Dated: 21 June 2013)

Since about 200 years it is generally realized how fluctuations and chance play a
prominent role in fundamental studies of science. The main examples have come from
astronomy (Laplace, Poincaré), biology (Darwin), statistical mechanics (Maxwell,
Boltzmann, Gibbs, Einstein) and the social sciences (Quetelet). The mere power
of numbers for large systems and the unavoidable presence of fluctuations for small
systems make the theory of chance very much part of basic physics. But today also
other domains like economy, business, technology and medicine increasingly demand
complex stochastic models. Stochastic techniques have led to a richer variety of
modeling accompanied by powerful computational methods. An important subclass
of stochastic processes are Markov processes, where memory effects are strongly
limited and to which the present notes are devoted.

I. INTRODUCTION

What follows is a fast and brief introduction to Markov processes. These are a class of
stochastic processes with minimal memory: the update of the system’s state is function only
of the present state, and not of its history. We know such type of evolutions well, as they
appear from the first order differential equations that we traditionally use in mechanics for
the autonomous evolution of a state (positions and momenta). Also in quantum mechanics,
with the Schrödinger equation for the wave function, the update is described by a first order
dynamics. Markov processes add noise to these descriptions, and such that the update is
not fully deterministic. The result is a class of probability distributions on the possible
trajectories.

*Electronic address: christian.maes@fys.kuleuven.be;itf.fys.kuleuven.be/~christ/
II. GENERALITIES, PERHAPS MOTIVATING

The theory of chances, more often called probability theory, has a long history. It has mostly come to us not via the royal road on which we find geometry, algebra or analysis but more via sideways. We need to cross market squares and quite a lot can be learnt from entering gambling houses.

Many departments of mathematics still today have no regular seminar or working group on probability theory, even though the axiomatization of probability is now already some 80 years old (Kolmogorov 1933). Very often probability theory gets mixed up with statistics or with measure theory, and more recently it gets connected with information theory. On the other hand, all in all, physicists have always appreciated a bit of probability. For example, we need a theory of errors and we need to understand how reliable can be our observations and experiments. In other important cases chances enter physics for a variety of (overlapping) reasons, such as

1. The world is big and nature is complex and uncertain to our crude senses and our little minds. By lack of certain information, we talk about plausibility, hopefully estimated in an objective scientific manner. Chances are the result, making quantitative degrees for plausibility.

2. The dynamics we are studying in physics are often very sensitive to initial and boundary conditions. The technical word is chaos. Chaotic evolutions appear very erratic and random. Practical unpredictability is the natural consequence; probabilistic considerations then become useful or even unavoidable.

3. The details at a very fine degree of accuracy, are not always relevant to our description. Some degrees of freedom can better be replaced by noise or chance. In so doing we get a simple and practical way of treating a much larger complex of freedoms.

4. Quantum mechanical processes often possess chance-like aspects. For example, there are fundamental uncertainties about the time of decay of a unstable particle, or we can only give the probabilities for the positions of particles. Therefore again, we use probability models. In fact two major areas of physics, statistical mechanics and quantum mechanics, depend on notions of probability for their theoretical foundation and deal explicitly with laws of chance.
One of the first to have the idea to apply probability theory in physics was Daniël Bernoulli. He did this in an article (1732) on the inclination of planetary orbits relative to the ecliptic. The chance that all the inclination angles are less than 9 degrees is so small that Daniël inferred that this phenomenon must have a definite cause. This conceptual breakthrough and this way of reasoning had a major influence on Pierre Simon de Laplace who has resumed that from his first works on. Daniël Bernoulli also applies these first statistical considerations to a gas model, the beginning of the kinetic theory of gases. That idea was also independently rediscovered by John Waterston and by August Krönig, while Maxwell and Boltzmann take over towards the end of the 19th century. Boltzmann will give a most important contribution to the statistical interpretation of thermodynamic entropy. For the first time a statistical law enters physics. Maxwell calls it a moral certainty and the second law of thermodynamics gets a statistical status. That way Maxwell repeats (and cites) another Bernoulli, Jacob Bernoulli and author of the Ars Conjectandi in 1713: we ought to learn how to guess because the true logic of the world is the calculus of Probabilities, which takes into account of the magnitude of the probability which is, or ought to be, in a reasonable man's mind. (J.C. Maxwell).

The goals of this course are typical for a physics education, modeling and analysis: (1) learning how to translate to and to model in the mathematics of stochastics, (2) learning to calculate with probabilistic models, here Markov processes.

III. REMINDER OF BASIC CONCEPTS — PROBABILITIES

A. The chance-trinity

Speaking of probabilities implies that we assign some number between zero and one to events. Let us start from the events. They can mostly be viewed as sets of more elementary outcomes. As an example, to throw an even number of eyes is an event $A$ and consists of the more elementary outcomes 2, 4 and 6, or $A = \{2, 4, 6\}$. The basic thing is thus to know the set (or space) of all possible outcomes; we call it the universe $\Omega$. Each outcome is an element $\omega \in \Omega$. An event $A$ is a set of such outcomes, hence a subset of $\Omega$. A probability
distribution gives numbers between 0 and 1 to these events. The universe, the possible events and their probabilities: well, that is the chance-trinity.

The set of events $\mathcal{F}$ deserves some structure. First of all we want that $\Omega$ and the empty set $\emptyset$ belong to it. But also, that if $A, B \in \mathcal{F}$, then also $A \cap B, A \cup B, A^c \in \mathcal{F}$, where the latter $A^c = \Omega \setminus A$ denotes the complement of the set $A$. Such a $\mathcal{F}$ is called a field. We say that it is a $\sigma$–field when $\mathcal{F}$ is also closed under countable unions (or intersections).

On that set $\mathcal{F}$ we put a (additive countable) probability law, which is a non-negative function $P$ defined on the $A \in \mathcal{F}$ for which $P(A) = 1 - P(A^c) \geq 0$, $P(\Omega) = 1 - P(\emptyset) = 1$, and secondly

$$P(A) = \sum_n P(A_n)$$

whenever $A = \cup_n A_n$ for each sequence of mutually disjoint $A_n \in \mathcal{F}$.

Let us have a look at some examples. If $\Omega$ is countable, it is less of a problem. Why not taking the simplest case: for a coin, we have $\Omega = \{\text{up, down}\}$, and $\mathcal{F}$ contains 4 elements. A probability law on it is completely determined by the the probability of up, $P(\text{up})$. It becomes more difficult for uncountable universes. Let us take the real numbers, $\Omega = \mathbb{R}$. The natural $\sigma$-field here is the so called Borel $\sigma$–field. That is defined as the smallest $\sigma$–field that contains all intervals, and in particular contains all open sets.

When we have an additive countable probability law on a space (called, measure space) $(\Omega, \mathcal{F})$ with a universe $\Omega$ and a $\sigma$–field $\mathcal{F}$, then the triple $(\Omega, \mathcal{F}, P)$ is called a probability space. That is the start. For example, a stochastic variable is a map $f : \Omega \to \mathbb{R}$ so that for all Borel-sets $B$, $f^{-1}B \in \mathcal{F}$. The meaning is simply that $f$ takes values that depend on the outcome of a chance-experiment.

For the purpose of these lectures we consider mostly two classes of probability spaces. There is first the state space $K$, say a finite set on which we can define probability distributions $\rho$ with

$$\rho(x) \geq 0, \quad \sum_{x \in K} \rho(x) = 1$$
We also write
\[ \text{Prob}_\rho[B] = \sum_{x \in B} \rho(x) \]
for the probability of event \( B \) and
\[ \rho(f) = \langle f \rangle_\rho = \sum_{x \in K} f(x) \rho(x) \]
for the expectation of a function \( f \) on \( K \) (an observable). Observables are random variables and can be added and multiplied because we take them real-valued. The subscripts \( \rho \) under \( \text{Prob} \) or under \( \langle \cdot \rangle_\rho \) will only be used when necessary.

Secondly there will be the space of possible trajectories (time-sequences of states) for which we will reserve the letter \( \Omega \), and that one will become uncountable in continuous time.

### B. State space

For our purposes it is sufficient to work with a finite number of states, together making the set \( K \). Obviously that corresponds to quite a reduced description — one forgets about the microscopic details, taking them into account effectively via the laws of probability. For example, these states can be possible energy levels of an atom or molecule, or some discrete set of chemo-mechanical configurations of an enzyme, or a collection of (magnetic) spins on a finite lattice, or the occupation variables on a graph etc. When we would allow a continuum of states we can include velocity and position distributions as from classical mechanics, or even probabilities on wave functions as for quantum mechanics.

The elements of \( K \) are called states. The \( \sigma \)-field (the events) are all possible subsets of \( K \). A probability distribution \( \mu \) on \( K \) specifies numbers \( \mu(x) \geq 0, x \in K \), that sum to one, \( \sum_x \mu(x) = 1 \). Here are some simple examples:

1. **spin configurations**: the state space is \( K = \{+1, -1\}^N \) where \( N \) is the number of spins. Each spin has two values (up or down) that we take as \( \pm 1 \). The elements of \( K \) are then \( N \)-tuples \( x = (x(1), x(2), \ldots, x(N)) \), with each \( x(i) = \pm 1 \). Here is a nice probability law

\[ \rho(x) = \frac{1}{Z} \exp \left( \frac{J}{N} \sum_{i,j=1}^N x(i)x(j) \right) \]
for a certain coupling constant $J$ and $Z$ is the normalization. This law defines the Curie-Weiss model for magnetism. The normalization (partition function) $Z$ is

$$Z = \sum_{x \in K} \exp \left( \frac{J}{N} \sum_{i,j=1}^{N} x(i) x(j) \right) = \sum_{m} e^{N \left( J m^2 + s_N(m) \right)}$$

where the last sum is over $m = -1, -1 + 2/N, \ldots, 1 - 2/N, 1$ (the $N + 1$ possible values of $\sum_{i=1}^{N} x(i)/N$), and $s_N(m)$ is the entropy per particle

$$s_N(m) = \frac{1}{N} \log \frac{N!}{(N^{1+m}/2)! (N^{1-m}/2)!}$$

Try to check that!

(2) lattice gas: the state space is $K = \{0, 1\}^\Lambda$ where $\Lambda$ is a finite graph, or a finite piece of a lattice. At each site (or vertex) $i \in \Lambda$ sits one or no particle. The state is thus specified upon saying which sites are occupied and which are vacant, $x(i) \in \{0, 1\}$. A product probability law is given by

$$\rho(x) = \frac{1}{Z} P^{\mathcal{N}(x)}$$

where $p \in [0, 1]$ is a parameter and $\mathcal{N}(x) = \sum_i x(i)$ is the total number of particles on the lattice.

FIG. 1: Particle configuration in a lattice gas.
C. Conditional probabilities/expectations

Additional information can be used to deform our probabilities. For example, if we know that the event $B$ happens (is true), then

$$\rho(x|B) = \frac{\rho(x)}{\text{Prob}[B]}, \quad \text{if } x \in B, \quad \rho(x) = 0 \text{ otherwise}$$

is a conditional probability (given the event $B$ for which we assume a priori that $\rho(B) > 0$). Now the probability of event $A$ is

$$\text{Prob}[A|B] = \frac{\text{Prob}[A \cap B]}{\text{Prob}[B]}$$

Of course that defines a bona fide probability distribution $\rho(\cdot|B)$ on $K$. We can also take expectations of a function $f$ as

$$\langle f|B \rangle_{\rho} = \sum_{x \in K} \rho(x|B) \ f(x)$$

which is a conditional expectation. If we have a partition $\{B_i\}$ of the state space $K$, then

$$\rho(x) = \sum_i \rho(x|B_i) \text{Prob}[B_i]$$

Finally, it is useful to remember Bayes’ formula

$$\text{Prob}[A|B] = \text{Prob}[B|A] \frac{\text{Prob}[A]}{\text{Prob}[B]} \quad \text{(III.1)}$$

for two events $A$ and $B$ that have positive probability.

We say that two random variables $f$ and $g$ are independent when their joint distribution factorizes. In other words

$$\text{Prob}[f(x) = a|g(x) = b] = \text{Prob}[f(x) = a]$$

for all $a, b$ possible values of respectively $f$ and $g$. As a consequence then, their covariance $\langle fg \rangle - \langle f \rangle \langle g \rangle = 0$.

D. Distributions

Some probability distributions got names, for good reasons. The mother is Bernoulli, for which $K = \{0, 1\}$ and the probability $\rho(0) = 1 - p, \rho(1) = p$ is completely determined by
the parameter \( p \in [0, 1] \) (probability of success). Next is the binomial distribution, which
asks for the number of successes out of \( n \) independently repeated Bernoulli experiments. We
now have two parameters, \( n \) and \( p \) and the probability to get \( k \) successes out of \( n \) trials is
\[
\rho_{n,p}(k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n
\]
If we here let \( p \downarrow 0 \) while \( n \uparrow +\infty \) keeping \( \lambda = pn \) fixed, we find
\[
\rho_{n,p}(k) \rightarrow e^{-\lambda} \frac{\lambda^k}{k!}, \quad k = 0, 1, \ldots
\]
which is the Poisson distribution with parameter \( \lambda > 0 \). If \( n \) is counted as time, then \( \lambda \)
would be running as (proportional to) a continuous time.
If, on the other hand, we take \( k = mn \pm \sigma \sqrt{n} \), then the binomial distribution can be seen
to converge to the normal (Gaussian) density,
\[
\rho(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-m)^2}, \quad x \in \mathbb{R}
\]
Finally we look at the exponential distribution. We say that a random variable \( \tau \) with values
in \([0, \infty)\) is exponentially distributed when its probability density is
\[
\rho(t) = \xi e^{-\xi t}, \quad t \geq 0
\]
for some parameter \( \xi > 0 \). In other words, for \( 0 \leq a \leq b \),
\[
\text{Prob}[a < \tau < b] = \xi \int_a^b e^{-\xi t} \, dt
\]
The exponential distribution can be seen as the limit of the geometric distribution, in the
same sense as the Poisson distribution emerges from the binomial distribution.

IV. EXERCISES

1. Monty Hall problem.
Suppose you’re on a game show and you’re given the choice of three doors [and will win
what is behind the chosen door]. Behind one door is a car; behind the others, goats
[unwanted booby prizes]. The car and the goats were placed randomly behind the doors
before the show. The rules of the game show are as follows: After you have chosen a door,
the door remains closed for the time being. The game show host, Monty Hall, who knows
what is behind the doors, now has to open one of the two remaining doors, and the door he opens must have a goat behind it. If both remaining doors have goats behind them, he chooses one [uniformly] at random. After Monty Hall opens a door with a goat, he will ask you to decide whether you want to stay with your first choice or to switch to the last remaining door. Imagine that you chose Door 1 and the host opens Door 3, which has a goat. He then asks you “Do you want to switch to Door Number 2?” Is it to your advantage to change your choice?


2. Boy and Girl paradox.

From all families with two children, at least one of whom is a boy, a family is chosen at random. What is the probability that both children are boys?

(Each child is either male or female. Each child has the same chance of being male as of being female. The sex of each child is independent of the sex of the other.)


Compute the approximate probability that in a room of \( n \) people, at least two have the same birthday. For simplicity, disregard variations in the distribution, such as leap years, twins, seasonal or weekday variations, and assume that the 365 possible birthdays are equally likely. For what \( n \) is that probability approximately 50 percent?


Given a needle of length \( \ell \) dropped on a plane ruled with parallel lines \( d \) units apart, what is the probability that the needle will cross a line? See Fig. 2.


5. Taxi problem.

In Redshift (North-Dakota) only green and blue cabs are allowed. There are 80 percent green taxi cars and 20 percent blue taxi cars. Some local taxi car there was implied in a
FIG. 2: Throwing needles.

deadly hit-and-run. Our only witness claims the taxi car was green. An investigation shows however that the witness always remembers green as green, but in half the cases also says green when blue is shown. Estimate the plausibility that the accident was caused by a blue taxi.

6. Poisson approximation
Suppose that $N$ electrons pass a point in a given time $t$ with $N$ following a Poisson distribution. The mean current is $I = e\langle N \rangle / t$. Show that we can estimate the charge $e$ from the current fluctuations.

7. Central limit theorem
Explain the claim in Section III D about the convergence (in distribution) of the binomial distribution to the Gaussian distribution.

8. Statistical independence
Suppose that $f$ and $g$ are two random variables with expectations $\langle fg \rangle = \langle f \rangle = 0$. Can
you find an example where \( f \) and \( g \) are not independent?

Suppose now that \( \langle fg \rangle = \sqrt{\langle f^2 \rangle \langle g^2 \rangle} \) with \( \langle f \rangle = \langle g \rangle = 0 \). Show that \( f \) and \( g \) are linearly dependent (\( f = ag \) for some constant \( a \)).

What is the variance of a sum of \( n \) independent random variables?

9. **Exponential distribution**

The exponential distribution is memoryless. Show that its probabilities satisfy

\[
\text{Prob}[\tau > t + s | \tau > s] = \text{Prob}[\tau > t]
\]

for all \( t, s \geq 0 \). The only memoryless continuous probability distributions are the exponential distributions, so memorylessness completely characterizes the exponential distributions among all continuous ones.

10. For the lattice gas near the end of Section III B, find the probability that a given specified site is occupied.

What is the normalization \( Z \) in the given example for \( \rho \)?

11. **Fair coin** (Bernoulli variable with \( p = q = 1/2 \)).

A fair coin is thrown repeatedly. What is the probability that on the \( n \)-th throw

a) a head appears for the first time? Moreover, prove that sooner or later a head shows up.

b) the numbers of heads and tails are equal?

c) exactly two heads have appeared altogether?

d) at least two heads have appeared?

12. **Gambler’s ruin**

A man is saving money to buy a new car at a cost of \( N \) units of money. He starts with \( k \) (\( 0 < k < N \)) units and tries to win the remainder by gambling with his bank manager. He tosses a coin (\( P\{H\} = p = 1 - P\{T\} \)) repeatedly; if it comes up heads \( H \) then the manager pays him one unit, but if it comes up tails \( T \) then he pays the manager one unit. He plays this game until one of two events occurs: either he runs out of money or he wins enough to buy the car. What is the probability that he goes bankrupt?

What is the mean number of steps before hitting one of the stop conditions?
13. Poisson distribution
An important probability distribution for physicists (and the like) was first introduced by Siméon Denis Poisson in 1837 in his *Recherches sur la probabilité des jugements en matière criminelle et en matière civile*. We have met that distribution already in Section III D: random variable $X$ with values in $\mathbb{N}$ is Poissonian if

$$P[X = k] = e^{-z} \frac{z^k}{k!}, \quad k = 0, 1, 2, \ldots$$

where $z > 0$ is a parameter. Show that its mean equals its variance equals $z$. The $X$ most often refers to a number (of arrivals, mutations, changes, actions, jumps, ...) in a certain space-time window, like the number of particles in a region of space in a gas where $z$ would be the fugacity.
The easiest way to imagine a trajectory over state space $K$ is to think of uniform steps (say of size one) over which the state can possibly change. The path space is then $\Omega = K^\mathbb{N}$ with elements $\omega = (x_0, x_1, x_2, \ldots)$ where each $x_n \in K$ (the state at time $n$). We can now build certain probability laws on $\Omega$. They are parameterized by two types of objects: (1) the initial distribution (the law $\mu$ from which we draw our initial state $x_0$), and (2) the updating rule (giving the conditional probability of getting some value for $x_n$ giving the state $x_{n-1}$ at time $n-1$). As the updating rule only uses the present state, and not the past history, we say that the process is Markovian.

In 1907, Andrei Andreyevich Markov started studying such exciting new types of chance processes, see A.A. Markov, *An Example of Statistical Analysis of the Text of Eugene Onegin Illustrating the Association of Trials into a Chain*. Bulletin de l’Académie Impériale des Sciences de St. Petersburg, ser. 6, vol. 7 (1913), pp. 153162. In these processes, the outcome of a given experiment affects or can affect the outcome of the next experiment. (From http://www.saurabh.com/Site/Writings.html, last accessed on 27 Oct.2011)

Suppose you are given a body of text and asked to guess whether the letter at a randomly selected position is a vowel or a constant. Since consonants occur more frequently than vowels, your best bet is to always guess consonant. Suppose we decide to be a little more helpful and tell you whether the letter preceding the one you chose is a vowel or consonant. Is there now a better strategy you can follow? Markov was trying to answer the above problem and analysed twenty thousand letters from Pushkin’s poem *Eugene Onegin*. He found that 43 percent letters were vowels and 57 percent were consonants. So in the first problem, one should always guess “consonant” and can hope to be correct 57 percent of the time. However, a vowel was followed by consonant 87 percent of the time. A consonant was followed by a vowel 66 percent of the time. Hence, guessing the opposite of the preceding letter would be a better strategy in the second case. Clearly, knowledge of the preceding letter is helpful.

The real insight came when Markov took the analysis a step further. Markov investigated whether knowledge about the preceding two letters confers any additional advantage. He found that there was no significant advantage to knowing the additional preceding letter. This leads to the central idea of a Markov chain — while the successive outcomes are not
independent, only the most recent outcome is of use in making a prediction about the next outcome.

### A. Example: the Ehrenfest model

Let us start with a famous example, Ehrenfest’s model, also called the *dog-and-flea* model, for reasons that will become clear. The state space is $K = \{0, 1, 2, \ldots, N\}$ with each state representing the number $x$ of particles in a vessel. For the updating we imagine there is another vessel containing $N - x$ particles. We randomly pick a particle from the two vessels, and we move it to the other vessel with probability $p$. Alternatively, we leave it where we found it with probability $1 - p$. In that way $x$ will change, by a stochastic rule.

Let us add the formulæ. We have $x \to x + 1$ at the next time with probability $p(N - x)/N$, $x \to x - 1$ with probability $px/N$ and $x$ remains $x$ with probability $1 - p$. We can write it in a matrix $p(x, y)$ where $x$ is the present state and $y$ is the new possible state. That matrix of transition probabilities will be abstracted in the next section. Here we simply write $p(x, x + 1) = p(N - x)/N$, $p(x, x - 1) = px/N$ and $p(x, x) = 1 - p$. The rest of the matrix elements are zero.

![FIG. 3: Two vessels and particles.](image)
path sequence \((2, 3, 4, 5)\) the path probability is \(\mu(2) p^3 \frac{(N-2)(N-3)(N-4)}{N^3}\) while the reversed path \((5, 4, 3, 2)\) has probability \(\mu(5) p^3 \frac{5 \times 4 \times 3}{N^3}\) which is only equal to the previous probability when

\[
\frac{\mu(2)}{\mu(5)} = \frac{5 \times 4 \times 3}{(N-2)(N-3)(N-4)} \quad \text{(V.1)}
\]

We can ask indeed whether such a probability distribution \(\mu\) exists, so that the Markov chain becomes reversible. Intuition tells us that in the long run we will not notice the arrow of time anymore — the system has equilibrated. The distribution of the number of particles in our vessel has then become stationary and giving rise to a time-reversal invariant dynamics. Some further thought suggests to try (for \(\mu\))

\[
\rho(x) = \frac{1}{2^N x!(N-x)!}, \quad x = 0, 1, 2, \ldots, N
\]

(fraction of subsets with \(x\) elements from a set with \(N\) elements) which gives equal probability to all particle configurations. In fact, this distribution is time-invariant (stationary) and exactly satisfies (V.1) above, and can be checked more generally to generate time-reversal invariance. In particular, it is easily checked that for all \(x = 1, \ldots, N-1,\)

\[
\rho(x) p(x, x \pm 1) = \rho(x \pm 1) p(x \pm 1, x)
\]

This \(\rho\) is called the equilibrium distribution. The Ehrenfest model of diffusion was proposed by Paul Ehrenfest to illustrate the relaxation to equilibrium and to understand the second law of thermodynamics.

### B. Stochastic matrix and Master equation

Assume that the state space has \(|K| = m\) elements. We say that an \(m \times m\)-matrix is a stochastic matrix \(P\) when all its elements \(p(x, y) \geq 0\) are non-negative and for each row \(\sum_y p(x, y) = 1\). That allows a probabilistic interpretation. Transition matrices specifying the updating rule for Markov chains are stochastic matrices, and their elements are the transition probabilities: \(p(x, y)\) is the transition probability to change to state \(y\) given that we now have state \(x\). That gives the building block for the probability law on discrete time trajectories: at every time \(n = 0, 1, \ldots\)

\[
\text{Prob}[x_{n+1} = y|x_n = x] = p(x, y), \quad x, y \in K \quad \text{(V.2)}
\]
and the Markov property can be expressed as saying

\[ \text{Prob}[x_{n+1} = y|x_n = x, x_{n-1} = a_{n-1}, x_{n-2} = a_{n-2}, x_0 = a_0] = p(x, y), \quad x, y \in K \quad (V.3) \]

no matter what earlier history \((a_0, a_1, \ldots, a_{n-1})\) up to time \(n-1\). In other words, a Markov process started with distribution \(\mu\) and with transition matrix \(P\) gives probability

\[ \text{Prob}_\mu[x_0 = a_0, x_1 = a_1, \ldots, x_{n-1} = a_{n-1}, x_n = a_n] = \mu(a_0) p(a_0, a_1) p(a_1, a_2) \ldots p(a_{n-1}, a_n) \]

to the trajectory \((a_0, a_1, \ldots, a_n) \in K^{n+1}\). Note the subscript \(\mu\) in the left-hand side indicating the initial distribution. Formulæ (V.2)–(V.3) are the defining properties of Markov chains.

Many properties follow. For example, we can ask what is the probability to find state \(x\) at time one if we started with distribution \(\mu\) at time zero. That is

\[ \text{Prob}_\mu[x_1 = x] = \sum_{a \in K} \text{Prob}_\mu[x_0 = a, x_1 = x] = \sum_{a \in K} \mu(a) p(a, x) \]

Let us call \(\mu_n\) the probability distribution at time \(n\); that is

\[ \mu_n(x) = \text{Prob}_\mu[x_n = x] \]

We have then shown that \(\mu_1(x) = \sum_{a \in K} \mu(a) p(a, x)\) where \(\mu = \mu_0\). Obviously, by the very same reasoning

\[ \mu_n(x) = \sum_{a \in K} \mu_{n-1}(a) p(a, x) \]

for all \(n = 1, 2, \ldots\). We can still rewrite that as

\[ \mu_n(x) - \mu_{n-1}(x) = \sum_{a \in K} [\mu_{n-1}(a) p(a, x) - \mu_{n-1}(x) p(x, a)] \quad (V.4) \]

which is called the **Master equation** for Markov chains. Observe that the change in probability (in the left-hand side) is given (in the right-hand side) by a source (the first) and a sink (the second term). It is What is written above for the evolution of probability distributions has an obvious dual for the evolution of observables. After all, the expected value of a function \(f\) at time \(n\) is

\[ \langle f(x_n) \rangle_\mu = \mu_n(f) = \sum_x f(x) \mu_n(x) = \sum_x f(x) \sum_{a \in K} \mu_{n-1}(a) p(a, x) \]
so that we can write

\[ \mu_n(f) = \sum_{a \in K} \mu_{n-1}(a) P f(a), \quad P f(a) = \sum_x p(a, x) f(x) \]

or, we could abbreviate

\[ \mu_n(f) = \mu_{n-1}(P f) \]

which is again equivalent with the Master equation above.

Note that we think of the new function \( P f \) as a column-vector obtained by multiplying the matrix \( P \) with the column-vector \( f \). In other words, observables (functions on \( K \)) can be written as column-vectors upon choosing an (arbitrary) order in \( K \). Thinking of probability distributions as row-vectors actually completes the picture: the expectation

\[ \mu(f) = \sum_x \mu(x) f(x) \]

then takes the form of a scalar product, multiplying the row-vector \( \mu \) with the column-vector \( f \). Never mind too much however that notation. The essential thing is that the transition probability \( P \) really determines everything. Its products \( P^n \) (under matrix multiplication) have also a definite meaning, as the matrix elements are

\[ (P^n)(x, y) = \text{Prob}[x_n = y| x_0 = x] \]

(see also the calculation below). Moreover,

\[ \mu_n(f) = \mu_{n-1}(P f) = \mu_{n-2}(P^2 f) = \ldots = \mu_0(P^n f) = \sum_x \mu(x) P^n f(x) \]

for \( \mu_0 = \mu \) (initially). In a way, that means that \( \mu_n = \mu P^n \), again identifying probability distributions with row-vectors. Clearly, if we could diagonalize the transition matrix \( P \), the Markov chain would be completely solved. (Time to get your linear algebra straightened.)
1. Calculation

Let us do $n = 2$. The probability to find $x_2 = y$ given that we started from $x_0 = x$ is

\[
\text{Prob}[x_2 = y|x_0 = x] = \sum_z \text{Prob}[x_2 = y, x_1 = z|x_0 = x] = \sum_z \frac{\text{Prob}[x_2 = y, x_1 = z, x_0 = x]}{\text{Prob}[x_0 = x]} \\
= \sum_z \text{Prob}[x_2 = y|x_1 = z, x_0 = x] \text{Prob}[x_1 = z, x_0 = x] \frac{1}{\text{Prob}[x_0 = x]} \\
= \sum_z \text{Prob}[x_2 = y|x_1 = z] \text{Prob}[x_1 = z|x_0 = x] = \sum_z p(z, y) p(x, z) = (P^2)(x, y)
\]

and we reached equality with the matrix elements of $P^2$.

2. Example

![Diagram](image)

FIG. 4: Three states and their connections.
Consider the $3 \times 3$ transition matrix

$$
P = \begin{pmatrix}
0 & 1 & 0 \\
0 & 1/2 & 1/2 \\
1/2 & 0 & 1/2
\end{pmatrix}
$$

for state space $K = \{1, 2, 3\}$, see Fig. 4. (Check that $P$ is a stochastic matrix.) We want to find the probability that at general time $n$ the state is 1 given that we started at time zero in 1. That is $P^n(1, 1)$. The best is to compute the eigenvalues of $P$. The characteristic equation is

$$
\det(\lambda - P) = 0 = \frac{1}{4}(\lambda - 1)(4\lambda^2 + 1)
$$

(We knew $\lambda = 1$ would be an eigenvalue, because $Pc = c$ for a constant vector). The eigenvalues are $1, i/2$ and $-i/2$. From linear algebra we thus have that

$$
P^n(1, 1) = a + b\left(\frac{i}{2}\right)^n + c\left(-\frac{i}{2}\right)^n
$$

(what would that become when some eigenvalue is repeated?) Since the answer must be real (a probability!), we can take

$$
P^n(1, 1) = \alpha + \frac{1}{2}n\left\{\beta \cos \frac{n\pi}{2} + \gamma \sin \frac{n\pi}{2}\right\}
$$

where the constants $\alpha, \beta, \gamma$ do not depend on $n$. We can compute by mere inspection that

$$
P^0(1, 1) = 1, P^1(1, 1) = 0, P^2(1, 1) = 0
$$

to find $\alpha = 1/5, \beta = 4/5, \gamma = -2/5$, which completely specifies $P^n(1, 1)$.

3. Time-correlations

So far we have been mostly interested in finding the distribution and expected values at a fixed time $n$. We are obviously also interested in time-correlations. That means to estimate the correlation and dependence between observations at various different times. Let us look at pair-correlations, say for observables $f, g$: for times $0 \leq u \leq n$,

$$
\langle g(x_n) f(x_u) | x_0 = x \rangle = \sum_{y,z} P^n(x, y) f(y) P^{n-u}(y, z) g(z)
$$
That follows from the Markov property, by thinking as follows
\[
\langle g(x_n) f(x_u) | x_0 = x \rangle = \sum_y \langle g(x_n) f(x_u) | x_u = y, x_0 = x \rangle \text{Prob}[x_u = y | x_0 = x] \quad \text{and}
\]
\[
\langle g(x_n) f(x_u) | x_u = y, x_0 = x \rangle = f(y) \langle g(x_n) | x_u = y \rangle
\]
\[
= f(y) \langle g(x_{n-u}) | x_0 = y \rangle
\]
The last equality uses time-homogeneity, the property that the updating-rule itself does not depend on time. The obtained formula can also be written as
\[
\langle g(x_n) f(x_u) \rangle_\mu = \mu(P^u(f P^{n-u}g))
\]
\[
= \mu_u(f P^{n-u}g)
\]
when starting at time zero from \(\mu\). In the same way we can treat more general correlation functions.

**C. Detailed balance and stationarity**

We say that a probability distribution \(\rho\) is time-invariant, or stationary when it does not change under the Markov evolution. In other words,
\[
\rho P = \rho
\]
or \(\rho\) is a left-eigenvector with eigenvalue 1 for \(P\). That means that \(\rho\) solves the stationary Master equation:
\[
\rho(x) = \sum_{a \in K} \rho(a) p(a,x), \quad \sum_{a \in K} [\rho(a) p(a,x) - \rho(x) p(x,a)] = 0
\]
(V.6)
If we start from \(\rho\) at time zero, we get \(\rho\) at time one, and we get \(\rho\) at all times, \(\rho_n = \rho\). That also makes the time-correlations stationary, i.e., by inspecting (V.5),
\[
\langle g(x_n) f(x_u) \rangle_\rho = \rho(f P^{n-u}g) = \langle g(x_{n-u}) f(x_0) \rangle_\rho
\]
(V.7)
only depending on the time-difference \(n - u\).

A special case of stationarity is equilibrium. We get it when each term separately in the second formula of (V.6) gives zero:
\[
\rho(a) p(a,x) - \rho(x) p(x,a) = 0, \quad a, x \in K
\]
(V.8)
That is very special, and is a strong requirement. In fact, it implies that you can reverse the order of time in (V.7) as then
\[ \rho(f P g) = \rho(g P f), \quad \langle g(x_n) f(x_u) \rangle_\rho = \langle g(x_0) f(x_{n-u}) \rangle_\rho \]

Of course, to check whether that holds does not truly require knowing the stationary distribution \( \rho \); here is how that goes. We say that the Markov chain satisfies the condition of detailed balance when there is a function \( V \) so that
\[ e^{-V(a)} p(a, x) = e^{-V(x)} p(x, a), \quad \forall a, x \in K \quad (V.9) \]

We call such a \( V \) a potential. Obviously, if detailed balance (V.9) holds, then
\[ \rho(x) = \frac{1}{Z} e^{-V(x)}, \quad x \in K \]
is stationary. In other words, if we can find a \( V \) solving (V.9), we know a stationary distribution. But it is not sure there is always such a potential \( V \). That relates to the following physical interpretation.

\section*{D. Time-reversal}

Suppose that \( \rho \) is a stationary distribution for the given transition probability matrix \( P \). Clearly then, it does not matter when we start the evolution, be it at time zero or at any other time \( n = -T \). In fact, we can now speak about the stationary Markov chain defined on (doubly-)infinite trajectories \( \omega = (a_n, n \in \mathbb{Z}) \). Any piece of such a trajectory has probability
\[ \Pr_\rho[x_{n_1} = a_1, x_{n_2} = a_2, \ldots, x_{n_k} = a_k] = \rho(a_1) P^{n_2-n_1}(a_1, a_2) \ldots P^{n_k-n_{k-1}}(a_{k-1}, a_k) \quad (V.10) \]
for any \( n_1 \leq n_2 \leq \ldots \leq n_k \) on \( \mathbb{Z} \). Of course that stationary stochastic process \( (x_n, n \in \mathbb{Z}) \) has marginal distribution, at each time \( n \), equal to \( \rho \):
\[ \Pr[x_n = a] = \rho(a), \quad a \in K, n \in \mathbb{Z} \]
That is what stationarity means really.
Let us then look at its time-reversal, the stochastic process \((y_n, n \in \mathbb{Z})\) defined from

\[ y_n = x_{-n} \]

It simply reverses the original trajectories. We could write down the probabilities

\[ \text{Prob}_p[y_{n_1} = a_1, y_{n_2} = a_2, \ldots, y_{n_k} = a_k] = \text{Prob}_p[x_{-n_1} = a_k, \ldots, x_{-n_2} = a_2, x_{-n_1} = a_1] \]

in general, just like in (V.10), but let us concentrate on two consecutive times. Say we ask how plausible it is to see in the time-reversed process the transition \(a \rightarrow b\) once we are in state \(a\):

\[ \text{Prob}_p[y_n = b|y_{n-1} = a] = \text{Prob}_p[x_{-n} = b|x_{-n+1} = a] \]

That asks for Bayes (III.1), and we continue

\[ \text{Prob}_p[y_n = b|y_{n-1} = a] = \text{Prob}_p[x_{-n} = b|x_{-n+1} = a] = p(b,a) \frac{\rho(b)}{\rho(a)} \]

An interesting conclusion follows: the transition probabilities for \(a \rightarrow b\) in the time-reversed process (left-hand side) are equal to the original one \(p(a,b)\) if there is detailed balance. Or, the stationary process is time-reversal invariant if and only if the process satisfies the condition of detailed balance (V.8)–(V.9). Detailed balance assures that in stationarity there is no arrow of time — that is why we then call that stationary distribution an equilibrium distribution. An equilibrium process (stationary -and- under detailed balance) is also called a reversible process.

**E. Relaxation**

At large times \(n\) we could hope that the distribution \(\mu_n\) becomes more constant. In a way, for very large \(n\), there should be little difference between \(P^n\) and \(P^{n+k}\) for any given \(k\). These things can be made precise. (We take here \(K\) finite.)

There is always a stationary distribution; since the column-vector of ones is an eigenvector with eigenvalue 1 for \(P\), then \(P\) must have a row-eigenvector with eigenvalue 1.

We say that the Markov chain is **probabilistically ergodic** when there is a probability distribution \(\rho\) such that for all initial distributions \(\mu = \mu_0\) the limit

\[ \lim_{n \uparrow +\infty} \mu_n = \rho \]
gives that $\rho$. Of course such $\rho$ is unique and is stationary. We could also write, equivalently, that

$$P^n f \rightarrow \rho(f) = \langle f \rangle_\rho$$

as time $n$ moves on, for all observables $f$. From here we see most clearly that this must be a property of the matrix $P$, and that linear algebra must be able to tell. That is right, and the theorem in algebra is that of Perron-Frobenius. We just describe the result here: the Markov chain is always probabilistically ergodic when the matrix $P$ is irreducible and aperiodic. Irreducible means that all states are eventually connected; you can reach all states from wherever in a finite time with positive probability. Aperiodicity on the other hand relates to the probability to return to the same state. For example, if we take $p = 1$ in the Ehrenfest-model we must wait at least two steps before we get back to the same state, and we cannot get back to the same state after an odd number of moves. It means that the Ehrenfest model with $p = 1$ is not aperiodic. But in general we do not worry too much about aperiodicity because when $P$ is an irreducible stochastic matrix, then for any $0 < p < 1$, the matrix $Q = pP + (1-p)I$ is stochastic, irreducible and aperiodic, and has the same stationary distribution as $P$. The matrix $Q$ is a lazy version of $P$ in the sense that now for sure there is the possibility to remain in the same state.

The relaxation to stationarity is exponentially fast for irreducible and aperiodic Markov chains. In other words, there is a typical time, the relaxation time, after which the initial data are essentially forgotten and a stationary regime is established. The fact that this relaxation is exponential is not so strange, because forgetting is multiplicative: at each step some information is lost and that accumulates in a multiplicative way; you even forget what you forgot.

For example, if we look back at time-correlations (V.7), we can say that for large enough times $n$

$$\langle g(x_n) f(x_0) \rangle_\rho \simeq \rho(g) \rho(f)$$

Another property of relaxation is that there is a function, called relative entropy, that is monotone in time:

$$s(\mu_n | \rho) = \sum_x \mu_n(x) \log \frac{\mu_n(x)}{\rho(x)} \geq 0$$

turns out to decay monotonically to zero as time $n$ runs. Of course, if you do not know $\rho$
you do not know that relative entropy. Therefore, this monotonicity is most explicitly useful when there is detailed balance (in which case physicists call it a sort of H-theorem).

F. Random walks

A very interesting example of Markov chains are random walks. There the state space contains the possible positions on a lattice or graph, with the edges between them indicating the possible moves. The transition matrix specifies the weights associated to each move. As a standard example we can consider the (standard) random walk on the integers (onedimensional lattice). The state space is $K = \mathbb{Z}$ (and there could be no stationary distribution). At time zero we start at some site $x_0 = x$ and the position at time $n \geq 1$ is

$$x_n = x_0 + v_1 + v_2 + \ldots v_n$$

where the $v_i$ are a collection of independent and identically distributed random variables, say $v_i = 0$ with probability $1 - p$ and $v_i = \pm 1$ with probability $p/2$ (for some $p \in (0,1)$). From here we can calculate the mean position at time $n$ (no net drift here) and its variance (proportional to time $n$).

![FIG. 5: A random walk in 2 dimensions, from close-by and from further away.](image)

That gives a simple model of diffusion. Let us see if we can find the diffusion equation. For this we look at the Master equation

$$\mu_n(x) - \mu_{n-1}(x) = \frac{p}{2}\mu_{n-1}(x-1) + \frac{p}{2}\mu_{n-1}(x+1) - p\mu_{n-1}(x)$$

which can be rewritten, suggestively, via a discrete Laplacian (in its right-hand side)

$$\mu_n(x) - \mu_{n-1}(x) = \frac{p}{2}[\mu_{n-1}(x-1) + \mu_{n-1}(x+1) - 2\mu_{n-1}(x)]$$
That indeed resembles a (discrete) diffusion equation with diffusion coefficient $p/2$. We imagine that $\mu(x)$ then corresponds to a density of independent walkers.

The same set-up can be used in any dimension, on $\mathbb{Z}^d$ for the $d-$dimensional regular lattice. The behavior can however be drastically different depending on the dimension. In one and two dimensions the probability of (ever) returning to the origin is one. That return probability decreases as the number of dimensions increases: for $d = 3$ the probability decreases to roughly 34 percent; in $d = 8$ that return probability is about 7 percent. As you can perhaps imagine, that very different behavior of the standard random walk depending on the dimension (recurrence versus transience) is at the origin of many physical effects. Or better, it summarizes and stochastically interprets the behavior of the Green’s function of the Laplacian, which is of course relevant to quite many physical contexts (electromagnetism and gravity, Bose-Einstein condensation,...)

G. Hitting probability [optional]

It is often important to estimate the probability to ever land in a certain state. Such states then have a special importance; for example, being in that state could stop the evolution. We call such states absorbing. Consider thus a random walk on the semi-infinite lattice $K = \{0, 1, 2, \ldots \}$ with transition probabilities $p(x, x + 1) = p_x, p(x, x - 1) = q_x, p_x + q_x = 1$ for $x \geq 1$ and $p(0,0) = 1$. The state zero is absorbing; the chain gets extinct when hitting it. We want to calculate the hitting probability $h_x$, i.e., the extinction probability starting from state $x$:

$$h_x := \text{Prob}[X_n = 0 \text{ for some } n > 0 | X_0 = x]$$

That satisfies a recurrence relation

$$h_0 = 1, \quad h_x = p_x h_{x+1} + q_x h_{x-1} \text{ for } x = 1, 2, \ldots$$

There could be more than one solution, but one can prove that the sought for $h_x$ is the minimal solution. Consider now $u_x := h_{x-1} - h_x$ for which the recurrence becomes $p_x u_{x+1} = q_x u_x$. Hence,

$$u_{x+1} = \left( \frac{q_x}{p_x} \right) u_x = \left( \frac{q_x q_{x-1} \ldots q_1}{p_x p_{x-1} \ldots p_1} \right) u_1$$
Combining that with \( u_1 + \ldots u_x = h_0 - h_x \) we get
\[
h_x = 1 - u_1(\gamma_0 + \ldots + \gamma_{x-1})
\]
for \( \gamma_x := \frac{q_x q_{x-1} \ldots q_1}{p_x p_{x-1} \ldots p_1}, x \geq 1 \) and \( \gamma_0 = 1. \) So we only need \( u_1. \) Suppose now first that the infinite sum \( \sum \gamma_x = \infty. \) Since \( h_x \in [0,1] \) we must then have \( u_1 = 0, h_x = 1 \) for all \( x. \) On the other hand, if \( \sum \gamma_x < \infty, \) then we can take \( u_1 > 0 \) so long as
\[
1 - u_1(\gamma_0 + \ldots + \gamma_{x-1}) \geq 0
\]
The minimal non-negative solution occurs for \( u_1 = \left( \sum \gamma_x \right)^{-1} \) and then
\[
h_x = \frac{\sum_{y=x}^{\infty} \gamma_y}{\sum_{y=0}^{\infty} \gamma_y}
\]
which is strictly less than one for all \( x \neq 0. \)

H. Example: a periodic Markov chain

The following is a paradoxical game invented by Juan Parrondo (1996); see http://www.eleceng.adelaide.edu.au/Groups/parrondo/intro.html for more explanations.

The state space is \( K = \{1,2,3\} \) and the state at time \( n \) is \( x_n. \) The Markov chain uses a different rule (A or B) at even and at odd times \( n. \) Alternatingly, the following two games are played. Game A is fair coin tossing: we simply move \( x \to x \pm 1 \mod 3 \) with equal probability at even times. Game B is played at odd times and with two biased coins, a good one and a bad one. In game B, the good coin is tossed when \( x_n \in \{1,2\} \) and the bad coin is used each time when \( x_n = 3. \) Winning takes \( x_{n+1} = x_n + 1; \) losing at time \( n \) means \( x_{n+1} = x_n - 1, \) always modulo 3. The transition probabilities are then
\[
\text{Prob}[x_{n+1} = x \pm 1 | x_n = x] = 1/2, \quad \text{when } n \text{ is even}
\]
\[
\text{Prob}[x_{n+1} = x + 1 | x_n = x] = 3/4, \quad \text{when } n \text{ is odd and } x \neq 3
\]
\[
\text{Prob}[x_{n+1} = x + 1 | x_n = x] = 1/10, \quad \text{when } n \text{ is odd and } x = 3 \quad \text{(V.12)}
\]
Let us check detailed balance when we would only play game B (at all times):
Consider the cycle \( 3 \to 1 \to 2 \to 3. \) Its stationary probability (always for game B alone) is \( \text{Prob}[3 \to 1 \to 2 \to 3] = \rho(3) \times 1/10 \times 3/4 \times 3/4 = 9\rho(0)/160. \) For the reversed cycle,
the probability \( \text{Prob}[3 \rightarrow 2 \rightarrow 1 \rightarrow 3] = \rho(3) \times 9/10 \times 1/4 \times 1/4 = 9\rho(3)/160 \) is the same. The equilibrium distribution for game \( B \) is then found to be \( \rho(1) = 2/13, \rho(2) = 6/13 \) and \( \rho(3) = 5/13 \). Obviously then, there is no current when playing game \( B \) and clearly, the same is trivially verified for game \( A \) when tossing with the fair coin. Yet, and here is the paradox, when playing periodically game \( B \) after game \( A \), a current arises... ( ...which you would like to check).

I. Example: one-dimensional Ising model

We consider the Markov chain on \( K = \{+1, -1\} \) with transition probability

\[
p(x, y) = \frac{1}{Z} e^{Jxy+ay+bx}
\]

where \( b \) and \( Z \) take care of the normalization, \( \sum_y p(x, y) = 1 \). For example, we can take

\[
P = \begin{pmatrix}
  e^{J+a} & e^{-J-a} \\
  2\cosh(J+a) & 2\cosh(J+a) \\
  e^{J-a} & e^{-J-a} \\
  2\cosh(J-a) & 2\cosh(J-a)
\end{pmatrix}
\]

in which case

\[
b = \frac{1}{2} \log \frac{\cosh(J - a)}{\cosh(J + a)}, \quad Z = \frac{2 \cosh b}{\cosh(J + a) + \cosh(J - a)}
\]

Clearly then, the probability of a trajectory \((x_0, x_1, \ldots, x_n)\) gives

\[
\frac{\mu(x_0)}{Z^n} \exp\{Jx_0x_1 + Jx_1x_2 + \ldots + Jx_{n-1}x_n + (a + b)(x_1 + x_2 + \ldots + x_n) + b(x_0 - x_n)\}
\]

which is up to boundary conditions the probability of a spin configuration in the one-dimensional Ising model (in a magnetic field \( a + b \)) with lattice sites replacing discrete time. Indeed, that Ising model is traditionally solved using the transfer matrix formalism, which is equivalent to the formalism of Markov chains.

![FIG. 6: Up and down spins in the one-dimensional Ising model.](image)

No thermal phase transition in one dimension for systems with short range interactions is the same as saying that Markov chains are probabilistically ergodic. The mathematical ground is the Perron-Frobenius theorem.
J. Exercises

1. [See the free e-book by Grinstead and Snell, from http://www.dartmouth.edu/]
According to Kemeny, Snell, and Thompson the Land of Oz is blessed by many things, but
not by good weather. They never have two nice days in a row. If they have a nice day, they
are just as likely to have snow as rain the next day. If they have snow or rain, they have
an even chance of having the same the next day. If there is change from snow or rain, only
half of the time is this a change to a nice day. With this information, model/construct a
Markov chain — write down the state space and the transition matrix.

2. Consider a discrete time Markov chain on state space $K = \{+1, 0, -1\}$. The transition
matrix has $p(x, x) = 0$ and
\[
p(-1, 0) = p(0, 1) = p(1, -1) = p \\
p(0, -1) = p(1, 0) = p(-1, 1) = 1 - p
\]
for parameter $p \in [0, 1]$.
Determine the stationary distribution.
What is the transition matrix for the time-reversed process in that stationary distribution?
For what value(s) $p$ is there detailed balance?

3. Show that the Ehrenfest model satisfies detailed balance, and find the potential.
Show that all Markov chains with two states, $|K| = 2$, satisfy detailed balance, at least
when the $p(x, y) > 0$.

4. Consider a container with green and red balls, in total $N$. At discrete times and
blindly two balls are picked out and we look at their color. They are then put back in the
container after we have changed their color (green becomes red and red becomes green).
Model this with a Markov chain, and write down the transition matrix. What could be the
stationary distribution?

5. Show that for detailed balance (V.9) to hold, we must have that for any three states
\[ x, y, z \]
\[ p(x, y)p(y, z)p(z, x) = p(x, z)p(z, y)p(y, x) \]

or, the probability of any cycle/loop should not depend on the order/orientation in which it is being traversed.

6. Carefully understand and prove formula (V.10).

7. Consider the most general two-state Markov chain (discrete time) and compute the \( n \)-th power \( P^n \) of its transition matrix.
Discuss when and how it converges, as \( n \uparrow +\infty \) via integers, to the stationary distribution.

8. Take the \( 3 \times 3 \) transition matrix
\[
\begin{pmatrix}
\frac{1}{3} & 1/6 & 1/2 \\
1/4 & 1/3 & 5/12 \\
1/2 & 1/4 & 1/4
\end{pmatrix}
\]
and consider the initial distribution \( \mu_0 = (1/10, 2/5, 1/2) \). Find the probability law \( \mu_1 \) at the next time. Find also the stationary distribution.

9. Consider the probability distribution \( \rho = (1/6, 1/2, 1/3) \) on \( K = \{-1, 0, 1\} \). Find a Markov chain on \( K \) which makes \( \rho \) stationary.

10. Show that the product of two stochastic matrices is again stochastic.

11. Which of the following, when stationary, are reversible Markov chains?
   a) The two-state chain having transition matrix \( \begin{pmatrix} 1 - \alpha & \alpha \\ \beta & 1 - \beta \end{pmatrix} \) where \( \alpha + \beta > 0 \).
   b) The chain having transition matrix \( \begin{pmatrix} 0 & p & 1-p \\ 1-p & 0 & p \\ p & 1-p & 0 \end{pmatrix} \) where \( 0 < p < 1 \).
12. Find the $n$-step transition probabilities $P^n(x, y)$ for the chain having transition matrix
\[
\begin{pmatrix}
0 & 1/2 & 1/2 \\
1/3 & 1/4 & 5/12 \\
2/3 & 1/4 & 1/12
\end{pmatrix}
\]
Compare $\lim_{n \to \infty} P^n(x, y)$ with the stationary distribution.

13. Consider the Markov chain on $\{0, 1, \ldots \}$ with transition probabilities given by
\[
p(0, 1) = 1, \quad p(x, x+1) + p(x, x-1) = 1, \quad p(x, x+1) = \left(\frac{x + 1}{x}\right)^2 p(x, x-1), \quad x \geq 1
\]
Show that if initially $x_0 = 0$ then the probability that $x_n \geq 1$ for all $n \geq 1$ is $6/\pi^2$. (Hint: use formula (V.11).)

14. Consider the transition matrix
\[
\begin{pmatrix}
1/2 & 1/2 & 0 & 0 & 0 \\
1/2 & 0 & 1/2 & 0 & 0 \\
1/2 & 0 & 0 & 1/2 & 0 \\
0 & 1/2 & 0 & 0 & 1/2 \\
0 & 0 & 1/2 & 0 & 1/2
\end{pmatrix}
\]
Find the stationary distribution. (Note the first five Fibonacci numbers.)

15. Lady Ann possesses $r$ umbrellas which she employs in going from home to office and back. If she is at home (resp. office) at the beginning (resp. end) of a day and it is raining, then she will take an umbrella with her to the office (resp. home), at least if there is one to be taken. If it is not raining, then she will not take an umbrella. Assuming that, independent of the past, it rains at the beginning (end) of a day with probability $p$, what fraction of the time does Lady Ann arrive soaked at the office?

16. Markov found the following empirical rule for the transition matrix in the vowel-consonant space in Pushkin’s novel:
\[
\begin{pmatrix}
0.128 & 0.872 \\
0.663 & 0.337
\end{pmatrix}
\]
Show that that is consistent with the vowel versus consonant frequency of $(0.432, 0.568)$. 
VI. MARKOV JUMP PROCESSES — CONTINUOUS TIME

We can embed a discrete time process, such as the Markov chains above, in continuous time. What then seems essential for discrete time processes is that the time step is always the same and independent of the state the system is in. In other words, one would say that the time between jumps is deterministic and fixed, no matter what state. One should add here however that the system can remain in the same state, i.e., the jumps can be to itself, or there is no real jump to another state. That implies that the state can remain identical over several time-steps. How long the system remains in that same state depends on the total probability to jump away from the state. Going to continuous time processes means to randomize these true waiting times between true jumps. The jump alarm rings at exponential times, after which the state changes.

A. Examples

**Example 1: The Poisson Process.** Events that occur independently with some average rate are modeled with a Poisson process. This is a continuous-time Markov process with state space $K = \{0, 1, 2, \ldots\}$. The states count the number of arrivals, successes, occurrences etc. If at any time we are in state $x$ we can only jump to state $x + 1$; there is only one exponential clock running with fixed rate $\xi$ (the intensity). We start at time zero from $x_0 = 0$ and inspect the state at time $t \geq 0$. The probability that $x_t = k$ is the probability that the clock has rung exactly $k$ times before time $t$. Not surprisingly, that is given by the Poisson distribution

$$\text{Prob}[x_t = k] = e^{-\xi t} \left(\xi t\right)^k \frac{1}{k!}$$

because the probability of a ring in $dt$ is $\xi dt$ and we try that $t/dt$ times, cf. the convergence of the binomial to the Poisson distribution. The (waiting) times between the events/arrivals have an exponential distribution, like the limit of a geometric distribution. In that way the Poisson process captures complete randomness of arrivals, as in radioactive decay.

**Example 2: Two-state jumps.** We have two possible states, $K = \{0, 1\}$. When the state is zero, we wait an exponentially distributed time with rate $\xi(0) > 0$ after which the jump $0 \rightarrow 1$ occurs. When the state is one, we wait an exponentially distributed time with
rate $\xi(1) > 0$ after which the jump $1 \rightarrow 0$ occurs. Trajectories are piecewise constant and switch between 0 and 1. In the (long time) stationary regime, the relative frequency of one over zero is $\xi(0)/\xi(1)$.

We can obtain this continuous time jump process from a discrete time approximation. Consider the Markov chain with transition matrix

$$P_\delta = \begin{pmatrix} 1 - \delta \xi(0) & \xi(0) \delta \\ \xi(1) \delta & 1 - \delta \xi(1) \end{pmatrix}$$

parameterized by small $\delta > 0$. Clearly, $P_\delta = I + \delta L$ with $I$ the unit matrix and

$$L = \begin{pmatrix} -\xi(0) & \xi(0) \\ \xi(1) & -\xi(1) \end{pmatrix}$$

Therefore, after $n$ discrete steps, by the definition of the exponential function,

$$(P_\delta)^n = (I + \delta L)^n \rightarrow e^{tL}$$

when $n\delta = 1$ for $n \uparrow +\infty$, $\delta \downarrow 0$. Hence, the matrix elements $e^{tL}(x,y)$ must give the transition probability to go from $x$ to $y$ in time $t$. The rates of change are read from the off-diagonal elements of $L$.

**B. Path-space distribution**

A continuous time Markov process on state space $K$ has piecewise constant trajectories $\omega = (x_t, t \geq 0)$ which are specified by giving the sequence $(x_0, x_1, x_2, \ldots)$ of states together with the jump times $(t_1, t_2, \ldots)$ at which the jumps, respectively, $x_0 \rightarrow x_1, x_1 \rightarrow x_2, \ldots$ occur. We take the convention that $x_t$ is right-continuous, so that $x_{t_i+\varepsilon} = x_{t_i}$ for sufficiently small $\varepsilon$, while $x_{t_i+\varepsilon} = x_{t_i}$ for all $0 < \varepsilon < t_{i+1} - t_i$, no matter how small.

To give the probability distribution over these possible paths means to give the distribution of waiting times $t_{i+1} - t_i$ when in state $x$, and to give the jump probabilities $x \rightarrow y$ when a jump actually occurs. For the first (the waiting times) we take an exponential distribution with rate $\xi(x)$, i.e., when in state $x$ at the jump time $t_i$, the time to the next jump is exponentially distributed as

$$\text{Prob}[t_{i+1} - t_i \in [s, s + ds] | x_{t_i} = x] = \xi(x) e^{-\xi(x)s} ds$$
Secondly, at the jump time $t_{i+1}$ the jump goes $x \rightarrow y$ with probability $p(x, y)$, for which we assume that $p(x, x) = 0$. We thus get

$$
\xi(x_0)e^{-\xi(x_0)t_1}p(x_0, x_1)\xi(x_1)e^{-\xi(x_1)(t_2-t_1)}p(x_1, x_2) \ldots
$$

as probability density on path space $\Omega$.

![Figure 7: A continuous time random walk on $\mathbb{Z}$](image)

The product

$$
k(x, y) = \xi(x)p(x, y), \quad \text{with then } \xi(x) = \sum_y k(x, y)
$$

are called the transition rates for the jumps $x \rightarrow y$. We can thus say that a path $\omega = (x_0, x_1, \ldots, x_n)$ over the time interval $[0, T]$ has probability density

$$
k(x_0, x_1)k(x_1, x_2) \ldots k(x_{n-1}, x_n) \exp\left\{-\int_0^T \xi(x_s)ds\right\}
$$

(VI.1)

The last integrand is

$$
\int_0^T \xi(x_s)ds = \xi(x_0)t_1 + \xi(x_1)(t_2-t_1) + \ldots + \xi(x_n)(T-t_n)
$$

for jump times $t_1 < t_2 < \ldots < t_n < T$. These $\xi$'s are called escape rates, quite appropriately.

The fact that the waiting times are exponentially distributed is equivalent with the Markov property, as we can see from the following argument.

Call $\tau$ the waiting time between two jumps, say while the state is $x$. For $s \geq 0$ the event
\( \tau > s \) is equivalent to the event \( \{ x_u = x \text{ for } 0 \leq u \leq s \} \). Similarly, for \( s, t \geq 0 \) the event \( \tau > s + t \) is equivalent to the event \( \{ x_u = x \text{ for } 0 \leq u \leq s + t \} \). Therefore,

\[
P[\tau > s + t | \tau > s] = P[x_u = x \text{ for } 0 \leq u \leq s + t | x_u = x \text{ for } 0 \leq u \leq s] \\
= P[x_u = x \text{ for } s \leq u \leq s + t | x_u = x \text{ for } 0 \leq u \leq s] \\
= P[x_u = x \text{ for } s \leq u \leq s + t | x_s = x] \\
= P[x_u = x \text{ for } 0 \leq u \leq t | x_0 = x] \\
= P[\tau > t]
\]

Thus, the distribution of \( \tau \) is memoryless, which means that it is exponential.

### C. Generator and semigroup

The previous section provides a probabilistic construction of continuous time Markov processes. The result is a probability distribution on path space. Nevertheless, for many calculations that path-space law is not very useful. A more analytic approach starts from the idea of transition matrix as we had it for Markov chains.

A continuous time Markov process is specified by giving the transition rates \( k(x, y) \geq 0 \) between \( x, y \in K \). They define the escape rates \( \xi(x) = \sum_y k(x, y) \). From these we make the so called (backward) generator \( L \) which is a matrix with elements

\[
L(x, y) = k(x, y), \quad x \neq y \\
L(x, x) = -\xi(x) \quad (\text{VI.2})
\]

For each row, \( \sum_y L(x, y) = 0 \). It acts on observables (column-vectors) \( f \) as

\[
L f(x) = \sum_y k(x, y) [f(y) - f(x)]
\]

Look at the structure: \( L f \) is like the change in the observable \( f \), multiplied with the rate of that change.

From there we make the semigroup \( S(t) = \exp(tL), t \geq 0 \). That semigroup takes over the role of the transition matrix, in the sense that we have

\[
S(t)f(x) = \langle f(x_t) | x_0 = x \rangle, \quad \mu_t = \mu S(t)
\]
The right-hand claim is equivalent to noting

$$\mu(S(t)f) = \langle f(x_t) \rangle_\mu = \mu_t(f)$$

Of course we can check that it corresponds to the limit of discrete time Markov chains when we construct the lazy process $$P_\delta = (1 - \delta)I + \delta P$$ and take $$(P_\delta)^n \to S(t)$$ for $$n\delta = t$$ with $$n \uparrow +\infty$$, $$\delta \downarrow 0$$, so that the path space distribution of the previous section is entirely compatible. But now we get many more analytic and algebraic tools, and there is no need to refer to discrete time at all. The main reason is that the generator $$L$$ has the structure (VI.2) which makes $$S(t)$$ a stochastic matrix for all times $$t$$.

D. Master equation, stationarity, detailed balance

The evolution of probability distributions follows the Master equation

$$\frac{d}{dt} \mu_t(x) + \sum_y j_{\mu_t}(x, y) = 0, \quad x \in K$$ (VI.3)

with probability currents

$$j_{\mu}(x, y) = \mu(x)k(x, y) - \mu(y)k(y, x), \quad x, y \in K$$

Again, this has the form of a balance equation: the probability at $$x$$ grows by jumps $$y \to x$$ and decreases by the jumps $$x \to y$$.

Obviously the Master equation (VI.3) is nothing more than writing out $$\mu_t = \mu S(t)$$ in differential form

$$\frac{d}{dt} \mu_t = \mu_t L = \mu S(t)L = \mu LS(t)$$

We say that a probability law $$\rho$$ is stationary when $$\rho L = 0$$, or

$$\sum_y j_{\rho}(x, y) = \sum_y [\rho(x)k(x, y) - \rho(y)k(y, x)] = 0$$

That means $$\rho S(t) = \rho$$. Then, for all observables $$f$$

$$\rho(Lf) = \langle Lf \rangle_\rho = \sum_x \rho(x)Lf(x) = 0$$

We say the dynamics satisfies the condition of detailed balance when there is a potential $$V$$ such that

$$k(x, y)e^{-V(x)} = k(y, x)e^{-V(y)}$$
Then, for all triples \( x, y, z \in K \), \( k(x, y)k(y, z)k(z, x) = k(x, z)k(z, y)k(y, x) \), and also, under detailed balance,
\[
\rho(x) = \frac{1}{Z} e^{-V(x)}, \quad Z = \sum_x e^{-V(x)}
\]
is stationary; we call it the equilibrium distribution. The condition of detailed balance is also here equivalent with time-reversibility: the equilibrium process is reversible.

When asked to find the explicit time-evolution of a Markov jump process, we must solve the Master equation (VI.3). That is a collection of linear coupled first order differential equations for the \( \mu_t(x) \). When there are few states, like \(|K| = 2\) or \(|K| = 3\), or when there are special symmetries, we can solve it almost directly by also using the normalization \( \sum_x \mu_t(x) = 1 \). In the more general case we need to diagonalize the generator \( L \), which of course also diagonalizes \( S(t) \).

When asked to find the stationary distribution we better first check whether there is perhaps detailed balance. If not, we must find the (left) row-eigenvector of \( L \) with eigenvalue zero, which means to solve the stationary Master equation.

E. Example: two state Markov process

Let \( K = \{-1, +1\} \), and transition rates \( k(-1, +1) = \alpha, k(+1, -1) = \beta \). Given the initial state \( x_0 = +1 \), find the probability for \( x_t = +1 \), written \( P[x_t = +1] \).

The Master equation gives
\[
\frac{dP[x_t = +1]}{dt} = \alpha(1 - P[x_t = +1]) - \beta P[x_t = +1] = \alpha - (\alpha + \beta)P[x_t = +1]
\]

Solving this differential equation and using the initial condition \( P[x_0 = +1] = 1 \) we find
\[
P[x_t = +1] = \frac{\alpha}{\alpha + \beta} + \left(1 - \frac{\alpha}{\alpha + \beta}\right) \exp[-(\alpha + \beta)t]
\]

Check that the initial condition has been implemented and that
\[
\lim_{t \to \infty} P(x_t = +1) = \frac{\alpha}{\alpha + \beta}
\]
which gives the equilibrium distribution (solution of the stationary Master equation, with detailed balance).
F. Exercises

1. Show that \( S(t) \) is a stochastic matrix.
Show that \( S(0) = I \) and \( S(t + s) = S(t)S(s) \).
Show that
\[
\frac{d}{dt} \langle f(x_t) \rangle = \langle Lf(x_t) \rangle
\]

2. Check the H-theorem (monotonicity of the relative entropy) for two state Markov processes.

3. Consider the following continuous time Markov process. The state space is \( K = \{0, +2, -2\} \) and the transition rates are \( k(0, +2) = \exp[-b], k(0, -2) = \exp[-a], k(-2, +2) = k(+2, -2) = 0, k(+2, 0) = \exp[b - h], k(-2, 0) = \exp[a + h] \)
Determine the stationary distribution. That asks for the time-invariant state occupation. Is there detailed balance (or, for what values of the parameters \( a, b, h \))?

4. We consider the following continuous time Markov process \( (x_t) \). The state space is \( K = \{-1, 1\} \), say up and down. The transition rates are specified via parameters \( v, H > 0 \):
\[
k(-1, 1) = v, \quad k(1, -1) = H/v
\]
We choose the initial condition \( x_0 = +1 \) after which the random trajectory \( (x_t) \) develops. Compute the expectation value of \( \exp x_t \),
\[
\langle \exp x_t | x_0 = 1 \rangle
\]
for arbitrary times \( t \geq 0 \), as a function of \( v \) and of \( H \).

5. Consider a continuous time Markov process with state space \( K = \{1, 2, \ldots, N\} \) and with transition rates
\[
k(x, x + 1) = p \text{ except for } x = N, \quad k(x, x - 1) = q \text{ except for } x = 1
\]
All other transition rates are zero. Determine the stationary distribution.
6. We wish to make a continuous time Markov process that relaxes to an equilibrium probability distribution
\[ \rho(x) = \frac{1}{Z} \exp(-\beta E(x)) \]
for some parameter \( \beta > 0 \), normalization \( Z \) and energy function \( E(x) \) for states \( x \in \{1, 2, \ldots, N\} \). Write down explicit transition rates \( k(x, y) \) that define a process so that \( \rho \) solves the stationary Master equation. That evolution would simulate \( \rho \).

7. Imagine that \( L \) is the generator of a continuous time Markov process with a finite state space.
   a) Describe an explicit example for the case of three states where the stationary distribution is not uniform and where there is no detailed balance. Give all details including the explicit form of the transition rates and the stationary distribution.
   b) Write \( \rho \) for the stationary distribution and suppose that \( \rho(x) \neq 0 \) for all states \( x \). Show that if the matrix \( H \) with elements
\[ H_{xy} = \sqrt{\rho(x)} L_{xy} \frac{1}{\sqrt{\rho(y)}} \]
is symmetric, that then detailed balance holds. (The choice of the letter \( H \) points to a symmetric Hamiltonian for quantum evolutions.)

8. Consider a network with four states \((x, v)\) where \( x \in \{0, 1\}, v \in \{-1, +1\}\). (Imagine \( x \) to be a position and \( v \) like a velocity.) We define a Markov process in continuous time via transition rates that depend on parameter \( b > 0 \),
\[ k((1, +1), (1, -1)) = k((1, -1), (1, +1)) = k((0, +1), (0, -1)) = k((0, -1), (0, +1)) = 1 \]
\[ k((1, -1), (0, -1)) = k((0, +1), (1, +1)) = b \]
All other transitions are forbidden. Make a drawing. Determine the stationary distribution on the four states as function of \( b \). Is there detailed balance?

9. Show that a continuous time Markov process on a finite set \( K \) satisfies the condition of detailed balance if and only if
\[ k(x, y) = \psi(x, y) e^{\frac{1}{2}[V(x)-V(y)]} \]
for some symmetric $\psi(x, y) = \psi(y, x)$ and potential $V$.

10. Show that for all observables $f$,

$$L(f^2) \geq 2 f L f$$

for the generator $L$ of a Markov process. That property is sometimes called dissipative, in contrast to a true derivation where the equality holds.

Show also that the differential operator

$$L f(x) = F(x) f'(x) + f''(x), \quad x \in \mathbb{R}$$

for all real functions $F$, is dissipative on smooth functions $f : \mathbb{R} \to \mathbb{R}$.

11. A radioactive source is measured by a Geiger counter. The distribution of detected particles obeys a Poisson distribution, i.e., with $N(t)$ the total number of detections after some time $t$:

$$\Pr[N(t) = j] = \frac{\lambda^j t^j}{j!} e^{-\lambda t}, \quad \text{with } j = 0, 1, 2, \ldots.$$ 

Now define the arrival times $T_n$ as

$$T_0 = 0, \quad T_n = \inf\{t : N(t) = n\}.$$ 

Show that the inter-arrival times $x_n := T_n - T_{n-1}$ are independent random variables with exponential distribution with parameter $\lambda$. [See the figure below for a typical realization of the experiment.]
12. Calculate $p_t(x,y)$ (transition probability over time $t$) for the continuous time Markov process on $N + 1$ states with the following diagram:

What is the stationary distribution?

13. Consider the three-state process with transition rates defined in the following diagram:

a) Write down the generator $L$.

b) Solve the equation $\rho L = 0$, to find the stationary distribution $\rho$.

c) Check the detailed balance condition. Explain your findings.

14. Let $\lambda, \mu > 0$ and consider the Markov process on $\{1, 2\}$ with generator

$$L = \begin{pmatrix} -\mu & \mu \\ \lambda & -\lambda \end{pmatrix}$$

a) Calculate $L^n$ and sum $\sum_{n=0}^{\infty} t^n/n! L^n$. Compare your answer with the matrix $\exp tL$.

b) Solve the equation $\rho L = 0$, to find the stationary distribution. Verify that $p_t(x,y) \to \rho(y)$ as $t \uparrow +\infty$.

15. Consider a three-level atom, in which each level has (a different) energy $E_x$ with $x = 1, 2, 3$. Now suppose that this atom is in contact with a thermal reservoir at inverse temperature $\beta$ such that the system’s energy jumps between two neighbouring levels $(1 \leftrightarrow 2$ or $2 \leftrightarrow 3)$ at a rate $k(x,y) = (E_x - E_y)^4 \exp[-\beta (E_y - E_x)/2]$, with $|x - y| = 1$.

a) Write down the generator, as it acts on the function/observable that expresses the population of the middle level.

b) Is there detailed balance? Find the stationary distribution.
VII. ON THE PHYSICAL ORIGIN OF JUMP PROCESSES

We have seen in the previous sections how jump processes can stand quite naturally as model for various phenomena. For example, descriptions in terms of random walks are ubiquitous and appear as good and relevant models for a great variety of processes in human, natural and mathematical sciences. In other cases, we readily feel that this or that model can be a useful description within chemical kinetics, or in population dynamics etc. Physicists often want more however than effective modeling or well-appreciated descriptions. In physics we also want to understand the physical origin and limitations of a model. We want to evaluate the model also with respect to more fundamental laws or equations. It is therefore interesting to spend some time on the question where jump processes could physically originate. We discuss below two possible origins: 1) via Fermi golden rule — so called, Van Hove weak coupling limit, 2) within Kramers’ theory of reaction rates.

A. Weak coupling regime

B. Reaction rate theory
Markov diffusion processes

VIII. BROWNIAN MOTION

There is a long and well-documented history of Brownian motion; see for example http://www.physik.uni-augsburg.de/theo1/hanggi/Duplantier.pdf. While observed by various people before (such as in 1785 by Jan Ingenhousz), the name refers to the systematic studies of Robert Brown (1773-1858) who observed the irregular, quite unpredictable, unhalted and universal motion of small particles suspended in fluids at rest. The motion becomes more erratic as the particle gets smaller, as the temperature gets larger or for lower viscosity. There is of course no vitalist source; the motion is finally caused by the collisions of the particle with the fluid particles. What we observe is a reduced dynamics, having no direct access to the mechanics of the fluid particles. In that sense the theory of Brownian motion provides a microscope to molecular motion. It was thus a convincing and an important ingredient in the understanding of the corpuscular nature of matter, say the atomic hypothesis:

I am now convinced that we have recently become possessed of experimental evidence of the discrete or grained nature of matter, which the atomic hypothesis sought in vain for hundreds and thousands of years. The isolation and counting of gaseous ions, on the one hand, which have crowned with success the long and brilliant researches of J.J. Thomson, and, on the other, agreement of the Brownian movement with the requirements of the kinetic hypothesis, established by many investigators and most conclusively by J. Perrin, justify the most cautious scientist in now speaking of the experimental proof of the atomic nature of matter. Wilhelm Ostwald, Grundriss der allgemeinen Chemie (4th ed., 1909)

Jumping ahead, a cloud of spherically shaped Brownian particles of diameter \( d \) (or a repeated measurement of the position on a single particle) in a fluid with viscosity \( \eta \) satisfies the diffusion equation

\[
\frac{\partial}{\partial t} \rho(x, t) = D \Delta \rho(x, t) \tag{VIII.1}
\]
with diffusion constant
\[ D = \frac{k_B T}{6\pi\eta d} \]
Here Boltzmann’s constant \( k_B \) determines Avogadro’s number \( N_A = R/k_B \) via the universal gas constant \( R \) (cf. ideal gas law). Comparing the so measured value of \( N_A \) with still other methods, Jean Baptiste Perrin in 1910 obtained a consistent unique number which was an important confirmation of the atomic hypothesis indeed.

A second revolutionary concept was the understanding of Brownian motion as a fluctuation-induced phenomenon, thus also correcting and extending thermodynamics. That is not at all restricted to motion in fluids; as an example, the work of S. Chandrasekhar in the 1938-1943 included an important model for dynamical friction in stellar dynamics based on Brownian motion. The study of Brownian motion is more largely at the beginning of stochastic modelling and the obvious role of fluctuations in physics phenomena, and beyond. Models of Brownian motion and diffusion processes have been used at least since the work of Louis Bachelier (PhD thesis *Théorie de la spéculaction*, 1900) to evaluate stock options and to use stochastic modeling in the study of finance, perhaps culminating in the most famous Black–Scholes–Merton model and Black–Scholes formula for the price of a European-style option.

### A. Random walk, continued

Let us consider a random walker on a line occupying sites 0, ±\( \delta \), ±2\( \delta \), …. After each time \( \tau \) the walker is displaced one step to the left or to the right with equal probability. We keep of course in mind to model the motion of a Brownian particle that is being pushed around at random. Then, the probability to find the particle at position \( n\delta \) at time \( k\tau \) equals
\[
P(n\delta, k\tau) = \frac{k!}{2^k (\frac{k-n}{2})! (\frac{k+n}{2})!}
\]
Taking \( n = x/\delta \), \( k = t/\tau \uparrow \infty \) while fixing \( D = \delta^2/2\tau \), we have the continuum limit
\[
\lim_{\delta \downarrow 0} \frac{1}{\delta} P(n\delta, k\tau) = \rho(x, t) = \frac{1}{\sqrt{4\pi Dt}} e^{-\frac{x^2}{4Dt}}
\]
for the (well-normalized) probability density \( \rho(x, t) \) which indeed satisfies the diffusion equation (VIII.1).
When considering initial conditions \( x_0 \neq 0 \) or initial time \( t_0 \neq 0 \) we have the fundamental solution of that diffusion equation reading
\[
\rho(x, t) = \frac{1}{\sqrt{4\pi D(t-t_0)}} e^{-\frac{(x-x_0)^2}{4D(t-t_0)}}, \quad \rho(x, t_0) = \delta(x-x_0)
\]

We can calculate all moments (take \( m \) even)
\[
\langle (x(t) - x_0)^m \rangle = \int_{\mathbb{R}} dx (x - x_0)^m \frac{1}{\sqrt{4\pi Dt}} e^{-\frac{(x-x_0)^2}{4Dt}} \sim t^{m/2}
\]
which means that while \( \langle (x - x_0)^2 \rangle \) is of order \( t \), the higher moments \( m > 2 \) tend to zero faster than \( t \) as \( t \downarrow 0 \).

There is of course another way to find back that solution. The point is that the random walk itself satisfies the Master equation
\[
P(n\delta, (k+1)\tau) = \frac{1}{2} P((n-1)\delta, k\tau) + \frac{1}{2} P((n+1)\delta, k\tau)
\]
or, again with \( x = n\delta, t = k\tau, D = \delta^2/(2\tau), \)
\[
\frac{P(x, t+\tau) - P(x, t)}{\tau} = \frac{\delta^2}{2\tau} \left\{ P((x+\delta, t) + P(x-\delta, t) - 2P(x, t) \right\}
\]
which tends to the diffusion equation in the limit \( \delta \downarrow 0 \) with \( D \) being kept fixed.

Finally we can look at the process in a more probabilistic way. After all, the position of the walker is the random variable (starting from \( X_0 = 0 \))
\[
X_{k\tau} = \delta(v_1 + v_2 + \ldots + v_k)
\]
where each \( v_i = \pm 1 \) with equal probability (independent random variables) and \( \delta = \sqrt{2\tau D} = \sqrt{2Dt/k} \). We let \( k \) go to infinity and consider the limit of the sum
\[
\sqrt{\frac{2Dt}{k}} (v_1 + v_2 + \ldots + v_k)
\]
The central limit theorem is saying that it converges in distribution to a Gaussian random variable with mean zero and variance \( 4Dt \). In fact, more is true. Not only do we get convergence to a Gaussian for fixed time \( t \); also the process itself has a limit and becomes a Gaussian process. That limiting process, rescaled limit of the standard random walk, is the so called Wiener process, also simply called after the basic phenomenon it models, Brownian motion. See more in Section XI A.
B. Einstein relation

We can consider the particle density as \( n(x, t) = N \rho(x, t) \) for a total of \( N \) particles. The number of particles is conserved and indeed the diffusion equation is a continuity equation

\[
\frac{\partial}{\partial t} n(x, t) + \frac{\partial}{\partial x} j_D(x, t) = 0, \quad j_D(x, t) := -D \frac{\partial}{\partial x} n(x, t)
\]

That diffusive current \( j_D \) is generated by density gradients. Suppose now however that there is also a constant external field, like gravity. We put the \( x \)-axis vertically upward. Using Newton’s equation we would write \( \dot{x} = v \) and

\[
m \frac{d}{dt} v(t) = -mg - m\gamma v(t)
\]

with \( m\gamma \) the friction (or drag) coefficient (and \( \gamma \) is called the damping coefficient). The particle current caused by the external field is \( j_g(x, t) = n(x, t)v(t) \) so that under stationary conditions, where \( dv/dt = 0 \), the contribution to the particle current from gravity equals \( j_g(x) = -gn(x)/\gamma \).

Yet, under equilibrium, there is no net particle current and the diffusive current \( j_D \) caused by concentration gradients cancels the “gravity” current. For these currents \( j_g \) and \( j_D \) we can use the equilibrium profile \( n(x) \). It is given by the Laplace barometric formula

\[
n(x) = n(x_0) e^{-\frac{V(x-x_0)}{k_BT}}
\]

where \( V \) is the external potential: \( V(x) = mgx \) for gravity. Whence, \( j_D(x) = D mg n(x)/(k_B T) \) and the total current \( j = j_g + j_D \) vanishes in equilibrium when

\[
j(x) = -\frac{gn(x)}{\gamma} + D mg n(x)/(k_B T) = 0
\]

which produces the diffusion constant as

\[
D = \frac{k_B T}{m\gamma}
\]

called, Einstein relation (Über die von molekülarkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierter Teilchen, 1905). The Stokes relation further specifies \( m\gamma = 6\pi \eta R \) in terms of the particle diameter \( R \) (at least for spherical particles) and the fluid viscosity \( \eta \).
Imagine again our random walker but now with a small bias, i.e., an external field $\alpha$, that also depends on the position. The walker moves to the right with probability $p(n\delta) = 1/2 + \alpha(n\delta)\delta$ and to the left with probability $q(n\delta) = 1/2 - \alpha(n\delta)\delta$. Then,

$$P(n\delta, (k+1)\tau) = p((n-1)\delta) P((n-1)\delta, k\tau) + q((n+1)\delta) P((n+1)\delta, k\tau)$$

which can be rewritten as

$$P(n\delta, (k+1)\tau) - P(n\delta, k\tau) = \frac{1}{2}\left\{\frac{1}{2} - \alpha((n+1)\delta) P((n+1)\delta, k\tau) - \alpha((n-1)\delta) P((n-1)\delta, k\tau)\right\}$$

which in the limit $\delta \downarrow 0, \tau \downarrow 0$ becomes

$$\frac{\partial}{\partial t}\rho(x, t) = -4D \frac{\partial}{\partial x}\left(\alpha(x)\rho(x, t)\right) + D \frac{\partial^2}{\partial x^2}\rho(x, t)$$

It is better to normalize it in the sense $4D\alpha(x) = F(x)/(m\gamma)$:

$$\frac{\partial}{\partial t}\rho(x, t) = -\frac{1}{m\gamma} \frac{\partial}{\partial x}\left(F(x)\rho(x, t)\right) + D \frac{\partial^2}{\partial x^2}\rho(x, t)$$

(IX.1)

which is called the Smoluchowski equation (1906).

Simple computations lead to

$$\frac{d}{dt}\langle x(t)\rangle = \frac{1}{m\gamma}\langle F(x(t))\rangle, \quad \frac{d}{dt}\langle x^2(t)\rangle = \frac{2}{m\gamma}\langle x(t)F(x(t))\rangle + 2D$$

These yield closed equations when $F$ is linear in which case the density $\rho(x, t)$ is Gaussian when starting from any fixed position. It is a good exercise to calculate $\rho(x, t)$ for $F(x) = -\kappa x$ (harmonic force).

Smoluchowski’s equation (IX.1) is a generalization of Brownian motion in the sense that the external field or potential is taking inside the equation for the evolution of the density. That equation can also be regarded as the Master equation for a Markov process:

$$\frac{\partial}{\partial t}\rho(x, t) = L^+\rho(x, t)$$

where $L^+$ is the adjoint (with respect to $dx$ in $\mathbb{R}$) of generator

$$Lf(x) = \frac{1}{m\gamma} f'(x) F(x) + D f''(x)$$

That can be guessed from its origin as a continuum description of a biased random walker. It is however useful to write more directly the stochastic updating, and that gives the Langevin equation, as in the next Section. We will then see it back in Section XII.
X. LANGEVIN DYNAMICS

If one considers a particle in a fluid as subject to friction $m\gamma$ and random forces $\xi(t)$, we can consider the Newtonian-type equation

$$m \frac{d}{dt} v(t) = -m\gamma v(t) + \xi(t) \quad (X.1)$$

This notation should not be taken extremely serious from the mathematical point of view. Its importance is foremost to suggest the physics of the dynamics. In a way, we want to describe a Markov process whose generator $L = L_1 + L_2$ is the sum of two contributions, drift with generator $L_1 f(v) = -\gamma v f'(v)$ and diffusion with generator $L_2 f(v) = C f''(v)$.

One question is how, more precisely, to think physically and mathematically about the process $\xi(t)$. Further technical issues will be how to calculate with it.

The general context is of course that of reduced dynamics. One has integrated out the many fluid particles and one has assumed that their initial conditions were statistically described. That can sometimes be done explicitly, but here we choose for more qualitative considerations; see Exercise 1 below.

The first thing that in general looks reasonable, certainly for a fluid at rest, is that the average $\langle \xi(t) \rangle = 0$. The average can first be understood as an arithmetic average over many samples, particle collisions etc., but in the mathematical model, the average will refer to a mathematical model for a stochastic process. From this hypothesis, we already deduce from (X.1) that

$$\langle v(t) \rangle = v_0 e^{-\gamma t}, \quad \langle x(t) \rangle_{v_0} = v_0 \frac{1 - e^{-\gamma t}}{\gamma}$$

with of course the notation $v(t = 0) = v_0$ and assuming $\langle x(0) \rangle = 0$.

Now to the covariance $\langle \xi(t_1)\xi(t_2) \rangle$ for which we can first assume that they reflect stationary correlations, only depending on the time-difference $|t_2 - t_1|$, as we assume the fluid is in a stationary condition and unaltered by the presence of the few bigger particles that we study here. In fact, a further simplification and physical hypothesis is that we model the force $\xi(t)$ as white noise. It means to suppose that

$$\langle \xi(t_1)\xi(t_2) \rangle = C \delta(t_1 - t_2)$$

for some constant $C$. In other words, the $\xi(t)$ in (X.1) varies much faster (on more microscopic time-scales) than the typical relaxation time $\gamma^{-1}$ of the velocity of the suspended
particle. That extra hypothesis helps because we can now compute the velocity correlations from

\[ v(t) = v_0 e^{-\gamma t} + \frac{1}{m} \int_0^t ds e^{-\gamma(t-s)} \xi(s) \]

Using the above statistics of the white noise we get

\[ \langle v(t_1)v(t_2) \rangle = \frac{C}{2\gamma m^2} \left( e^{-\gamma|t_1-t_2|} - e^{-\gamma(t_1+t_2)} \right) + v_0^2 e^{-\gamma(t_1+t_2)} \]  \hspace{1cm} (X.2)

In particular,

\[ \langle v^2(t) \rangle = \frac{C}{2\gamma m^2} (1 - e^{-2\gamma t}) + v_0^2 e^{-2\gamma t} \]

Note that we can determine the constant \( C \) from requiring that asymptotically in time

\[ \lim_{t \to \infty} \frac{1}{2} m \langle v(t)^2 \rangle = \frac{1}{2} k_B T \]

as from the equipartition rule (or, Maxwell distribution of speeds). Therefore we wish to choose

\[ C = 2m\gamma k_B T \]

We can think of it as a diffusion constant for the velocity process. To conclude, and slightly renormalizing the notation from (X.1), we have the Langevin equation

\[ \frac{dv(t)}{dt} = -\gamma v(t) + \sqrt{\frac{2\gamma k_B T}{m}} \xi(t), \quad \langle \xi(t)\xi(s) \rangle = \delta(t-s) \]

for the velocity thermalization of a particle in a heat bath (for all possible initial \( v(t=0) = v_0 \)).

There is of course also the position process, which is however completely enslaved by the velocity process. For example

\[ \langle x^2(t) \rangle = \int_0^t dt_1 \int_0^t dt_2 \langle v(t_1)v(t_2) \rangle \]

which can be computed from (X.2) to be

\[ \langle x^2(t) \rangle = \frac{2k_B T}{m\gamma} t + \frac{k_B T}{m\gamma^2} \left( 4e^{-\gamma t} - e^{-2\gamma t} - 3 \right) + v_0^2 \left( \frac{1}{\gamma} (e^{-\gamma t} - 1) \right)^2 \]

say for \( x_0 = 0 \). See that for small times, \( \langle x^2(t) \rangle \simeq (v_0 t)^2 \) (ballistic motion), while for large times \( t \) we find \( \langle x^2(t) \rangle \simeq 2k_B T t/(m\gamma) = 2Dt \) (diffusive motion with Einstein relation). Small and large times are physically distinguished in terms of the magnitude of the
dimensionless \( \gamma t \).

A final remark concerns the equilibrium dynamics, i.e., when \( v_0 \) is distributed with a Maxwellian distribution, \( \langle v_0^2 \rangle = k_B T/m \). That makes for one extra average in the above formula. For example,

\[
\langle x^2(t) \rangle = \frac{2k_B T}{m\gamma^2} (\gamma t + e^{-\gamma t} - 1), \quad \langle v(t_1)v(t_2) \rangle = \frac{k_B T}{m} e^{-\gamma |t_2-t_1|}
\]

XI. FOKKER-PLANCK EQUATION

In analogy with Brownian motion we call a time-independent Markov process diffusive if there are functions \( a(x) \) and \( b(x) \) so that the transition probability density \( p_t(x_0, x) \) satisfies

\[
\int_{\mathbb{R}} dx \ (x - x_0) p_t(x_0, x) = a(x_0)t + o(t)
\]

\[
\int_{\mathbb{R}} dx \ (x - x_0)^2 p_t(x_0, x) = b(x_0)t + o(t)
\]

\[
\int_{\mathbb{R}} dx \ (x - x_0)^m p_t(x_0, x) = o(t), \quad m > 2
\]

as \( t \downarrow 0 \). The notation \( o(t) \) means to indicate any term \( A(t) \) for which \( A(t)/t \to 0 \) as \( t \to 0 \).

We will show that then necessarily \( p_t(x_0, x) \) satisfies the Fokker-Planck equation (where Adriaan Fokker was a Dutch physicist, cousin of but not be confused with the aeronautical engineer Anthony Fokker)

\[
\frac{\partial}{\partial t} p_t(x_0, x) = -\frac{\partial}{\partial x} (a(x)p_t(x_0, x)) + \frac{1}{2} \frac{\partial^2}{\partial x^2} (b(x)p_t(x_0, x)) \tag{XI.1}
\]

In what follows, we will then simply write \( p_t(x_0, x) = \rho_t(x) \) for the probability density at time \( t \), ignoring the specific initial condition.

The proof of (XI.1) goes as follows. As the process is Markovian, we must have that at time \( t + \delta \)

\[
p_{t+\delta}(x_0, x) = \int_{\mathbb{R}} dy \ p_t(x_0, y) p_\delta(y, x)
\]

By integrating the previous identity with a smooth function \( \phi(x) \) and rearranging the integration (variables) we obtain

\[
\int_{\mathbb{R}} dy \ p_{t+\delta}(x_0, y) \phi(y) = \int_{\mathbb{R}} dy \ \int_{\mathbb{R}} dx \ p_t(x_0, y) p_\delta(y, x) \phi(x)
\]
Using a Taylor-expansion for \( \phi \),
\[
\int_{\mathbb{R}} dy p_{t+\delta}(x_0,y) \phi(y) = \int_{\mathbb{R}} dy p_{t}(x_0,y) \int_{\mathbb{R}} dx p_{\delta}(y,x) \left( \phi(y)+(x-y) \phi'(y)+\frac{1}{2}(x-y)^2 \phi''(y)+O((x-y)^3) \right)
\]
so that we get
\[
\int_{\mathbb{R}} dy p_{t+\delta}(x_0,y) \phi(y) = \int_{\mathbb{R}} dy p_{t}(x_0,y) \left( \phi(y) + a(y) \phi'(y) + \frac{1}{2} b(y) \phi''(y) + o(\delta) \right)
\]
Dividing by \( \delta \) and taking the limit \( \delta \downarrow 0 \) makes
\[
\int_{\mathbb{R}} dy \phi(y) \frac{\partial}{\partial t} p_{t}(x_0,y) = \int_{\mathbb{R}} dy p_{t}(x_0,y) \left( \phi'(y)a(y) + \frac{1}{2} \phi''(y) b(y) \right)
\]
The conclusion is reached by partial integration and by considering that these equalities hold for all functions \( \phi \).

The Fokker-Planck equation (XI.1) is of course also a Master equation with current \( j(x,t) \),
\[
\frac{\partial}{\partial t} \rho_{t}(x) + \frac{\partial}{\partial x} j(x,t) = 0, \quad j(x,t) = a(x) \rho_{t}(x) - \frac{1}{2} \frac{\partial}{\partial x} \left( b(x) \rho_{t}(x) \right)
\]
We can also write it in terms of the forward generator \( L^{+} \),
\[
\frac{\partial}{\partial t} \rho_{t} = L^{+} \rho_{t}
\]
The backward generator (adjoint of \( L^{+} \) with respect to \( dx \)) is then
\[
L f(x) = a(x) f'(x) + \frac{b(x)}{2} f''(x)
\]
working on functions (observables) \( f \). There will be conditions on \( a(x), b(x) \) that assure that \( p_{t}(x_0,x) \rightarrow \rho(x) \) (stationary density) as \( t \uparrow +\infty \).

We have detailed balance and an equilibrium distribution \( \rho \) when the current in the stationary regime vanishes; that is when
\[
a(x) \rho_{t}(x) = \frac{1}{2} \frac{\partial}{\partial x} \left( b(x) \rho_{t}(x) \right)
\]
Apparently, on \( \mathbb{R} \) that is always possible, as long as \( \rho \) is normalizable. For example, we can take \( F(x) = (2a(x) - b'(x))/b(x) \) and find the potential \( V \) such that \( -V'(x) = F(x) \). Then, \( \rho(x) = \frac{1}{Z} \exp -V(x) \) if normalizable: \( Z = \int_{\mathbb{R}} \exp -V(x) < \infty \). That need however not be true in higher dimensions; detailed balance can easily be broken whenever non-conservative forces are present. More on that later.
A. Wiener process

The Wiener process is the Brownian motion process, called after and in honor of Norbert Wiener. It corresponds to the special case where \( a = 0, b = 2D = 2k_B T/(m\gamma) \). It is a non-stationary time-independent Gaussian process with mean zero and covariance

\[
\langle x(t_1) x(t_2) \rangle = 2D \min\{t_1, t_2\}
\]

The density at each time \( t \) is Gaussian with

\[
\rho_t(x) = \frac{1}{\sqrt{4\pi D t}} e^{-\frac{x^2}{4Dt}}, \quad t > 0
\]

and the transition probability density is

\[
p_t(x, y) = \frac{1}{\sqrt{4\pi D t}} e^{-\frac{(x-y)^2}{4Dt}}, \quad t > 0
\]

Note finally the relation with white noise. If we take

\[
x(t) = \int_0^t ds \xi(s)
\]

with white noise \( \xi(s) \) having covariance \( \langle \xi(s)\xi(t) \rangle = \Gamma \delta(t-s) \), then

\[
\langle x(t_1) x(t_2) \rangle = \Gamma \int_0^{t_1} ds_1 \int_0^{t_2} ds_2 \delta(s_1-s_2) = \Gamma \min\{t_1, t_2\}
\]

B. Ornstein-Uhlenbeck process

The Ornstein-Uhlenbeck process (named after Dutch theoretical physicists Leonard Ornstein and George Uhlenbeck) is the unique Markov process which is both stationary and Gaussian. Physically it is then better to have in mind the thermalization of velocities. We thus change (with respect to the previous sections) \( x \to v \) to emphasize that the process deals with speed instead of position. We choose \( a(v) = -\gamma v \) and \( b(v) = 2k_B T\gamma/m = 2\gamma^2 D \).

The Fokker-Planck equation (XI.1) is then

\[
\frac{\partial}{\partial t} \rho_t(v) = \gamma \frac{\partial}{\partial v} (v \rho_t(v)) + \gamma^2 D \frac{\partial^2}{\partial v^2} \rho_t(v)
\]

and the Markov process that corresponds to it has transition probability density

\[
p_t(v, w) = \sqrt{\frac{m}{2\pi k_B T}} \frac{1}{\sqrt{1-e^{-2\gamma t}}} \exp \left( -\frac{m}{2k_B T} \frac{(w - ve^{-\gamma t})^2}{1-e^{-2\gamma t}} \right)
\]
with stationary distribution \( \rho(v) = \sqrt{\frac{m}{2\pi k_B T}} e^{\frac{mv^2}{2k_BT}} \). Therefore in the stationary regime \( \langle v(t) \rangle = 0 \) and another calculation shows

\[
\langle v(t_0) v(t) \rangle = \frac{k_B T}{m} e^{-\gamma(t-t_0)}
\]

We can also calculate \( \langle v(t) \rangle \) for fixed initial speed \( v_0 \) at an earlier time, to find that these coincide exactly with the Langevin dynamics. In other words, the Langevin dynamics (X.1) is an Ornstein-Uhlenbeck process. After all, in the Langevin dynamics, the speed \( v(t) \) must be a Gaussian process, and hence is completely fixed by its mean and covariances in time.

## XII. OVERDAMPED DIFFUSIONS

We take a real variable \( x \) undergoing a drift \( F \) and white noise \( \xi \)

\[
\frac{d}{dt} x(t) = F(x(t)) + \xi(t), \quad \langle \xi(t)\xi(s) \rangle = \Gamma \delta(t-s)
\]

That is clearly a time-homogeneous Markov diffusion process because the increments are independent and given by the white noise. Its probability density will therefore satisfy a Fokker-Planck equation (XI.1) with \( a(x) = F(x) \) and \( b(x) = \Gamma \).

There are various natural ways to obtain such a diffusion process, and one important approximation is to start from the dynamics

\[
\frac{d}{dt} x(t) = v(t), \quad \frac{d}{dt} v(t) = \frac{1}{m} F(x(t)) - \gamma v(t) + \xi(t)
\]

and to suppose that we can neglect the acceleration with respect to the other terms. Such overdamped approximation then gives an evolution equation for the position:

\[
\frac{d}{dt} x(t) = \frac{1}{m\gamma} F(x(t)) + \xi(t), \quad \langle \xi(s)\xi(t) \rangle = \frac{\Gamma}{\gamma^2} \delta(t-s)
\]

which has \( a(x) = F(x)/(m\gamma) \) and \( b(x) = \Gamma/\gamma^2 = D \), or with Fokker-Planck equation

\[
\frac{\partial}{\partial t} \rho_t(x) = -\frac{1}{m\gamma} \frac{\partial}{\partial x} (F(x) \rho_t(x)) + \frac{D}{2} \frac{\partial^2}{\partial x^2} \rho_t(x)
\]

which is the Smoluchowski equation (IX.1). When \( F(x) = -V'(x) \), then the equilibrium distribution is

\[
\rho(x) = \frac{1}{Z} \exp[-\beta V(x)], \quad \beta = \frac{2\gamma}{m\Gamma}
\]
XIII. MORE DIMENSIONAL DIFFUSION PROCESSES

The main logic and derivations do not change when considering multiple variables or more dimensional variables. Yet, the physics becomes more rich. We now put \( x(t) = (x_1(t), \ldots, x_n(t)) \) and suppose that, as \( t \downarrow 0 \), the transition probability density satisfies, componentwise,

\[
\int_{\mathbb{R}^n} dx (x - x_0)_i p_t(x_0, x) = a_i(x_0) t + o(t)
\]

\[
\int_{\mathbb{R}^n} dx (x - x_0)_i (x - x_0)_j p_t(x_0, x) = b_{ij}(x_0) t + o(t)
\]

for \( a \in \mathbb{R}^n \) and \( b \) a real symmetric \( n \times n \) matrix, and all higher moments vanishing more rapidly than \( t \) as \( t \downarrow 0 \).

The corresponding Fokker-Planck equation is then

\[
\frac{\partial}{\partial t} \rho_t(x) = -\sum_{i=1}^n \frac{\partial}{\partial x_i} (a_i(x) \rho_t(x)) + \frac{1}{2} \sum_{i,j=1}^n \frac{\partial^2}{\partial x_i \partial x_j} (b_{ij}(x) \rho_t(x))
\]

A basic example is Kramers’ equation where, in the simplest version, \( n = 2 \) and with variables position and velocity:

\[
\frac{d}{dt} x(t) = v(t), \quad \frac{d}{dt} v(t) = -\gamma v(t) + \frac{1}{m} F(x(t)) + \sqrt{ \frac{2k_B T}{m}} \xi(t)
\]

Correspondingly, \( a(x,v) = (v, -\gamma v + F(x)/m) \) and \( b_{xx} = b_{xv} = b_{vx} = 0, b_{vv} = 2k_B T/\gamma m \).

That special case of the multi-dimensional Fokker-Planck equation gives

\[
\frac{\partial}{\partial t} \rho_t(x,v) + v \frac{\partial}{\partial x} \rho_t(x,v) + \frac{F(x)}{m} \frac{\partial}{\partial v} \rho_t(x,v) = \gamma \left( \frac{\partial}{\partial v} (v \rho_t(x,v)) + \frac{k_B T}{m} \frac{\partial^2}{\partial v^2} \rho_t(x,v) \right)
\]

which is called Kramers’ equation (after yet another Dutch theoretical physicist Hans Kramers). It can be regarded as a special case of a kinetic equation (of which the Boltzmann equation is the most famous representative), but with the great simplification that its right-hand side is linear in the density. That simplification is the result of considering only single (or independent) particles interacting with the fluid at rest.

XIV. DIFFUSION ON THE CIRCLE

We consider a particle undergoing an overdamped motion subject to some drift and to white noise:

\[
\dot{X}_t = [E - U'(X_t)] + \sqrt{2/\beta} \xi_t
\]
The dynamical variable $X_t \in S^1$ with $0 \leq t =$ time, takes values on the circle of length 1. The constant $E$ represents the (nonequilibrium) driving. Note that on the circle, a constant cannot be seen as the gradient of a potential function. The potential $U$ is periodic and assumed sufficiently smooth to assure the existence of a unique and stationary and smooth probability density $\rho$ with respect to $dx$. The $\beta$ refers to the inverse temperature of the medium in which the particle moves.

The Master equation is the usual Fokker-Planck equation (XI.2) for the time-dependent probability density $\rho_t$:

$$\frac{d\rho_t(x)}{dt} + j'_t(x) = 0, \quad j_t = (E - U')\rho_t - \frac{1}{\beta} \rho'_t$$  \hspace{1cm} (XIV.2)

as started from some initial density $\rho_0$. Stationarity implies that $j'_t = 0$, or the stationary current $j$ is a constant (in $x$). The current $j$ can also be identified with the actual flow of the particle around the circle.

The (backward) generator $L$ of the process on smooth functions $f$ is

$$\frac{d}{dt}(f(X_t))_{X_0 = x} = Lf(x) = \frac{1}{\beta} f''(x) + (E - U'(x)) f'(x)$$

Here are some standard possibilities:

- **Detailed balance.** When $E = 0$, then (XIV.1) is an equilibrium dynamics with reversible density

  $$\rho(x) = \frac{1}{Z} \exp -\beta U(x), \quad Z = \int dx \exp -\beta U(x)$$

  That the dynamics is time-reversible in the steady state is apparent by the fact that here $j = 0$.

- **Uniform field.** The case $U = 0$ is the simplest nonequilibrium case. The stationary distribution is given by the uniform density $\rho(x) = 1$, but there is no detailed balance. The stationary current equals $j = E$.

- **Non-uniform driving.** When $E \neq 0$ and $U \neq 0$ we can combine them as $F = E - U'$.

  The stationary density obtains the form

  $$\rho(x) = \frac{1}{Z} \iint e^{\beta W(y,x)} dy, \quad Z = \iint e^{\beta W(y,x)} dy dx$$  \hspace{1cm} (XIV.3)
where
\[ W(y, x) = \begin{cases} \int_y^x F \, dz & \text{for } y \leq x \\ \int_y^1 F \, dz + \int_0^x F \, dz & \text{for } y > x \end{cases} \]
is the work performed by the applied forces along the positively oriented path \( y \to x \).

In this model the stationary current can be computed by dividing the stationary Master equation by \( \rho \) and by integration over the circle:
\[
j = \frac{W}{\bar{f}(\rho)^{-1} dx}
\]
where \( \bar{W} = \int_0^1 F \, dx \) is the work carried over a completed cycle. The non-zero value of this stationary current indicates that time-reversibility is broken.

**XV. PATH-INTEGRAL FORMULATION**

As for jump processes, we can ask for the plausibility of paths or trajectories \( \omega = (x_s, 0 \leq s \leq t) \) in a time-interval \([0, t]\). The trajectories of diffusion processes are continuous but not differentiable, so that it becomes a more technical issue how to select the space of trajectories and how to write down the action. Here we will rather formal, and start with the most singular object of all, which is the white noise \( \xi \) appearing in the Langevin and diffusion processes discussed before.

The useful input here is to remember the distribution of a multivariate normal variable; that is a higher dimensional Gaussian random variable. Say for \( k \) components, \( \eta = (\eta_1, \eta_2, \ldots, \eta_k) \) is multivariate Gaussian (with mean zero) when its density on \((\mathbb{R}^k, d\eta_1 \ldots d\eta_k)\) equals
\[
f(\eta_1, \eta_2, \ldots, \eta_k) = \frac{1}{\sqrt{(2\pi)^k \det A}} \exp \left( -\frac{1}{2} (\eta, A^{-1} \eta) \right)
\]
where \( \det A \) is the determinant of the positive definite matrix \( A \), and the scalar product in the exponential is \( (\eta, A^{-1} \eta) := \sum_{i,j} (A^{-1})_{ij} \eta_i \eta_j \). It is not hard to show that their covariance then equals
\[
\langle \eta_i \eta_j \rangle = A_{ij}
\]
determined by the matrix elements of \( A \).

Imagine now that the label \( i \) for the component \( \eta_i \) in reality refers to a time-instant, so that we are dealing with a Gaussian stochastic process with mean zero and for which the
time-correlations are specified using a kernel \( A(s,t) \). Indeed, a Gaussian process is but a stochastic process such that every finite collection of fixed-time realizations has a multivariate normal distribution. Thus, a Gaussian process is the infinite-dimensional realization of a multivariate Gaussian random variable. White noise is a degenerate version of this where the covariance is just the unit operator. Formally then, the path probability density of white noise \( \xi \) must be proportional to

\[
\exp \left( -\frac{1}{2} \int_0^t ds \, \xi_s^2 \right) \tag{XV.2}
\]

where the sums in the scalar product in the exponential of (XV.1) have now been replaced by an integral over time, and corresponding \( A(s,t) = \delta(t-s) \).

The next step is to take back stochastic differential equations such as

\[
\dot{x}_t = F(x_t) + \sqrt{2T} \xi_t \tag{XV.3}
\]

and to insert the density for the white noise (XV.2), obtaining path density

\[
P(\omega) \propto \exp \left( -\frac{1}{4T} \int_0^t ds \, (\dot{x}_s - F(x_s))^2 \right)
= \exp \left( -\frac{1}{4T} \int_0^t ds \, \dot{x}_s^2 \right) \exp \left( \frac{1}{2T} \int_0^t dx_s \, F(x_s) - \frac{1}{4T} \int_0^t ds \, F^2(x_s) \right) \tag{XV.4}
\]

All kinds of things must be further specified here, mostly having to do with the stochastic calculus of Section XVI, but the main usage of such formulæ is for obtaining the density of one process in terms of another process. In a way, the first exponential factor in (XV.4) refers again to white noise and is associated to the Brownian process \( B_t \) for which \( \dot{B}_t = \sqrt{2T} \xi_t \).

We can thus compare the path density of the process \( x_t \) with that of \( B_t \) and write

\[
P(\omega) = e^{-A(\omega)} P_B(\omega)
\]

with respect to that Brownian process \( P_B \) and with “action”

\[
A(\omega) := -\frac{1}{2T} \int F(x_s) \, dx_s + \frac{1}{4T} \int_0^t ds \, F^2(x_s) \tag{XV.5}
\]

More generally, if we have two diffusion processes at the same temperature \( T \) such as (XV.3) with respectively drift \( F = F_1 \) and \( F = F_2 \), then the expectation of a general observable \( O(\omega) \) in the first process can be related to another expectation in the second process via

\[
\langle O(\omega) \rangle_1 = \langle O(\omega) e^{A_2(\omega) - A_1(\omega)} \rangle_2
\]
which can also be summarized as the so called Girsanov formula

\[
\frac{dP_1}{dP_2}(\omega) = \exp \frac{1}{2T} \int_0^T [F_1(x_s) - F_2(x_s)] \, dx_s - \frac{1}{4T} \int_0^T ds \, [F_1 - F_2]^2(x_s)
\]  

(XV.6)

which can be formally obtained from (XV.4) by taking the ratio. We skip here the further extensions to include \(x\)-dependent temperature, and higher dimensional versions which however basically follow the same logic as above, while their physics may become more rich.

**XVI. ELEMENTS OF STOCHASTIC CALCULUS**

The formulæ (XV.5) or (XV.6) contain stochastic integrals such as \(\int_0^T F(x_s) \, dx_s\). Their meaning and manipulation is the subject of stochastic calculus. Stochastic differential and integration calculus differs from the ordinary Newtonian calculus because of the diffusive nature of the motion in which the path \(x_t\) does not allow a velocity; diffusive motion implies \((dx_t)^2 \propto dt\).

Let us first look back at, what for physics, is a natural point of departure, the stochastic differential equation (XV.3). We think of determining the next position \(x_{t+dt}\) from knowledge of \(x_t, F(x_t)\) and the random force \(\xi_t\). Such an interpretation of (XV.3) is called reading the equation in the Itô sense. The resulting stochastic integral is then thought of as a limit where we discretize the time-interval \([0,t]\) via many intermediate times \(0 = t_0 < t_1 < \ldots t_n = t\)

\[
\int_0^T F(x_s) \, dx_s = \lim_{n \to \infty} \sum_i F(x_{t_i})(x_{t_{i+1}} - x_{t_i})
\]  

(XVI.1)

and where all the time-intervals \(t_{i+1} - t_i \simeq 1/n\) tend to zero. The limit has to be understood in the sense of square integrable functions.

The Itô-integral is thus characterized by the fact that the function \(F\) in (XVI.1) is each time evaluated in the left point of the time-interval. There is another more time-symmetric way of doing the integral, referred to as the Stratonovich sense and denoted by

\[
\int_0^T F(x_s) \circ dx_s = \lim_{n \to \infty} \sum_i F\left(\frac{x_{t_i} + x_{t_{i+1}}}{2}\right) (x_{t_{i+1}} - x_{t_i})
\]

which evaluates the function in the middle of the time-integral. The Itô and Stratonovich integrals can be related via the expansion

\[
\sum_i F\left(\frac{x_{t_i} + x_{t_{i+1}}}{2}\right) (x_{t_{i+1}} - x_{t_i}) = \sum_i \{F(x_{t_i}) + F'(x_{t_i}) \frac{x_{t_{i+1}} - x_{t_i}}{2} + \ldots \} (x_{t_{i+1}} - x_{t_i})
\]
which leads to
\[
\int_0^t F(x_s) \circ dx_s = \int_0^t F(x_s) \, dx_s + T \int_0^t F'(x_s) \, ds \tag{XVI.2}
\]
where we have used that \((dx_s)^2 = 2Tds\).

The Stratonovich calculus has by its time-symmetric formulation a number of mathematical niceties, also making it closer to ordinary integration. For example, taking \(F = f'\) into (XVI.2) yields
\[
f(x_t) - f(x_0) = \int_0^t f'(x_s) \, dx_s + T \int_0^t f''(x_s) \, ds
\]
In other words,
\[
df(x_t) = f'(x_t) \, dx_t + T f''(x_t) \, dt
\]
which is called the Itô-lemma. In fact, if we insert here \(dx_t = F(x_t)\, dt + \sqrt{2T} \, dB_t\) (which is (XV.4) with \(B_t\) Brownian motion), we get
\[
df(x_t) = \{ f'(x_t) F(x_t) + T f''(x_t) \} \, dt + \sqrt{2T} \, dB_t
\]
\[
= Lf(x_t) \, dt + \sqrt{2T} \, dB_t
\]
where we recognize the backward generator \(L\) acting on \(f\) in the term with \(dt\).

Finally, we still note that also the action can be rewritten as
\[
\mathcal{A}(\omega) := -\frac{1}{2T} \int F(x_s) \circ dx_s + \frac{1}{4T} \int_0^t ds \{ F^2(x_s) - 2TF'(x_s) \} \tag{XVI.3}
\]
which is now nicely split up in a time-antisymmetric term and a time-symmetric term for the time-reversal transformation. Physicists recognize the antisymmetric term
\[
\sigma := \frac{1}{T} \int F(x_s) \circ dx_s
\]
as the time-integrated entropy flux, or the dissipated heat by temperature from the forcing \(F\) in (XV.4). If the force \(F\) is gradient, \(F = -U'\) for system energy \(U\), then \(\sigma = [U(x_0) - U(x_t)]/T\) equals the energy change of the environment over its temperature. If however \(F\) contains non-gradient parts, like when considering (XV.4) on the circle in Section XIV.3, \(\sigma\) is time-extensive Joule heating over bath temperature.
XVII. STOCHASTIC FIELDS

So far we have mostly been looking at single (or independent) particles diffusing in some environment. We can however also build interacting Brownian motions or even interacting stochastic fields. In this way we consider models that are spatially extensive, and the basic variables are fields (to be completed...)

XVIII. EXERCISES

1. Try to follow and to add the details to the following derivation of a Langevin-type dynamics starting from a Newton dynamics. We want the effective or reduced dynamics for a bigger particle in a sea (heat bath) of smaller particles.

The bigger particle (e.g. a colloid) is described by a coordinate \( q \) and its conjugate momentum \( p \). The heat bath consists of harmonic oscillators described by a set of coordinates \( \{ q_j \} \) and their conjugate momenta \( \{ p_j \} \). For simplicity, all oscillator masses are set equal to 1. The system Hamiltonian is

\[
H_s = \frac{p^2}{2m} + U(q)
\]

and the heat bath Hamiltonian \( H_B \) includes harmonic oscillator Hamiltonians for each oscillator and a very special coupling to the system,

\[
H_B = \sum_j \left[ \frac{p_j^2}{2} + \frac{1}{2} \omega_j^2 (q_j - \frac{\gamma_j}{\omega_j} q)^2 \right]
\]

in which \( \omega_j \) is the frequency of the \( j \)th oscillator and \( \gamma_j \) measures the strength of coupling of the system to the \( j \)th oscillator. \( H_B \) consists of three parts: the first is just the ordinary harmonic oscillator Hamiltonian, specified by its frequencies; the second contains a bilinear coupling to the system, \( \sum \gamma_j q_j q \), specified by the coupling constants; and the third contains only \( q \) and could be regarded as part of the arbitrary \( U(q) \). The bilinear coupling is what here makes the derivation manageable. The equations of motion for the combined Hamiltonian \( H_s + H_B \) are

\[
\frac{dq}{dt} = \frac{p}{m}, \quad \frac{dp}{dt} = -U'(x) + \sum_j \gamma_j (q_j - \frac{\gamma_j}{\omega_j^2} q) \]

\[
\frac{dq_j}{dt} = p_j, \quad \frac{dp_j}{dt} = -\omega_j^2 q_j + \gamma_j q
\]
Suppose that the time dependence of the system coordinate \( q(t) \) is known. Then it is easy to solve for the motion of the heat bath oscillators, in terms of their initial values and the influence of \( q(t) \),

\[
q_j(t) = q_j(0) \cos \omega_j t + p_j(0) \sin \omega_j t + \gamma_j \int_0^t ds q(s) \frac{\sin \omega_j (t - s)}{\omega_j}.
\]

Integration by parts leads to a more useful form:

\[
q_j(t) = q_j(0) \cos \omega_j t + p_j(0) \sin \omega_j t + \gamma_j \int_0^t ds p(s) \frac{\cos \omega_j (t - s)}{m}.
\]

When this is put back into the equation for \( dp/dt \), we obtain

\[
\frac{dp}{dt} = -U'(x) - \int_0^t ds K(s) \frac{p(t - s)}{m} + \xi(t)
\]

in which the “memory function” \( K(t) \) in the friction is explicitly

\[
K(t) = \sum_j \frac{\gamma_j^2}{\omega_j^2} \cos \omega_j t
\]

and the “noise” \( \xi(t) \) is given explicitly by

\[
\xi(t) = \sum_j \gamma_j p_j(0) \frac{\sin \omega_j t}{\omega_j} + \sum_j \gamma_j (q_j(0) - \gamma_j \omega_j q(0)) \cos \omega_j t
\]

By carefully choosing the spectrum \( \{\omega_j\} \) and coupling constants \( \{\gamma_j\} \), the memory function can be given any assigned form. For example, if the spectrum is continuous, and the sum over \( j \) is replaced by an integral, \( \int d\omega g(\omega) \), where \( g(\omega) \) is a density of states, and if \( \gamma \) is a function of \( \omega \), then the memory function \( K(t) \) becomes a Fourier integral,

\[
K(t) = \int d\omega g(\omega) \frac{\gamma(\omega)^2}{\omega^2} \cos \omega t
\]

Further, if \( g(\omega) \) is proportional to \( \omega^2 \) and \( \gamma \) is a constant, then \( K(t) \) is proportional to \( \delta(t) \) and the integral disappears from (XVIII.1).

The noise \( \xi(t) \) is defined in terms of the initial positions and momenta of the bath oscillators and is therefore in principle a known function of time. However, if the bath has a large number of independent degrees of freedom, then the noise is a sum containing a large number of independent terms, and because of the central limit theorem, we can expect that its statistical properties are simple. Suppose, for example, that a
large number of computer simulations of this system are done. In each simulation, the bath initial conditions are taken from a distribution,

\[ \rho_{eq}(p_j, q_j) \sim \exp \left( -\frac{H_B}{k_B T} \right) \]

in which the bath is in thermal equilibrium with respect to a frozen or constrained system coordinate \( q(0) \). Then the averages of \( q \) and \( p \) are

\[ \langle q_j(0) - \frac{\gamma_j}{\omega_j^2} q(0) \rangle = 0, \quad \langle p_j(0) \rangle = 0 \]

Since the noise is a linear combination of these quantities, its average value is zero.

The second moments are

\[ \langle (q_j(0) - \frac{\gamma_j}{\omega_j^2} q(0))^2 \rangle = \frac{k_B T}{\omega_j^2}, \quad \langle p_j^2(0) \rangle = k_B T \]

There are no correlations between the initial values for different \( j \)'s. Then by direct calculation, using trigonometric identities, one sees immediately that there is (what is called) a fluctuation-dissipation relation,

\[ \langle \xi(t) \xi(t') \rangle = k_B T K(t - t') \]

Because the noise is a linear combination of quantities that have a Gaussian distribution, the noise is itself a Gaussian random variable. If the heat bath has been constructed so that the memory function is a delta function, then the noise is *white*.

2. Fix parameters \( \alpha > 0, B \in \mathbb{R}, D > 0 \) and look at the linear Langevin dynamics for a global order parameter \( M \in \mathbb{R} \),

\[ \dot{M}(t) = -\alpha M(t) + h_t B + \sqrt{2D} \xi(t) \]

for standard white noise \( \xi(t) \). The \( h_t, t > 0 \), is a small time-dependent field. We can think of a Gaussian approximation to a relaxational dynamics of the scalar magnetization \( M \) (no conservation laws and no spatial structure).

Show that the equilibrium (reversible stationary density on \( \mathbb{R} \) for perturbation \( B = 0 \)) is

\[ \rho(M) = \frac{1}{Z} \exp \left\{ -\frac{\alpha M^2}{2D} \right\} \]

with zero mean and variance \( \langle M^2 \rangle = D/\alpha \).
3. We now start from an initial fixed \( M(t = 0) = M_0 \).

Show that the response function (to linear order in the term \( h_t B \)) is obtained from

\[
\langle M(t) \rangle = M_0 e^{-\alpha t} + B \int_0^t ds \, h_s \, e^{-\alpha (t-s)}
\]

(XVIII.2)

Show also that for \( h_t \equiv 0 \), the correlation function for \( 0 < s < t \) is

\[
\langle M(s) M(t) \rangle = M_0^2 e^{-\alpha (t+s)} + \frac{D}{\alpha} \left[ e^{-\alpha (t-s)} - e^{-\alpha (t+s)} \right]
\]

and hence

\[
\frac{d}{ds} \langle M(s) M(t) \rangle = -\alpha M_0^2 e^{-\alpha (t+s)} + D \left[ e^{-\alpha (t-s)} + e^{-\alpha (t+s)} \right]
\]

4. Show next that when we average that last expression over the equilibrium density we find

\[
\frac{d}{ds} \langle M(s) M(t) \rangle_{\rho} = D e^{-\alpha (t-s)}
\]

and compare with the response (XVIII.2).

5. Solve the Langevin equation also in higher dimensions. Take for example the evolution equation for a charged particle in a magnetic field

\[
m \dot{v}(t) = -m \gamma v(t) + e (v \times B) + \xi(t)
\]

where \( B \) is the magnetic field, \( e \) is the charge, \( \gamma \) is the friction, \( m \) is the mass and \( \xi(t) \) are three independent standard white noises. There is a stationary equilibrium. Compute the velocity correlation matrix for \( t \geq 0 \) in equilibrium

\[
\langle v_i(0) v_j(t) \rangle_{eq}
\]

where \( i, j \) are the three spatial directions.

6. Take again the standard Langevin dynamics for a particle in a large box with friction coefficient \( \gamma \) at inverse temperature \( \beta \). The stationary distribution for the velocities in Maxwellian, while the position diffuses uniformly over the box. Imagining a large time-interval \( t \uparrow +\infty \) we define the diffusion constant as

\[
D = \lim_{t \to +\infty} \frac{1}{t} \langle (x_t - x_0)^2 \rangle
\]
where we average over the equilibrium trajectories. Let us now add a perturbation, taking a constant but small external field $E$, switched on at time zero:

$$
\dot{v}_t = -\gamma v_t + E + \sqrt{\frac{2\gamma}{\beta}} \xi_t, \quad t > 0
$$

(XVIII.3)

and we enquire on the mobility

$$
\chi = \frac{d}{dE} \langle v \rangle |(E = 0)
$$

The average is over the new stationary distribution (with $E$) but we look at the linear order. Verify the relation

$$
\chi = 2d \beta D
$$

(with $d$ the dimension). In other words the mobility satisfies

$$
\chi = 2d \beta \lim_{t \to \infty} \frac{1}{\sqrt{t}} \int_0^t v_s ds^2
$$

in equilibrium.

7. In order to understand the frequency dependence consider changing the Langevin dynamics into

$$
\frac{d}{dt} v_t = -\gamma v_t + \frac{1}{m} F(t) + \sqrt{\frac{2\gamma}{\beta m}} \xi(t), \quad t > 0
$$

(XVIII.4)

with $\xi(t)$ standard white noise and $F(t)$ some time-dependent external force. We have

$$
\langle v(t) \rangle = \int_{-\infty}^{+\infty} ds R(t - s) F(s)
$$

Show that

$$
R(\omega) = \frac{1}{m \gamma - i\omega}
$$

Show that the imaginary part

$$
\Im R(\omega) = \frac{\beta \omega}{2\gamma} G(\omega)
$$

where $G$ is the Fourier transform of the velocity autocorrelation function

$$
G(\omega) = \int_{-\infty}^{+\infty} dt e^{i\omega t} \langle v(t) v(0) \rangle_{eq}
$$

8. Verify the claim (XIV.3), and the expression for the current (XIV.4). See also how it reduces to the equilibrium form in case of detailed balance ($E = 0$).
Large Deviations

XIX. INTRODUCTION

“Many important phenomena, in physics and beyond, while they cannot be shown to hold without exception, can be shown to hold with very rare exception, suitably understood.” (S. Goldstein in Typicality and Notions of Probability in Physics, in: Probability in Physics, edited by Y. Ben-Menahem and M. Hemmo, 59-71 (Springer, 2012)).

The theory of large deviations is a theory about fluctuations around some typical (limiting) behavior. It depends on the physical context what that typical (limiting) behavior refers to but invariably, it mathematically corresponds to some (generalized) law of large numbers. Here are a number of possible examples:

1. **Macroscopic averages**: We start from identifying the degrees of freedom in the system, like what are the basic dynamical variables in the given description. We are then interested in macroscopic properties. These are given in terms of arithmetic means of the individual contributions. These are typically constant as we vary over most of the \textit{a priori} possibilities. That is sometimes called macroscopic reproducibility. For example, for a dilute gas we can ask for the fraction of particles having their position and velocity around a given position and velocity (with some tolerance). The evolution of that coarse-grained description is provided by the Boltzmann equation. When we randomly throw the particles in a volume for a given fixed total energy, we expect to find the homogeneous Maxwell distribution for a temperature that corresponds to that energy. That is typical. Fluctuations refer to deviations from this law. Here is how Boltzmann first wrote in 1896 about that case: \textit{One should not forget that the Maxwell distribution is not a state in which each molecule has a definite position and velocity, and which is thereby attained when the position and velocity of each molecule approach these definite values asymptotically.... It is in no way a special singular distribution which is to be contrasted to infinitely many more non-Maxwellian distributions; rather it is characterized by the fact that by far the largest number of possible velocity distributions have the characteristic properties of the Maxwell distribution, and compared to these there are only a relatively small number of possible distributions that deviate significantly from Maxwells.} (Ludwig Boltzmann, 1896)
2. **Time-averages**: We can also ask for time-averages, such as what is the fraction of time that the system has spent in some part of its phase space. When the dynamics is ergodic, that is given by a phase-space average in the limit for large times. Deviations (for intermediate times) are possible and are referred to as dynamical fluctuations.

3. **Deterministic limits**: In some cases we start from a noisy or stochastic dynamics and we are interested to take the zero noise limit as reference. Typical behavior would then correspond to the zero noise dynamics and deviations from that trajectory are possible but increasingly unlikely as the noise gets smaller. Noise can be thermal in which case we are interested in ground state properties and low temperature behavior.

4. **Classical limits**: The previous case can be restated in the case of quantum noise. When discussing quantum dynamics we have in mind that there are limiting cases of observations or situations where the classical (Newtonian) dynamics should emerge. We have in mind the case where the de Broglie wavelength gets very small with respect to spatial variations. Informally speaking, for say large objects moving at high energies, the typical behavior is classical. Around that situation we get quantum fluctuations and the possibly multiple and looped Feynman paths start to contribute.

The above already indicate various good reasons for the study of large deviations. They describe the probability of risks, of rare events, but they also connect various levels of descriptions. It is perhaps also not surprising that through them, we will obtain variational principles characterizing the limiting typical behavior. Even when we never see the rare events, their plausibility is governed by fluctuation functionals that do have a physical and operational meaning. Their derivatives will for example matter in response theory around that typical behavior. On a more analytic side the theory of large deviations connects with asymptotic evaluation of integrals as in the Laplace formula and approximation.

We start explaining all that next with the simplest case where we consider independent random variables.

### XX. INDEPENDENT VARIABLES

Suppose that \(X_1, X_2, \ldots, X_i, \ldots\) are independent and identically distributed real-valued random variables. Their arithmetic mean is \(m_N := (X_1 + \ldots X_N)/N\). Unless otherwise
stated we will indicate the \textit{a priori} distribution of \( X_i \) by \( \rho \) and expectation values (averages) with \( \langle \cdot \rangle \). Limits typically refer to \( N \uparrow +\infty \).

\section*{A. Law of large numbers}

The strong law of large numbers asserts that \( m_N \to \langle X_i \rangle \) almost surely if \( \langle |X_i| \rangle < \infty \). The “almost surely” makes it \textit{strong}: the convergence (of the random mean) to the constant (expectation) is with probability one. It is however much easier to prove the (weaker) convergence in mean square, using the assumption of finite second moment \( \langle X_i^2 \rangle < \infty \): as \( N \uparrow +\infty \),

\[ \langle (m_N - \langle X_i \rangle)^2 \rangle \to 0 \]

\section*{B. Central limit theorem}

The first correction to the strong law of large numbers is given by the central limit theorem. That governs the small fluctuations, as we will see below. The central limit theorem speaks about convergence in distribution,

\[ \frac{1}{\sqrt{N}} \sum_i (X_i - \langle X_i \rangle) \to \mathcal{N}(0,1) \]

when \( \text{Var} \ X_i = 1 \). The notation \( \mathcal{N}(0,1) \) stands for a Gaussian random variable with mean zero and variance one (standard normal distribution). Note the rescaling, exactly right to keep the variance at one. We have blown up something that goes to zero \( (m_N - \langle X_i \rangle) \) by multiplying with something that tends to infinity \( (\sqrt{N}) \).

Let us rewrite that statement in terms of deviations: for \( a \leq b \) as \( N \uparrow +\infty \),

\[ \text{Prob} \left[ \frac{a}{\sqrt{N}} < m_N - \langle X_i \rangle < \frac{b}{\sqrt{N}} \right] \to \frac{1}{\sqrt{2\pi}} \int_a^b dx \ e^{-x^2/2} \]

The left and right boundaries in the probability go to zero with \( N \), and we are therefore looking at rather small fluctuations around zero (= law of large numbers).
C. Coin tossing

We start with an example of large deviations. Suppose that $X_i = 0, 1$ with equal probability (fair coin, $\rho(0) = \rho(1) = 0$), and that we inquire about the probability

$$p_N(a, b) := \text{Prob}[a < m_N < b]$$

for some $0 \leq a < b \leq 1$. We will show that

$$\lim_{N \to +\infty} \frac{1}{N} \log p_N(a, b) = -\inf_{a < m < b} [\log 2 - s(m)], \quad s(m) := -m \log m - (1-m) \log(1-m), m \in [0, 1]$$

The reason goes by Stirling’s formula. After all,

$$p_N(a, b) = 2^{-N} \sum_{aN < j < bN} \frac{N!}{(N-j)!j!}$$

so that for $q_N(a, b) := \max_{aN < j < bN} \frac{N!}{(N-j)!j!}$,

$$2^{-N} q_N(a, b) \leq p_N(a, b) \leq 2^{-n(b-a)} q_N(a, b)$$

Therefore, $\lim \frac{1}{N} \log p_N(a, b) = \lim \frac{1}{N} \log q_N(a, b) = \sup_{a < m < b} s(m)$ where the last equality employs Stirling’s approximation. In fact, the relation between entropy functions and binomial coefficients comes from the simple fact that

$$\frac{n!}{(n-k)!k!} \approx \sqrt{\frac{n}{2\pi k(n-k)}} \exp\{n[-\frac{k}{n} \log \frac{k}{n} - \frac{n-k}{n} \log \frac{n-k}{n}]\}$$

So we conclude that (informally),

$$\text{Prob}[m_N \simeq m] \simeq e^{-NI(m)}, \quad I(m) = \log 2 - s(m)$$

where we call $I(m)$ the fluctuation functional. Note immediately that $I(m) \geq 0$ with equality only in $m = 1/2$ and maxima at $I(m = 0, 1) = \log 2$. In that way we have estimated the corrections to the law of large numbers.

A next natural question is to ask about an unfair coin, that is where $X_i = 0$ with probability $1 - p$ and $X_i = 1$ with probability $p$.

Then, the probability for $k$ successes is obtained exactly in the same way as before, but with
a different \textit{a priori} weight: instead of the $2^{-N}$ for the fair coin, we now get $p^N m (1-p)^{N(1-m)}$ and hence

$$\text{Prob } [m_N \simeq m] \simeq e^{-NI_p(m)}, \quad I_p(m) = -m \log p - (1-m) \log (1-p) - s(m)$$

$$= m \log \frac{m}{p} + (1-m) \log \frac{1-m}{1-p} \quad (XX.1)$$

Obviously, $I_p(m) \geq 0$ with equality only if $m = p$.

\textbf{D. Multinomial distribution}

An immediate generalization is to the multinomial case. Here we take a finite state space $K$ so that the random variables take $n = |K|$ possible values (instead of two). We take $N$ copies of the random variable (with distribution $\rho$) and we ask for the probability that a number $m_x$ have the value $x$ (i.e., $m_1$ have a first value, $m_2$ have a second value,... till $m_n$ have the $n$th value, and of course $N = m_1 + m_2 + \ldots + m_n$.) We then again apply Stirling’s formula in the form

$$\lim_{N \uparrow +\infty} \frac{1}{N} \log \frac{N!}{m_1! \ldots m_n!} = - \sum_{i=1}^n p_i \log p_i, \quad \text{under } \frac{m_i}{N} \to p_i \quad (XX.2)$$

We can formulate the outcome of the $N$-sampling using the empirical distribution $m^n$; that is the random probability distribution

$$m^n(x) := \frac{1}{N} \sum_{j=1}^N \delta_{X_j, x}, \quad x \in K$$

(Random because it is function of the $X_1, X_2, \ldots, X_N$ and $m^n$ is a probability distribution-valued random variable). Note that its first moment is

$$\sum_x x m^n(x) = m_N = \frac{1}{N} \sum_{j=1}^N X_j$$

just the arithmetic mean. The probability to observe the proportions $\mu(x)$ when the \textit{a priori} probabilities are $\rho(x)$ is

$$\text{Prob}_\rho [m^n(x) \simeq \mu(x), x \in K] \simeq e^{-NI_\rho(\mu)}$$

with fluctuation functional

$$I_\rho(\mu) = \sum_x \mu(x) \log \frac{\mu(x)}{\rho(x)} =: S(\mu|\rho)$$
the relative entropy between $\mu$ and $\rho$. That result is called Sanov’s theorem. It is a computation that goes back to Boltzmann and follows as before for coin tossing but now from the multinomial formula (XX.2) combined with the a priori probabilities of $\rho(x)^N \mu(x)$ to throw always $x$ in $N \mu(x)$ trials:

$$I_\rho(\mu) = -\sum_x \mu(x) \log \rho(x) + \sum_x \mu(x) \log \mu(x)$$

That extends to the case of the empirical distribution on $N$ identically distributed random variables on $\mathbb{R}$ with common distribution $\rho$, summarizing

$$\text{Prob}\left[\frac{1}{N} \sum_j \delta_{X_j} \simeq \mu\right] \simeq e^{-NS(\mu|\rho)}$$

**XXI. FREE ENERGY METHOD**

It is of course difficult to extend the above methods of explicit calculation to arbitrary a priori distributions. But we can illustrate a more general procedure already for independent variables. Indeed, a method to compute the fluctuation functional $I(m)$ in

$$\text{Prob}_\rho[m_N \simeq m] \simeq e^{-NI(m)}$$

for quite general distributions will be based on the observation that then for all functions $F: K \to \mathbb{R},$

$$\frac{1}{N} \log \langle e^{NF(m_N)} \rangle \simeq \frac{1}{N} \log \sum_m e^{NF(m)} e^{-NI(m)} \simeq \sup_m \{F(m) - I(m)\}$$

In particular, for $F(m) = \theta m$ for some $\theta$, as $N \uparrow \infty$,

$$\frac{1}{N} \log \langle e^{\theta(X_1 + X_2 + \ldots X_N)} \rangle \simeq \sup_m \{\theta m - I(m)\}$$

Therefore, $I(m)$ is the Legendre transform of the (convex) function

$$\psi(\theta) := \lim_N \frac{1}{N} \log \langle e^{\theta(X_1 + X_2 + \ldots X_N)} \rangle = \log \sum_x \rho(x) e^{\theta x}$$

where only the last identity has used the fact that we deal with independent random variables. We have used the notation for discrete random variables having exponential moments (so that $\psi(\theta)$ exists and is finite for all $\theta$):

$$I(m) = \sup_{\theta} [m \theta - \psi(\theta)]$$
The supremum is reached at \( \theta^* = \theta^*(x) \) such that \( \psi'(\theta^*) = m \). Then, the tilted probability distribution

\[
\rho^*(x) := e^{\theta^* x - \psi(\theta^*)} \rho(x)
\]

has average \( \sum_x \rho^*(x) x = m \). Thinking of \( x \) as possible spin values, it is like changing/fixing the magnetization \( m \) by adjusting the magnetic field \( \theta \). The function \( \psi(\theta) \) is then the change in free energy under the application of a magnetic field. In the more mathematical language, \( \psi(\theta) \) is called the log-generating function; its derivatives are given by connected correlation functions.

The statement above, for quite general random variables, is known as the Gärtner-Ellis theorem: if the \( \psi(\theta) \) exists and is finite and differentiable for all \( \theta \), then the fluctuation functional \( I(m) \) can be found as its Legendre transform. In the case that the distribution \( \rho \) does not allow exponential moments, we need to refer to other methods.

Let us combine again with the result that the relative entropy is a large deviation function. Note that \( a \log a \geq a - 1, a > 0 \), so that

\[
\sum_x \mu(x) \log \mu(x) \geq \sum_x \mu(x) \log \rho(x)
\]

(by choosing \( a = \mu(x)/\rho(x) \)). That means that the relative entropy \( S(\mu|\rho) \) is never negative. Let us now choose \( \rho(x) = e^{-V(x)/Z} \) to be an equilibrium distribution for potential \( U \): we note that

\[
s(\mu|\rho) = \sum_x \mu(x) \log \frac{\mu(x)}{\rho(x)} = \sum_x U(x) \mu(x) - s(\mu) + \log Z \geq 0
\]

In other words then, the relative entropy is a difference of the free energy functional \( \mathcal{F}(\mu) := \sum U(x) \mu(x) - s(\mu) \), evaluated in \( \mu \) and \( \rho \) respectively:

\[
S(\mu|\rho) = \mathcal{F}(\mu) - \mathcal{F}(\rho), \quad \mathcal{F}(\rho) = -\log Z
\]

As a consequence we recover the equilibrium distribution from minimizing \( \mathcal{F} \), since \( \mathcal{F}(\mu) \geq \log Z \) for all probability distributions \( \mu \):

\[
\text{Prob}_{\rho}\left[ \frac{1}{N} \sum_j \delta_{X_j \sim \mu} \right] \simeq e^{-N[\mathcal{F}(\mu) - \mathcal{F}(\rho)]}
\]
XXII. FROM LARGE TO SMALL FLUCTUATIONS

Suppose we in fact have a well-defined large deviation property for random variables \(X_1, X_2, \ldots\),

\[
\text{Prob}[m_N \simeq m] \simeq e^{-NI(m)}
\]

Let us assume for simplicity that their mean is \(\langle X_i \rangle = 0\) so that \(I(0) = 0\) and \(I'(0) = 0\) with smooth large deviation functional \(I\). We are now asked to evaluate the small fluctuations

\[
\text{Prob}[m_N \simeq x/\sqrt{N}] \simeq e^{-NI(x/\sqrt{N})}
\]

for which we expand around the mean zero, \(I(x/\sqrt{N}) = x^2 I''(0)/(2N) + o(1/N)\) or,

\[
\text{Prob}[\sqrt{N} m_N \simeq x] \simeq e^{-I''(0)x^2/2}
\]

which is the central limit theorem if the variance of \(\sqrt{N} m_N\) is indeed \(I''(0)\). That is the subject of an exercise. In light of the previous section it is also the second derivative of a free energy, which is in general related to an equilibrium response function or susceptibility.

XXIII. CONTRACTION PRINCIPLE

We can ask for large deviations on several levels. One can ask for the fluctuations in arithmetic means like \(m_N\) or also for

\[
\frac{1}{N}(X_1^2 + X_2^2 + \ldots X_N^2)
\]

etc. but we can also ask for large deviations for the empirical distribution \(m^N\). One can however move between the different levels. We compare for example the results of Sections XX C and XX D. We can indeed verify that

\[
I_p(m) = \inf_{\mu: \langle X \rangle_{\mu} = m} S(\mu|\rho)
\]

where \(\rho(0) = 1 - p, \rho(1) = p\). The infimum is obtained when \(\mu = \rho^*\):

\[
\mu(x) = e^{\theta x - \psi(\theta)} \rho(x)
\]

with \(\theta\) chosen such that

\[
\frac{d}{d\theta} \log(e^{\theta X}) = m
\]
XXIV. ENTROPIES

The expression
\[ S(\mu) = - \sum_x \mu(x) \log \mu(x) \]
is called the Shannon entropy of the probability distribution \( \mu \) on \( K \). The work of Claude Shannon is situated in the upcoming information and communication science of the 1940-1950.

Imagine we have lost our keys. If there are four possible rooms where the keys can be found and with equal chances, then we only need to ask two yes/no questions to find back our keys (or at least to know the room where the keys are waiting for us). We first ask whether the keys are (yes or no) in rooms 1 or 2. When the answer is “yes,” a second question “Are they in room 1?” suffices, and when the answer is “no” the second question is “Are they in room 3?” That two can be written as \( 2 = \log_2 4 = - \log_2 1/4 \), and the information is expressed in terms of a “binary bit.” Generalizing, it appears to make sense to express the information content of \( m \) equally plausible possibilities as \( \log_2 m \). That number is not an integer in general, but we can still think of it as a measure of how many questions we need to ask to locate our keys when lost in one of \( m \) possible rooms. Yet, sometimes we know more about the possibilities and these are not equally plausible. Indeed, it can happen that we estimate the chance to be 1/2 to find the keys in room 1 and to be each time 1/6 to find the keys in room 2,3 or 4. The minimal number of questions can then be expected to be completed...

XXV. MATHEMATICAL FORMULATION

For possible reference and further study we present here the mathematical definition of ”satisfying a large deviation principle.”

A sequence of probability distributions \( \{P_N\} \) on a Polish space (complete separable metric space) \( \mathcal{X} \) satisfies a “large deviation principle” with rate \( N \) and rate function \( I(m) \) if

1. \( I(m) \) is lower semi-continuous, nonnegative and has compact level sets \( \{m : I(m) \leq \ell\} \); 

2. 
\[
\limsup_{N \to \infty} \frac{1}{N} \log P_N(C) \leq - \inf_{m \in C} I(m) \quad \text{for closed } C \subset \mathcal{X}
\]
\[
\liminf_{N \to \infty} \frac{1}{N} \log P_N(G) \geq - \inf_{m \in G} I(m) \quad \text{for open } G \subset X
\]

Note that there is a general strategy for determining the rate function, as a generalization of the “magnetic field” in Section XXI. The point is that we should see how to change the probabilities so that the event \( A \) whose probability is sought becomes typical. If we find a deformation that makes \( A \) typical, then the “price to pay” is the large deviation rate function for \( A \). That is also why large deviation functionals are in fact always relative entropies.

**XXVI. LARGE DEVIATIONS AND PHASE TRANSITIONS**

**XXVII. QUANTUM LARGE DEVIATIONS**

**XXVIII. DYNAMICAL LARGE DEVIATIONS**

Consider time-averages
\[
p_T(x) = \frac{1}{T} \int_0^T \delta[X_t = x] \, dt
\]
(the fraction of time spent in state \( x \)) for an irreducible Markov jump process on a finite set \( K \). We are again interested in a large deviation, but now in time and with the \( T \) growing to infinity. The law of large times says that the time-average \( p_T(x) \to \rho(x) \) to the unique stationary distribution. The large deviations are then written as
\[
P[p_T \simeq \mu] \simeq e^{-TD(\mu)}
\]
where \( D(\mu) \) is called the Donsker-Varadhan functional (1975) and will of course depend on the transition intensities \( k(x,y) \) of the Markov process.

We can actually compute the functional \( D(\mu) \). The point is that we can make any stationary distribution \( \mu > 0 \) from modifying the transition rates into
\[
k_V(x,y) = k(x,y) \exp [V(x) - V(y)]/2
\]
where \( V = V_\mu \) is unique (up to a constant) The dependence on \( \mu \) can be quite complicated and in a way, finding \( V \) is like an inverse problem: we know the stationary \( \mu \) and we must find \( V \) so that
\[
\sum_y [k_V(x,y)\mu(x) - k_V(y,x)\mu(y)] = 0, \quad x \in K
\]
Accepting that, the computation proceeds via the Girsanov formula for the path probabilities (VI.1). Calling the original process $P$ and the modified process $P_V$ we have

$$
\frac{dP}{dP_V}(\omega) = \frac{k(x_0, x_1)}{k_V(x_0, x_1)} \cdots \frac{k(x_{n-1}, x_n)}{k_V(x_{n-1}, x_n)} \exp\left\{ \int_0^T [\xi_V(x_s) - \xi(x_s)] ds \right\}
$$

with modified escape rate $\xi_V(x) := \sum_y k_V(x, y)$. For the ratio’s of the transition rates we simply get

$$
\frac{k(x_0, x_1)}{k_V(x_0, x_1)} \cdots \frac{k(x_{n-1}, x_n)}{k_V(x_{n-1}, x_n)} = \exp[V(x_n) - V(x_0)]
$$

which will not contribute extensively in the $T$–limit for large deviations. On the other hand,

$$
\int_0^T [\xi_V(x_s) - \xi(x_s)] ds = T \sum_x p_T(x)[\xi_V(x) - \xi(x)]
$$

so that

$$
D(\mu) = \sum_x \mu(x) [\xi(x) - \xi_V(x)] = \sum_{x,y} \mu(x) k(x, y) [1 - \exp [V_\mu(x) - V_\mu(y)]/2]
$$

which is determined by the modification of expected escape rates. There is another, more mathematical way, to write the same thing:

$$
D(\mu) = -\inf_{g>0} \sum_x \mu(x) \frac{1}{g(x)} Lg(x)
$$

where the infimum replaces solving the above mentioned inverse problem (take $g(x) = e^{-V(x)/2}$). We do not explain that here further.

Instead, let us first look in equilibrium (under detailed balance): explicitly,

$$
D(\mu) = - \sum_x \rho(x) \sqrt{f(x)} L \sqrt{f(x)}
$$

for $f = d\mu/d\rho$. In fact, here it is easy to find $V_\mu$ by asking detailed balance

$$
\mu(x) k_V(x, y) = \mu(y) k_V(y, x)
$$

and using that already $\rho(x) k(x, y) = k(y, x) \rho(y)$.

The situation gets more complicated (and more interesting) under nonequilibrium. There the $D(\mu)$ becomes related to the entropy production rate, but we ignore further details.
XXIX. FREIDLIN-WENTZEL THEORY

Freidlin-Wentzel theory (1970) develops the theory of random perturbations around deterministic dynamics. Here is how some of it fits in the theory of large deviations. Consider the stochastic differential equation

$$x_\varepsilon(t) = x + \int_0^t b(x_\varepsilon(s)) \, ds + \sqrt{\varepsilon} \, B_t$$

where $B_t$ is $d-$dimensional Brownian motion. It is a small perturbation of a deterministic dynamics. As the noise strength $\varepsilon \to 0$, $x_\varepsilon(t)$ converges to the solution of the ordinary differential equation

$$x(t) = x + \int_0^t b(x(s)) \, ds$$

We are interested in the probabilities of large deviations around the zero noise limit,

$$\text{Prob}[x_\varepsilon(\cdot) \sim f(\cdot)]$$

where $f : [0, T] \to \mathbb{R}^d$ is an arbitrary function with $f(0) = x$.

The answer is again similar to the free energy method. Replace $b(x)$ with a new $c(t, x)$ and the corresponding stochastic differential equation is

$$dx(t) = c(t, x(t)) \, dt + \sqrt{\varepsilon} \, dB_t$$

in which we choose/find $c$ such that $f$ is a solution of

$$\dot{x} = c(t, x(t)) \quad \text{or} \quad c(t, f(t)) = f'(t)$$

We then proceed as in the previous section by computing the density on path space, applying a Girsanov formula but now for diffusions as in Section XV.

An obvious application of such Freidlin-Wentzel estimates is in the theory of nucleation and metastability, where the escape from local minima in a potential landscape is sought due to fluctuations. More generally, transition rate theory is based on it.

XXX. QUESTIONS

1. Discuss large deviations for sums of independent and identically distributed random variables that (sometimes) fail to have exponential moments, such as for sums of exponential
or Gaussian random variables.

2. Verify that $I''(0)$ is related to the variance of the random variable as suggested under Section XXII.

3. Verify the claims of the contraction principle in Section XXIII.

4. Complete the calculation of Section XXIX for finding the large deviation rate function.
XXXI. REFERENCE BOOKS

The material presented so far is so elementary that most text books go quite beyond. Books especially dedicated to the relation between stochastics and physics are also rare. Here are some general texts but much material and lecture notes can also be found on the web.


**Interpretative and foundational aspects** of the role of chance and probability in the natural sciences can be picked up from many sources, such as


Andrey N. Kolmogoroff, *Grundbegriffe der Wahrscheinlichkeitsrechnung*, Julius Springer,
Berlin (1933).


For the theory of large deviations, there are various references such as:

James Bucklew, Large Deviation Techniques in Decision, Simulation, and Estimation.

Thomas Cover and Joy Thomas, Elements of Information Theory.


R. S. Ellis, Entropy, Large Deviations, and Statistical Mechanics (Springer, 2005).

