Applied Stochastic Processes

Imperial College London

Mathematics Department

a.y. 2013/2014

M. Ottobre

Contents

1 Basics of Probability **5**
- 1.1 Conditional probability and independence. 8
- 1.2 Convergence of Random Variables 10

2 Stochastic processes **14**
- 2.1 Stationary Processes 18

3 Markov Chains **19**
- 3.1 Defining Markovianity 19
- 3.2 Time-homogeneous Markov Chains on countable state space . 23

4 Markov Chain Monte Carlo Methods **36**
- 4.1 Simulated Annealing 39
- 4.2 Accept-Reject method 41
- 4.3 Metropolis-Hastings algorithm 42
- 4.4 The Gibbs Sampler 46

5 The Ergodic Theorem **49**
- 5.1 Dynamical Systems 49
- 5.2 Stationary Markov Chains and Canonical Dynamical Systems 51

6 Continuous time Markov processes **54**

7 Brownian Motion **56**
- 7.1 Heuristic motivation for the definition of BM. 57
- 7.2 Rigorous motivation for the definition of BM. 58
- 7.3 Properties of Brownian Motion 59

8 Elements of stochastic calculus and SDEs theory **62**
- 8.1 Stochastic Integrals. 64
  - 8.1.1 Stochastic integral in the Itô sense. 67
  - 8.1.2 Stochastic integral in the Stratonovich sense. 72
- 8.2 SDEs 74
  - 8.2.1 Existence and uniqueness. 74
  - 8.2.2 Examples and methods of solution. 76
  - 8.2.3 Solutions of SDEs are Markov Processes 79
- 8.3 Langevin and Generalized Langevin equation 81

9 Markov Semigroups and their Generator **86**
9.1 Ergodicity for continuous time Markov processes. 91

10 Diffusion Processes **94**
10.1 Definition of diffusion process 95
10.2 Backward Kolmogorov and Fokker-Planck equations 100
10.2.1 Time-homogeneous case 103
10.3 Reversible diffusions and spectral gap inequality 106
10.4 The Langevin equation as an example of Hypocoercive diffusion 112

11 Anomalous diffusion - - - STILL DRAFT **121**

Appendices **126**

A Miscellaneous facts **126**
A.1 Why can we think of white noise as the derivative of Brownian Motion 126
A.2 Gronwall's Lemma 126
A.3 Kolmogorov's Extension Theorem 127

B Elements of Functional Analysis **127**
B.1 Adjoint operator 128
B.2 Strong, weak and weak-* convergence. 130
B.3 Groups of bounded operators and Stone's Theorem 131
B.4 Functional Inequalities in their basic form 131

C Laplace method **132**

12 A Few Exercises **141**

13 Solutions **151**

Warning. The material of Section 1 is very sketchy and it is there mainly to recall some basic definitions and fix the notation for the rest of the course, as standard notions from probability theory are assumed to be a prerequisite. I don't recommend attending this course if you haven't taken a course in Probability before.

Standing assumptions for the rest of the course.

1. We assume that all the measures we deal with have a density, unless otherwise stated.

2. All the Markov processes are time homogeneous (again, unless otherwise stated).

ACKNOWLEDGMENTS

I would like to thank the students of the course that I taught at Imperial College in the Autumn 2013 for comments and remarks that helped improving these notes. In particular, A. Corenflos, J. Kunz and M. R. Majorel. The changes suggested by A.C. have not yet been incorporated in this version...I'm working on it!

1 Basics of Probability

For background material on probability theory see for example [79]. We start by recalling some elementary definitions.

• Let E be a set and $\mathcal{E}$ a collection of subsets of E. $\mathcal{E}$ is a *σ-algebra* of subsets of E if

- $E \in \mathcal{E}$,
- if $A \in \mathcal{E}$ then $A^c \in \mathcal{E}$,
- if $\{A_j\}_{j\in I} \subset \mathcal{E}$ then $\bigcup_{j\in I} A_j \in \mathcal{E}$, where I is an at most countable set of indices.

The pair $(E, \mathcal{E})$ is a *measurable space*. If μ is a measure [1] on E, the triple $(E, \mathcal{E}, \mu)$ is called a *measure space*.
If Γ is a class of subsets of E, we denote by $\sigma(\Gamma)$ the *σ-algebra generated by* Γ, i.e. the smallest σ-algebra containing Γ, which is the intersection of all the σ-algebras containing Γ.
If E is endowed with a topology then the σ-algebra generated by the open sets of E is called the *Borel* σ-algebra and denoted $\mathcal{B}(E)$ or simply $\mathcal{B}$ when there is no risk of confusion. We will often use the σ-algebra $\mathcal{B}(\mathbb{R}^n)$. It is a good exercise to compare the definition of σ-algebra with the definition of topology, which are completely different and come from two very different needs, but they are often mixed up.

• If $\mathbb{P}$ is a measure on a set Ω such that $\mathbb{P}(\Omega) = 1$, we call $\mathbb{P}$ a *probability measure*. The triple $(\Omega, \mathcal{F}, \mathbb{P})$, with $\mathcal{F}$ a σ-algebra on Ω, is then a *probability space*. Ω is the *sample space*, points $\omega \in \Omega$ are *sample points* and elements $A \in \mathcal{F}$ are *events*.

• Let $(E, \mathcal{E})$, $(E', \mathcal{E}')$ be two measurable spaces. A map $X : E \to E'$ is *$\mathcal{E}/\mathcal{E}'$-measurable* if the preimage of any $A' \in \mathcal{E}'$ is a set $A \in \mathcal{E}$, i.e.

$$X^{-1}(A') \in \mathcal{E}, \qquad \text{for all } A' \in \mathcal{E}'.$$

When the measurable space at hand is indeed a probability space then a $\mathcal{F}/\mathcal{E}'$-measurable map $X : \Omega \to E'$ is a *E'-valued random variable*. To fix ideas, from now on we will mainly work with $\mathbb{R}^n$-valued random variables, where $\mathbb{R}^n$ is assumed to be endowed with the Borel σ-algebra $\mathcal{B}(\mathbb{R}^n)$. Also, I will not repeat every time that $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space.

[1] Countably additive and non-negative set function with $\mu(\emptyset) = 0$.

• Let $X : \Omega \to \mathbb{R}^n$ be a random variable. The collection of sets

$$\sigma(X) := \{X^{-1}(B), B \in \mathcal{B}\}$$

is a σ-algebra (check), called the σ- *algebra generated by* X. It is the smallest sub-σ-algebra of $\mathcal{F}$ with respect to which X is measurable.

• Let $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathbb{R}^n, \mathcal{B})$ be a r.v. The *law* of X is the map $\mathcal{L}_X : \mathcal{B} \to [0, 1]$ defined as follows:

$$\mathcal{L}_X(B) = \mathbb{P}(X^{-1}(B)), \qquad \text{for all } B \in \mathcal{B}.$$

$\mathcal{L}_X$ is a probability measure on $\mathbb{R}^n$. Therefore a random variable induces, through its law, a probability measure on $\mathbb{R}^n$.
The function $F_X : \mathbb{R}^n \to [0, 1]$ defined as

$$F_X(x) := \mathbb{P}(X \leq x),$$

is the *distribution function* of X. If the range of X is discrete then X is a discrete r.v. and in this case its distribution function will be discontinuous. If instead the distribution function of X is continuous then it is customary to say that X is a continuous r.v. However, in order to avoid confusion with the terminology for stochastic processes, we will more often simply say that X is a r.v. with continuous distribution.

• If there exists a nonnegative integrable function $f : \mathbb{R}^n \to \mathbb{R}$ such that

$$F_X(x) = \int_{-\infty}^{x_n} \dots \int_{-\infty}^{x_1} f(z_1, \dots, z_n) dz_1 \dots dz_n,$$

then f is the *density* of X. In the above we used the notation $\mathbb{R}^n \ni x = (x_1, \dots, x_n)$.
If X admits a density function then

$$\mathbb{P}(X \in B) = \int_B f(x) dx.$$

• If $X_1, \dots, X_m$ are $\mathbb{R}^n$ valued r.v., their *joint distribution function* is the function $F_{X_1,\dots,X_m} : (\mathbb{R}^n)^m \to [0, 1]$,

$$F_{X_1,\dots,X_m} := \mathbb{P}(X_1 \leq x_1, \dots, X_m \leq x_m),$$

where this time $x_j \in \mathbb{R}^n$.

- We also recall that the *expectation* of X, $\mathbb{E}(X)$ is

$$\mathbb{E}(X) := \int_\Omega X \, d\mathbb{P} \tag{1}$$

so that

$$\mathbb{P}(X \in B) = \mathbb{E}(\mathbf{1}_B),$$

where $\mathbf{1}_B$ is the characteristic function of the set B. X is *integrable* if $\mathbb{E}(|X|) < \infty$. As you can imagine, formula (1) tends to be quite unpractical to use for calculations. However, if X has a density, then the following holds.

Lemma 1.1. *With the notation introduced above, let $g : \mathbb{R}^n \to \mathbb{R}^m$ be a measurable function such that $Y = g(X)$ is integrable. Then*

$$\mathbb{E}(Y) = \int_{\mathbb{R}^n} g(x) f(x) dx$$

and in particular

$$\mathbb{E}(X) = \int_{\mathbb{R}^n} x f(x) dx.$$

If $\mu = \mathbb{E}(X)$ then we define the *variance* of X to be

$$Var(X) := \mathbb{E}\,|X - \mu|^2 = \int |x - \mu|^2 \, f(x) dx,$$

where $|\cdot|$ denotes the euclidean norm. The *covariance* of two r.v. X and Y is instead

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

Recall also that if two r.v. have joint distribution $F_{X,Y}$ and joint density function $f_{X,Y}$ then

$$\mathbb{E}(g(X, Y)) = \iint g(x, y) f_{X,Y}(x, y) dx dy.$$

Moreover, the *Convolution Theorem* states that in such a case the density function of $Z = X + Y$ is given by

$$f_Z(z) = \int f_{X,Y}(x, z - x) \, dx \, .$$

Example 1.2. If $X : \Omega \to \mathbb{R}$ is a r.v. with density

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

then X is a *Gaussian (normal)* r.v. with mean m and variance σ^2. In this case we use the notation $X \sim \mathcal{N}(m, \sigma^2)$.

Example 1.3. Analogously in higher dimension. If $X : \Omega \to \mathbb{R}^n$ is a r.v. with density

$$f(x) = \frac{1}{\sqrt{(2\pi)^n \det(C)}} e^{-\frac{1}{2}(x-m)\cdot C^{-1}(x-m)},$$

for some $m \in \mathbb{R}^n$ and some positive definite matrix C then X is a *Gaussian (normal)* r.v. with mean m and covariance matrix C.

Finally, we recall that the space $L^1((\Omega, \mathcal{F}, \mathbb{P}); \mathbb{R}^n)$ (most often I will just write $L^1(\Omega, \mathcal{F}, \mathbb{P})$ or $L^1(\mathbb{P})$) is the space of $\mathcal{F}$-measurable integrable $\mathbb{R}^n$-valued . Analogously, for all $p \geq 1$, $L^p((\Omega, \mathcal{F}, \mathbb{P}); \mathbb{R}^n)$ is the space of $\mathcal{F}$-measurable functions $X : \Omega \to \mathbb{R}^n$ such that

$$\int_\Omega |X(\omega)|^p \, d\mathbb{P}(\omega) < \infty \,.$$

If the random variable X has a density (i.e. if the distribution function of X has a density) $f(x)$, then the above integral can just be rewritten as

$$M_p(f) := \int_{\mathbb{R}^n} |x|^p \, f(x) dx.$$

$M_p(f)$ is the *p-th (non-centered) moment* of the r.c. with density f.

1.1 Conditional probability and independence.

If $A, B \in \mathcal{F}$ and $\mathbb{P}(B) > 0$ we define the conditional probability of A given B to be

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

Any two events $A, B \in \mathcal{F}$ are *independent* if

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

A collection $\{X_i\}_i$ of $\mathbb{R}^n$ valued r.v. is independent if

$$\mathbb{P}(X_1 \in B_1, \ldots, X_k \in B_k) = \mathbb{P}(X_1 \in B_1) \cdot \ldots \cdot \mathbb{P}(X_k \in B_k)$$

for any $k \geq 2$ and any Borel sets B_i. If two real valued r.v. X and Y are independent then

$$\mathbb{E}(XY) = \mathbb{E}(X) \cdot \mathbb{E}(Y) \qquad \text{hence } \; Cov(X, Y) = 0$$

and

$$Var(X + Y) = Var(X) + Var(Y).$$

Let us now recall the basic facts about conditional expectation.

Definition 1.4. *Let X be a $\mathbb{R}^n$ valued random variable, $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ and let $\mathcal{G}$ be a sub-sigma algebra of $\mathcal{F}$. Then there exists a unique integrable and $\mathcal{G}$-measurable random variable Y such that*

$$\mathbb{E}(Y\mathbf{1}_G) = \mathbb{E}(X\mathbf{1}_G) \qquad \forall\, G \in \mathcal{G}. \tag{2}$$

The r.v. Y is (a version of) the conditional expectation of X given $\mathcal{G}$; we use the notation $Y =: \mathbb{E}(X|\mathcal{G})$. Also, if Z is a r.v. the notation $\mathbb{E}(X|Z)$ is a short notation for $\mathbb{E}(X|\sigma(Z))$.

If $X \in L^2(\Omega, \mathcal{F}, \mathbb{P})$ then there is a clear interpretation for the conditional expectation in terms of projections; indeed in this case if $\mathcal{G}$ is a sub-sigma algebra of $\mathcal{F}$ then $L^2(\Omega, \mathcal{G}, \mathbb{P})$ is a proper closed subspace of $L^2(\Omega, \mathcal{F}, \mathbb{P})$. Therefore for any $X \in L^2(\Omega, \mathcal{F}, \mathbb{P})$ there exists a unique $Y \in L^2(\Omega, \mathcal{G}, \mathbb{P}) =: \mathcal{K}$ such that

$$\|X - Y\|_{L^2} = \inf_{W \in \mathcal{K}} \|X - W\|_{L^2} \qquad \text{and} \qquad X - Y \perp W \quad \forall\, W \in \mathcal{K},$$

i.e. Y is the unique orthogonal projection of X on $\mathcal{K}$. In other words, among all the $\mathcal{G}$-measurable functions, Y is the best (in the L^2 sense) estimator of X. Rephrasing: if the information contained in the σ-algebra $\mathcal{G}$ is available, Y is our best guess on X. We list the main properties of the conditional expectation:

i) If $Y = \mathbb{E}(X|\mathcal{G})$ then $\mathbb{E}(Y) = \mathbb{E}(X)$ (which follows from (2))

ii) Take out what is known: if Z is $\mathcal{G}$-measurable then $\mathbb{E}(ZX|\mathcal{G}) = Z\mathbb{E}(X|\mathcal{G})$ and in particular $\mathbb{E}(Z|\mathcal{G}) = Z$

iii) Tower property: if $\mathcal{H} \subset \mathcal{G} \subset \mathcal{F}$ then

$$\mathbb{E}[\mathbb{E}(X|\mathcal{G})|\mathcal{H}] = \mathbb{E}[\mathbb{E}(X|\mathcal{H})|\mathcal{G}] = \mathbb{E}(X|\mathcal{H})$$

iv) If X is independent of Z then $\mathbb{E}(X|Z) = \mathbb{E}(X)$.

In the same way, for any event $A \in \mathcal{F}$ and any r.v. X, we can define

$$\mathbb{P}(A|X) := \mathbb{P}(A|\sigma(X)) := \mathbb{E}(\mathbf{1}_A|X).$$

It is a general fact that, for any two r.v., $\mathbb{E}(Z|X)$ can be written as a measurable function of X, i.e. there exists a measurable function h such that

$$\mathbb{E}(Z|X) = h(X).$$

However the function h is defined only up to a set of measure zero. When we apply this reasoning to $\mathbb{P}(A|X)$, we obtain that

$$\mathbb{P}(A|X) = h(X), \tag{3}$$

for some measurable function h. We might suggestively write, and in fact we shall do so,

$$\mathbb{P}(A|X = x) = h(x);$$

however the above expression is intended to hold for almost every x.

1.2 Convergence of Random Variables

First, let us list the main modes of convergence that we will be using.

Definition 1.5. *For simplicity let $X_n, X : \Omega \to \mathbb{R}$ be real valued random variables, (but clearly all the definitions below apply to $\mathbb{R}^n$-valued r.v.) and let ν_n and ν be the law of X_n and X. We will assume that ν_n and ν have a density* [2] *and with abuse of notation we will write $d\nu_n(x) = \nu_n(x)dx$ and $d\nu(x) = \nu(x)dx$.*

- $X_n \overset{a.s.}{\to} X$ almost surely *if*

$$X_n(\omega) \longrightarrow X(\omega) \quad \textit{for a.e. } \omega, \quad \textit{i.e.} \quad \mathbb{P}\left(\lim_{n\to\infty} X_n = X\right) = 1.$$

- $X_n \overset{L^p}{\to} X$, *where* $L^p := L^p(\Omega; \mathbb{R})$, $p \geq 1$, *if*

$$\mathbb{E}\,|X_n - X|^p \to 0.$$

- $X_n \overset{p}{\to} X$ in probability *if for any* $\epsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \epsilon) = 0.$$

- $X_n \overset{d}{\to} X$ *(or also* $X_n \overset{\mathcal{D}}{\longrightarrow} X$ *or* $X_n \rightharpoonup X$*)* in distribution *or* weakly *if*

$$\int_{\mathbb{R}} h(x)\nu_n(x)dx \longrightarrow \int_{\mathbb{R}} h(x)\nu(x)dx \qquad \textit{for any } h \in C_b(\mathbb{R}).$$

[2] We say that a probability measure μ on $(\mathbb{R}^n, \mathcal{B})$ has a density if there exists a non-negative function $f(x)$ such that $\nu(B) = \int_B f(x)dx$ for any $B \in \mathcal{B}$.

An important fact to notice about weak convergence is that this definition makes sense even if the random variables that come into play are not defined on the same probability space.

The above modes of convergence of r.v. are related as follows:

$$a.s. \implies \text{in probability} \implies \text{in distribution}$$

and

$$\text{in } L^p \implies \text{in probability} \quad .$$

The latter implication is a consequence of the **Markov Inequality**:

$$\mathbb{P}(X \geq c) \leq \frac{1}{g(c)} \mathbb{E}[g(X)], \tag{4}$$

for any r.v. X, for all $c \in \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}_+$ measurable and non-decreasing. A particular case of the Markov Inequality is the **Chebyshev's inequality**: for all $c \geq 0$ and for any square integrable r.v. X,

$$\mathbb{P}(|X - \mu| > c) \leq \frac{Var(X)}{c^2}, \quad \text{where } \mu := \mathbb{E}X \, .$$

In general $a.s.$ convergence and convergence in L^p are not related at all. However, we recall the following classic results.

- **Monotone Convergence Theorem (MCT).** If $0 \leq X_n \uparrow X$ almost surely, then $\mathbb{E}X_n \uparrow \mathbb{E}(X)$.

- **Dominated Convergence Theorem (DCT).** If $X_n \to X$ $a.s.$ and there exists a integrable r.v. Y s.t. $|X_n| \leq Y$ then

$$\mathbb{E}(|X_n - X|) \to 0 \tag{5}$$

  and therefore also

$$\mathbb{E}(X_n) \to \mathbb{E}(X). \tag{6}$$

- **Bounded Convergence Theorem (BCT).** If $X_n \to X$ $a.s.$ and there exists a constant $K > 0$ independent of n such that $|X_n| \leq K$ then $\mathbb{E}(|X_n - X|) \to 0$.

- **Uniform Integrability criterion for L^1 convergence (UIC).** Let X_n, X be integrable r.v. Then $X_n \to X$ in L^1 if and only if the following two conditions hold:

  i) $X_n \to X$ in probability

ii) X_n is uniformly integrable, i.e. for all $\epsilon > 0$ there exists $K > 0$ s.t.

$$\mathbb{E}(|X_n| \mathbf{1}_{\{|X_n|>K\}}) < \epsilon, \qquad \forall n \in \mathbb{N}.$$

- **Scheffé's Lemma.** Let X_n, X be integrable and suppose $X_n \to X$ *a.s.* Then

$$\mathbb{E}|X_n| \to \mathbb{E}|X| \iff \mathbb{E}(|X_n - X|) \to 0.$$

Remark 1.6 (On the above theorems)**.** In the DCT and in the BCT the inequalities $|X_n| \leq Y$ and $|X_n| \leq K$ are meant to hold pointwise, i.e. for all n and ω. In the DCT, (6) follows from (5) simply because $|\mathbb{E}(X_n - X)| \leq \mathbb{E}|X_n - X|$. The BCT is clearly a consequence of the DCT. Both the BCT and the DCT hold even if the assumption on the *a.s.* convergence of X_n to X is replaced by convergence in probability. In this respect, if X_n, X are assumed to be integrable, it is clear that DCT and BCT are a consequence of the UIC. Last, because $||a| - |b|| \leq |a - b|$, the implication " $\Leftarrow$ " in Scheffé's Lemma is trivial. The opposite implication is the nontrivial one.

Now some remarks about weak convergence. According to the famous Portmanteau Theorem (see for example [8]) convergence in distribution can be equivalently formulated as follows:

Theorem 1.7 (Portmanteau Theorem)**.** *Let X_n be a sequence of real valued r.v. with distribution functions $F_n(x)$ and $F(x)$, respectively. Then*

$$X_n \xrightarrow{d} X \quad \Leftrightarrow \quad \lim_{n\to\infty} F_n(x) = F(x)$$

for every point of continuity of F.

While we do not show the proof of the above theorem, we do hope that the following example is somewhat enlightening.

Example. If $X_n = 1/n$ then ν_n, the law of X_n, is the unit mass at $1/n$. Clearly $X_n \rightharpoonup X$, where the law of X, ν, is the unit mass at 0; indeed

$$\int_{\mathbb{R}} h(x) d\nu_n = h(1/n) \to h(0) = \int_{\mathbb{R}} h(x) d\nu \quad \forall h \in C_b(\mathbb{R}).$$

However 0 is a discontinuity point for F and

$$F_n(0) = \mathbb{P}(X_n \leq 0) = 0 \neq 1 = F(0) = \mathbb{P}(X \leq 0).$$

Other facts that you might want to bear in mind:

- If $X_n \xrightarrow{d} X$ and X is a constant (i.e. it is deterministic) then convergence in probability and weak convergence are equivalent.
- **Slutsky's Theorem.** If $X_n \xrightarrow{p} c$, where c is a constant, then $g(X_n) \xrightarrow{p} g(c)$ for any function g that is continuous at c.
- **Cramér's Theorem.** Suppose $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$, where c is deterministic. Then

  i) $X_n + Y_n \xrightarrow{d} X + c$

  ii) $X_n Y_n \xrightarrow{d} cX$

  iii) $X_n/Y_n \xrightarrow{d} X/c$ if $c \neq 0$.
- **Continuous Mapping Theorem.** Let g be a continuous function and suppose $X_n \xrightarrow{d} X$. Then $g(X_n) \xrightarrow{d} g(X)$. The same thing holds for almost sure convergence and for convergence in probability as well.

Definition 1.8 (Weak convergence of probability measures)**.** *Let (M, d) be a separable metric space. A sequence of probability measures μ_n on M is said to converge* weakly *to μ (probability measure on M) if for every continuous and bounded function h on M one has*

$$\int_M h(x)\mu_n(dx) \longrightarrow \int_M h(x)\mu(dx)\,.$$

Remark 1.9. Observe the two following facts.

- Taking μ_n to be the law of a random variable X_n, it is clear that μ_n converges weakly to μ if and only if X_n converges in distribution to X, where the law of the r.v. X is μ.
- In view of Remark 9.10 and the functional analytic definition of weak-* convergence (see Appendix B.2), what probabilists call weak convergence of probability measures should be referred to (and analysts do in fact refer to it) as *weak-* convergence.*

As it is customary, we say that $\{X_i\}_i$ are i.i.d. random variables if they are independent and identically distributed. We recall the following two fundamental results, which you will have already seen in some probability course. Later on we will build on these results and reread them in a more general context.

Theorem 1.10 (Strong Law of Large numbers)**.** *Let $\{X_i\}_{i\in\mathbb{N}}$ be a sequence of i.i.d. integrable random variables with $\mathbb{E}(X_i) = \mu$ and consider*

$$\bar{S}_n := \frac{1}{n}\sum_{i=1}^{n} X_i\,.$$

Then

$$\bar{S}_n \xrightarrow{a.s.} \mu, \quad i.e. \quad \mathbb{P}\left(\lim_{n\to\infty} \bar{S}_n = \mu\right) = 1.$$

Example 1.11 (The simplest example of Monte Carlo Method)**.** Let f be a bounded measurable function $f : [0,1] \to \mathbb{R}$ and X_n be a sequence of i.i.d r.v., uniformly distributed on $[0,1]$. Then the sequence $\{f(X_n)\}_n$ is still a sequence of i.i.d integrable r.v. Applying the LLN we get

$$\frac{1}{n}\sum_{k=1}^{n} f(X_k) \xrightarrow{a.s.} \mathbb{E}(f(X_1)) = \int_0^1 f(x)\,dx.$$

Therefore, once we can generate samples from the variables X_j, the quantity $\frac{1}{n}\sum_{k=1}^{n} f(X_k)$ is, for large n, a good approximation of $\int_0^1 f(x)\,dx$.

Theorem 1.12 (Central Limit Theorem)**.** *Let $\{X_i\}_{i\in\mathbb{N}}$ be a sequence of i.i.d. square integrable random variables with $\mathbb{E}(X_i) = \mu$ and $Var(X_i) = \sigma^2$. Then*

$$\sqrt{n}(\bar{S}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2).$$

Comment. Notice that if we set $S_n = \sum_{j=1}^{n} X_j$, then $\mathbb{E}(S_n) = n\mu$ and $Var(S_n) = n\sigma^2$. Therefore the above theorem is equivalently restated by

$$\frac{S_n - \mathbb{E}(S_n)}{[Var(S_n)]^{1/2}} \xrightarrow{d} \mathcal{N}(0,1),$$

i.e.

$$\lim_{n\to\infty} \mathbb{P}\left(a < \frac{S_n - n\mu}{\sigma\sqrt{n}} < b\right) = \frac{1}{\sqrt{2\pi}}\int_a^b e^{-\frac{x^2}{2}}\,dx, \quad \forall\, a \le b \in \mathbb{R}.$$

2 Stochastic processes

A family $\{X_t\}_{t\in I}$ of S-valued random variables $X_t : \Omega \to S$ is a stochastic process (s.p.). If $I = \mathbb{Z}$ or $I = \mathbb{N}$ then X_t is a *discrete time* stochastic process; if $I = \mathbb{R}$ or $I = \mathbb{R}_+ := \{x \in \mathbb{R} : x \ge 0\}$, it is a *continuous time* stochastic process. During this course the *state space* S will be either $\mathbb{R}^n$

or $\mathbb{Z}$, unless otherwise stated. For a reason that will possibly be more clear after reading Remark 2.1 below, we sometimes consider the process as a whole and use the notation $\mathbf{X} := \{X_t\}_{t\in T}$.

Now notice that

$$X_t(\omega) = X(t,\omega) : I \times \Omega \longrightarrow S.$$

If we fix ω and look at the (non-random) map

$$I \ni t \to X(t,\omega) \in S \quad \text{for fixed } \omega \tag{7}$$

then we are looking at the *path* $X_t(\omega) =: \omega(t)$, i.e. we are observing everything that happens to the sample point ω from time say 0 to time t. If instead we fix t and we look at the map

$$\Omega \ni \omega \to X(t,\omega) \in \mathbb{R}^n \quad \text{for fixed } t,$$

then this is a random variable, which gives us a snapshot of what is happening (although clearly not in a deterministic way) to all the sample points $\omega \in \Omega$ at the time t when we took the picture. The old chestnut of the Eulerian vs Lagrangian point of view. It is sometimes convenient to think of Ω as a set of particles and of the state space S as the physical state space (for example position-velocity), so that $\omega(t)$ is the path in state space followed by the particle ω and $X_t(\omega)$ represents, for fixed t, the state of the system at time t. This will be the point of view that we shall adopt when we talk about non-equilibrium statistical mechanics. In other cases it is more convenient to think of ω as an experiment, so that $X_t(\omega)$ is a realization of such an experiment at time t.

Remark 2.1. Formula (7) offers another perspective on stochastic process: we can look at a stochastic process as a random map from the sample space Ω to the path space $(S)^I := \{\text{maps from } I \text{ to } S\}$. More explicitly

$$\begin{aligned} \mathbf{X} : \Omega &\longrightarrow (S)^I \\ \omega &\longrightarrow \omega(t), \end{aligned}$$

i.e. to each sample point, or particle, we associate a function, its path.

Definition 2.2. *Two stochastic processes $\{X_t\}_t$ and $\{Y_t\}_t$ taking values in the same state space are* (stochastically) equivalent *if $\mathbb{P}(X_t \neq Y_t) = 0$ for all $t \in T$. If $\{X_t\}_t$ and $\{Y_t\}_t$ are stochastically equivalent then $\{X_t\}_t$ is said to*

be a version *of* $\{Y_t\}_t$ *(and the other way around as well). Given a stochastic process* $\{X_t\}_t$ *the family of distributions*

$$\mathbb{P}(X_{t_1} \in B_1, \dots, X_{t_k} \in B_k),$$

for all $k \in \mathbb{N}$, $t_1, \dots, t_k \in T$ *and* $B_1, \dots, B_k \in S$, *are the* finite dimensional distributions *of the process* $\{X_t\}$.

If two stochastic processes are equivalent then they have the same finite dimensional distributions, the converse is not true. Let us make some remarks about these two notions, starting with an example.

Example 2.3. Stochastically equivalent processes can have different realizations. Indeed, consider the processes $\{X_t\}_{t\in[0,1]}$ and $\{Y_t\}_{t\in[0,1]}$ defined as follows

$$X_t \equiv 0 \quad \forall t \quad \text{and} \quad Y_t = \begin{cases} 0 & \text{if } t \neq \tau \\ 1 & \text{if } t = \tau \end{cases}$$

where τ is a random variable with continuous distribution $\tau : \Omega \to [0,1]$, so that $\mathbb{P}(\tau = a) = 0$ for all $a \in [0,1]$. In this case $\mathbb{P}(X_t \neq Y_t) = \mathbb{P}(t = \tau) = 0$. Therefore these two processes are stochastically equivalent but the trajectory of X_t is continuous while the trajectory of Y_t has a discontinuity at $t = \tau$.

One might wonder whether the finite dimensional distributions determine the process uniquely; in general the answer is no, indeed also the finite dimensional distributions don't say anything about the paths. However the Kolmogorov extension Theorem (see for example [56, Theorem 2.1.5]) gives a consistency condition in order for a family of distributions to be the finite dimensional distributions of some stochastic process.

Definition 2.4 (Continuous processes)**.** *Let* $\{X_t\}$ *be a continuous-time s.p. We will say that* $\{X_t\}$ *is continuous if it has continuous paths, i.e. if the maps* $t \to X_t(\omega)$ *are continuous for a.e.* ω.

Given a function we know how to check whether it is continuous or not. Such an exercise might not be so obvious when it comes to stochastic processes. Luckily for us, Kolmogorov provided us with a very useful criterion, which we present for real valued processes but it holds in more generality.

Theorem 2.5 (Kolmogorov's continuity criterion)**.** *Let* $\{X_t\}_{t\geq 0}$ *be a s.p. with state space* $\mathbb{R}$. *If there exist constants* $\alpha, \beta > 0$ *and* $C \geq 0$ *s.t.*

$$\mathbb{E}\,|X(t) - X(s)|^\beta \leq C\,|t - s|^{1+\alpha} \quad \forall\, t, s \geq 0$$

then the process is continuous. More precisely, for any $\gamma \in (0, \alpha/\beta)$ *and* $T > 0$ *and for almost every* ω *there exists a constant* $K = K(\omega, \gamma, T)$ *such that*

$$|X(t,\omega) - X(s,\omega)| \leq K\,|t-s|^{\gamma}\,, \quad \forall\, 0 \leq s,t \leq T.$$

Definition 2.6. *A* filtration *on* $(\Omega, \mathcal{F})$ *is a family of* σ*-algebras* $\{\mathcal{E}_t\}_{t\in I}$ *such that* $\mathcal{E}_t \subset \mathcal{F}$ *for all* $t \in I$ *and*

$$\mathcal{E}_s \subset \mathcal{E}_t \quad \text{if } s \leq t.$$

The process $\{X_t\}_t$ *is* $\mathcal{E}_t$-adapted *if the r.v.* X_t *is* $\mathcal{E}_t$*-measurable for every* t*.*

The most natural filtration to consider is the one generated by the process itself, i.e. the filtration $\mathcal{F}_t = \sigma(\{X_s\}_{0\leq s\leq t})$, as X_t is clearly $\mathcal{F}_t$-adapted. The σ-algebra $\mathcal{F}_t$ contains all the information available to us about the process up to and including time t.

Example 2.7 (Standard Brownian Motion)**.** We define a *Wiener Process* or *Brownian Motion* (BM) to be a real-valued stochastic process $\{B(t)\}_{t\geq 0}$ such that

i) $B(0) = 0$

ii) $B(t) - B(s) \sim \mathcal{N}(0, t-s)$ for all $0 \leq s \leq t$

iii) Increments over non-overlapping time intervals are independent: for all $n \in \mathbb{N}$ and $t_1, \ldots, t_n$ such that $0 \leq t_1 < t_2 < \cdots < t_n$, the increments $B(t_1), B(t_2) - B(t_1), \ldots, B(t_n) - B(t_{n-1})$ are independent.

A few remarks about this definition:

1. From the definition it follows that

   $$\mathbb{E}(B(t)B(s)) = \min(t,s).$$

   Indeed suppose $s \leq t$, then

   $$\mathbb{E}(B(t)B(s)) = \mathbb{E}([B(t) - B(s) + B(s)]B(s)) = \mathbb{E}B(s)^2 = s.$$

   Therefore the process has *stationary increments* (it is not itself stationary though).

2. From ii) it follows that

   $$\mathbb{P}(B(t) \in (a,b)) = \frac{1}{\sqrt{2\pi t}} \int_a^b e^{-\frac{x^2}{2t}}\, dx\,. \tag{8}$$

3. It is possible to prove that $B(t_i) - B(t_{i-1})$ is independent of $\mathcal{F}_{t_{i-1}} = \sigma\{B(s); s \leq t_{i-1}\}$ as a consequence of 1., ii) and iii) above.

A natural extension of the definition of (standard, one dimensional BM) is the following: a *n-dimensional standard BM* is a n-vector $(B_1(t), \ldots, B_n(t))$ of independent one dimensional BMs.

BM is possibly the most important example of this course. For the moment we just give this formal definition. Later on we will introduce BM in a more physically motivated way and justify the definition that we have just given.

2.1 Stationary Processes

We start by working with continuous time stochastic processes $\{X_t\}_{t\in I}$, so for the time being $I \subseteq \mathbb{R}$.

Definition 2.8. *A continuous time stochastic process $\{X_t\}_{t\in I}$ is* strictly stationary*, or simply* stationary*, if its finite dimensional distributions are invariant under time shifts:*

$$\mathbb{P}(X_{t_1} \in B_1, \ldots, X_{t_k} \in B_k) = \mathbb{P}(X_{t_1+h} \in B_1, \ldots, X_{t_k+h} \in B_k)$$

for all $h \in \mathbb{R}$ (such that $h+t_j \in I$), for all $k \in \mathbb{N}$, $t_1, \ldots, t_k \in I$ and $B_1, \ldots, B_k \in S$, where S is the state space of the process X_t.

The intuitive meaning of this definition is readily seen when we take $k = 1$, so that the Definition 2.8 implies that the law of X_t does not depend on t. Stationary processes are therefore used to describe phenomena which happen under conditions that do not change in time.

Definition 2.9. *A continuous time stochastic process $\{X_t\}_{t\in I}$ is* wide sense stationary *(WSS) if it has finite first and second moments and*

1. *$\mathbb{E}(X_t)$ is constant, i.e. it does not depend on t;*
2. *$Cov(X_t X_s)$ is a function of the difference $t - s$.*

The function $Cov(X_t X_s)$ is also called the *autocovariance function* of the process X. To motivate the Definition 2.9, observe that if a process is strictly stationary then it is also WSS; indeed if X_t is stationary and, for simplicity, takes values in $\mathbb{R}$, then we know that $\mathbb{P}(X_t \leq x)$ does not depend on t, hence

$$\mathbb{E}(X_t) = \int x \, d\mathbb{P}(X_t \leq x) \quad \text{is constant in time.}$$

We can now assume, without loss of generality, that $\mathbb{E}(X_t) = 0$ for all t. Then, analogously,

$$\begin{aligned}\mathbb{E}(X_{t+h}X_{s+h}) &= \iint xy\, d\mathbb{P}(X_{t+h} \le x, X_{s+h} \le y) \\ &= \iint xy\, d\mathbb{P}(X_t \le x, X_s \le y) = \mathbb{E}(X_t X_s),\end{aligned}$$

from which we deduce that $\mathbb{E}(X_t X_s)$ depends only on the difference $t - s$. So strictly stationary $\Rightarrow$ WSS. The converse is in general not true. However for Gaussian processes strictly stationary is equivalent to WSS.

Clearly, the definition of stationarity can be given also for discrete-time processes and it is completely analogous to the one given for continuous time processes.

Definition 2.10. *A sequence of random variables $\{X_n\}_{n\in\mathbb{N}}$ is* strictly stationary*, or simply* stationary*, if for every $m \in \mathbb{N}$ and every $k \in \mathbb{N}$, the vector $(X_0, X_1, \dots, X_m)$ has the same distribution as $(X_k, X_{1+k}, \dots, X_{m+k})$.*

The notion of stationarity will pop up several times during this course, as stationary processes enjoy good ergodic properties.

3 Markov Chains

One of the most complete accounts on Markov chains is the book [53]. For MC on discrete state space we suggest [15], which is the approach that we will follow in this section.

3.1 Defining Markovianity

Definition 3.1 (Markov chain)**.** *A discrete-time stochastic process $\{X_n\}_{n\in\mathbb{N}}$, $X_n : (\Omega, \mathcal{F}, \mathbb{P}) \to (S, \mathcal{S})$ on a state space S is a Markov chain if*

$$\mathbb{P}(X_{n+1} \in B | X_0, X_1, \dots, X_n) = \mathbb{P}(X_{n+1} \in B | X_n), \quad \forall\, B \in \mathcal{S},\, n \in \mathbb{N}. \quad (9)$$

If the state space is discrete (countable or finite) we assume that $\mathcal{S}$ is the σ-algebra of all the subsets of S. In the discrete case it is customary to write (9) as

$$\mathbb{P}(X_{n+1} = x_{n+1} | X_0 = x_0, , \dots, X_n = x_n) = \mathbb{P}(X_{n+1} = x_{n+1} | X_n = x_n). \quad (10)$$

However if the state space is countable the above still holds almost everywhere. If the state space is finite then it holds pointwise.

Denoting by $\mathcal{F}_n$ the σ-algebra generated by $X_0, \ldots, X_n$, by definition of conditional expectation (9) can be rewritten as

$$\mathbb{P}(X_{n+1} \in B | \mathcal{F}_n) = \mathbb{P}(X_{n+1} \in B | X_n), \quad \forall\, B \in \mathcal{S},\, n \in \mathbb{N}. \tag{11}$$

Comment. Whether we look at (9) or (10), the moral of the above definition is always the same: suppose that X_n represents the position of a particle moving in state space and that at time n our particle is at point x in state space. In order to know where the particle will be at time $n+1$ (more precisely, the probability for the particle to be at point y at time $n+1$) we don't need to know the history of the particle, i.e. the positions occupied by ω before time n. All we need to know is where we are at time n. In other words, given the present, the future is independent of the past. It could be useful, in order to better understand the idea of Markovianity, to compare it with its deterministic counterpart; indeed the concept of Markovianity is the stochastic equivalent of Cauchy's determinism: consider the deterministic system

$$\dot{z}(t) = f(z), \quad z(t_0) = z_0. \tag{12}$$

We all know that under technical assumptions on the function f there exists a unique solution $z(t)$, $t \geq t_0$, to the above equation; i.e. given an evolution law f and an initial datum z_0 we can tell the future of $z(t)$ for all $t \geq t_0$. And there is no need to know what happened to $z(t)$ before time t_0. Markovianity is the same thing, just reread in a stochastic way: for the deterministic system (12) we know *exactly* where z will be a time t; for a Markov chain, given an initial position (or an initial distribution) at time n_0, we will know the probability of finding the system in a certain state for every time $n \geq n_0$.

Notation. If the chain is started at x then we use the notation

$$\mathbb{P}_x(X_n \in B) := \mathbb{P}(X_n \in B | X_0 = x);$$

if instead the initial position of the chain is not deterministic but drawn at random from a certain probability distribution μ (μ is a probability on S) then we write $\mathbb{P}_\mu(X_n \in B)$. Clearly $\mathbb{P}_x = \mathbb{P}_{\delta_x}$.

If the MC has initial distribution μ, the finite dimensional distributions of

a Markov Chain can be expressed through the relation

$$\begin{aligned}&\mathbb{P}_\mu(X_0 \in B_0, X_1 \in B_1, \dots, X_n \in B_n)\\ &= \int_{B_0} \mu(dy_0) \int_{B_1} \mathbb{P}(X_1 \in dy_1 | X_0 = y_0) \dots \int_{B_n} \mathbb{P}(X_n \in dy_n | X_{n-1} = y_{n-1}).\end{aligned} \tag{13}$$

During this course we will be mainly concerned with a special class of MC, i.e. time-homogoneous MC. For these processes the probability of going from x to y in one time step depends only on x and y and not on when we are in x. In other words, the one-step transition probabilities do not depend on time.

Definition 3.2 (Time-homogeneous Markov Chain)**.** *A Markov chain (MC) $\{X_n\}$ is time-homogeneous if*

$$\mathbb{P}(X_{n+1} \in B | X_n) = \mathbb{P}(X_1 \in B | X_0) \quad \forall n \geq 0. \tag{14}$$

Remark 3.3. To be more precise, one can prove that the above formula (13) is *equivalent* to the Markov property (9):

$$(9) \iff (13)\,.$$

In words: a time homogeneous Markov process has finite dimensional distributions of the form (13). Viceversa, if a stochastic process has finite dimensional distributions of the form (13), then it is a time-homogeneous Markov process .

In the rest of Section 3 (actually, in the rest of this entire course), we will always assume that we are dealing with time-homogeneous MC, unless otherwise explicitly stated. The Markov property and the time-homogeneity imply that we can write

$$\mathbb{P}(X_{n+1} \in B | X_n = x) =: p(x, B) \tag{15}$$

for some function $p : S \times \mathcal{S} \to [0, 1]$.

Definition 3.4. *A map $p : S \times \mathcal{S} \to [0, 1]$ enjoying the following properties*

1. *For fixed $x \in S$, $p(x, \cdot) : \mathcal{S} \to [0, 1]$ is a probability measure, meaning that $p(x, S) = 1$,*
2. *For fixed $B \in \mathcal{S}$, $p(\cdot, B)$ is a measurable map,*

is called a Markov transition function *or a family of* transition probabilities*. If, in addition to 1 and 2, the relation* (15) *holds, p are the transition probabilities of the Markov chain* X_n*.*

Using the transition probabilities we can rewrite the finite dimensional distributions (13) of the MC as

$$\mathbb{P}_\mu(X_0 \in B_0, X_1 \in B_1, \dots, X_n \in B_n) = \int_{B_0} \mu(dy_0) \int_{B_1} p(y_0, dy_1) \dots \int_{B_n} p(y_{n-1}, dy_n).$$

Therefore, to a Markov process we can associate a family of transition probabilities i.e. a family of functions fulfilling the requirements of Definition 3.4. The above formula also says that once we have an initial distribution, the transition probabilities are all we need in order to know the evolution of the chain. This means that we could have introduced Markov processes working the other way around, i.e. starting from the transition probabilities, as the following theorem states.

Theorem 3.5. *For any initial measure* μ *on* $(S, \mathcal{S})$ *and for any family of transition probabilities* $\{p(x, A) : x \in S, A \in \mathcal{S}\}$*, there exists a stochastic process* $\{X_n\}_{n\in\mathbb{N}}$ *such that*

$$\mathbb{P}_\mu(X_0 \in B_0, X_1 \in B_1, \dots, X_n \in B_n) = \int_{B_0} \mu(dy_0) \int_{B_1} p(y_0, dy_1) \dots \int_{B_n} p(y_{n-1}, dy_n). \tag{16}$$

Thanks to Theorem 3.5, an alternative way - alternative to Definition 3.1 - of defining a MC is as follows: we first assign a family of transition probabilities and then we say that X_n is a Markov Chain if

$$\mathbb{P}(X_{n+1} \in B | \mathcal{F}_n) = p(X_n, B), \quad \forall n \in \mathbb{N} \text{ and } B \in \mathcal{S}.$$

At this point, as before (see Remark 3.3), one can prove that the finite dimensional distributions of X_n are given by (16) and also, viceversa, that a process X_n with finite dimensional distributions given by (16) is a Markov chain with transition probabilities $p(x, A)$.

Remark 3.6. Before reading this Remark you should revise the content of the Kolmogorov extension Theorem in Appendix A.

Once we assign an initial distribution μ and a family of transition probabilities, the finite dimensional distributions (16) of the chain X_k (say each r.v. of the chain is real valued) are a consistent family of probability measures on $\mathbb{R}^n$. Therefore the Kolmogorov extension Theorem applies and

there exists a probability measure P on sequence space $\mathbb{R}^{\mathbb{N}}$ (equipped with the σ-algebra $\mathcal{R}^{\mathbb{N}}$ generated by the cylinder sets) such that for all $m \in \mathbb{N}$ and all $A_i \in \mathcal{B}(\mathbb{R})$

$$P(\omega \in \mathbb{R}^{\mathbb{N}} : \omega_i \in A_i, i = 1, \dots, m) = \mathbb{P}(X_1 \in A_1, \dots, X_m \in A_m).$$

In other words the coordinate maps of $(\mathbb{R}^{\mathbb{N}}, \mathcal{R}^{\mathbb{N}}, P)$, i.e. the maps $Y_n : (\mathbb{R}^{\mathbb{N}}, \mathcal{R}^{\mathbb{N}}, P) \to \mathbb{R}$ defined as $Y_n(\omega) = Y_n(\omega_0, \omega_1, \dots) = \omega_n$, have the same finite dimensional distributions as the chain X_n on $(\Omega, \mathcal{F}, \mathbb{P})$. We have therefore found a representation of our chain in sequence space.

3.2 Time-homogeneous Markov Chains on countable state space

In the remainder of this chapter we will assume that the Markov Chain X_n that we are dealing with is time-homogeneous and that it takes values on a countable state space S. Each $x \in S$ is called a *state* of the chain. We endow S with the σ-algebra $\mathcal{S}$ of all the subsets of S. To fix ideas you can think of $S = \mathbb{Z}$. The proofs that we will skip can be found in [15]. For this class of chains we have:

- Transition probabilities: in this case it suffices to assign the *transition matrix* [3] $p = \{p(x, y),\ x, y \in S\}$ where each map $p : S \times S \to [0, 1]$ satisfies

  $$\sum_{y \in S} p(x, y) = 1 \quad \text{and} \quad p(x, y) \geq 0, \quad \forall x, y \in S.^{4}$$

  Clearly, $p(x, y) := \mathbb{P}(X_1 = y | X_0 = x) = \mathbb{P}_x(X_1 = y)$ and for all $n \geq 1$ we denote $p^n(x, y) := \mathbb{P}_x(X_n = y)$.

- Markov property: $\mathbb{P}(X_{n+1} = x_{n+1} | X_n = x_n, \dots, X_0 = x_0) = p(x_n, x_{n+1})$.[5]

- Finite dimensional distributions:

  $$\mathbb{P}_\mu(X_0 = x_0, \dots, X_n = x_n) = \mu(x_0) p(x_0, x_1) \cdot \dots \cdot p(x_{n-1}, x_n).$$

[3] p will actually be a matrix if the state space is finite, it will be an "infinite matrix" if the state space is countable.

[4] The first condition says that p is a *stochastic matrix*. Here we don't need to specify that $p(x, y)$ is measurable in the first argument because with the chosen σ-algebra this is automatically true.

[5] Observe that this equality contains both the loss of memory and the time-homogeneity.

The next theorem gives an extremely important property of the MC.

Theorem 3.7 (Chapman-Kolmogorov equation)**.** *Let X_n be a time-homogeneous Markov chain with discrete state space. Then for any $m, n \geq 0$,*

$$\mathbb{P}_x(X_{n+m} = y) = \sum_{z \in S} \mathbb{P}_x(X_n = z)\mathbb{P}_z(X_m = y).$$

Proof. We recall the conditional version of the law of total probability:

$$\mathbb{P}(A|B) = \sum_j \mathbb{P}(A|C_j \cap B)\mathbb{P}(C_j|B)\,.$$

Using this fact,

$$\begin{aligned}\mathbb{P}(X_{n+m} = y|X_0 = x) &= \sum_{z \in S} \mathbb{P}(X_{n+m} = y|X_m = z, X_0 = x)\mathbb{P}(X_m = z|X_0 = x) \\ &= \sum_{z \in S} \mathbb{P}(X_{n+m} = y|X_m = z)\mathbb{P}(X_m = z|X_0 = x),\end{aligned}$$

having used the Markov property in the second equality. □

Example 3.8 (Random Walk on the integers)**.** Let $\xi_1, \xi_2, \ldots$ be i.i.d. random variables taking values in $\mathbb{Z}$ and with $\Gamma(j) = \mathbb{P}(\xi_n = j)$. The random walk Φ_n is defined as $\Phi_n = \Phi_{n-1} + \xi_n$, $n \geq 1$. Let us calculate the transition probabilities of the chain:

$$\begin{aligned}\mathbb{P}(\Phi_1 = y|\Phi_0 = x) &= \mathbb{P}(\Phi_0 + \xi_1 = y|\Phi_0 = x) \\ &= \mathbb{P}(x + \xi_1 = y) = \Gamma(y - x).\end{aligned}$$

Therefore $p(x, y) = \Gamma(x - y)$. Notice that the transition probability of going from x to y depends only on the increment $x - y$ and not on x and y, i.e. the random walk is translation invariant.

Example 3.9 (Renewal Chain)**.** This time the state space is $\mathbb{N}$. We define the chain through its transition probabilities as follows: given a sequence of positive numbers $a_k \geq 0$ such that $\sum_{k \geq 0} a_k = 1$,

$$\begin{aligned}&p(k, k-1) = 1 \quad \text{if} \quad k \geq 1 \\ &p(0, k) = a_k \qquad \text{for all} \quad k \geq 0 \\ &p(j, k) = 0 \qquad \text{otherwise.}\end{aligned}$$

Example 3.10 (Ehrenfest chain)**.** A box contains N air molecules. The box is divided in two chambers, that communicate through a small hole. The state of the system is determined once we know the number k of molecules that are contained say in the left chamber at each moment in time. Ehrenfest modelled the evolution of the system through a Markov chain defined as follows: suppose that at time n there are k molecules in the left chamber. Assuming that only one molecule per time step can go through the hole, at time $n+1$ either one molecule has gone from left to right (in which case we will end up with $k-1$ particles on the left) or one molecule has gone from right to left (leaving us with $k+1$ particles on the left.) In other words, Ehrenfest modelled the behaviour of gas molecules by using one of the classical "urn problems" of probability theory. With this in mind, the transition probabilities of the chain on state space $S = \{0, 1, \dots, N\}$, are given by

$$p(k, k-1) = k/N \quad \text{and} \quad p(k, k+1) = (N-k)/N, \quad \text{for all } k \geq 0,$$

while $p(j, k) = 0$ otherwise.

Definition 3.11. *For every set $A \subset S$ we define*

$$\tau_A := \inf\{n \geq 0 : X_n \in A\} \quad \textit{and} \quad T_A := \inf\{n \geq 1 : X_n \in A\},$$

to be the hitting time of A *and the* time of first return to A*, respectively. With obvious extension of notation, $\tau_x := \tau_{\{x\}}$ and $T_x := T_{\{x\}}$, for all $x \in S$. For all $k \geq 0$, the* time of k-th return to x *can be defined recursively:*

$$T_x^0 := 0 \quad \textit{and} \quad T_x^k := \inf\{n > T_x^{k-1} : X_n = x\}, \quad \forall k \geq 1.$$

In this way $T_x^1 = T_x$. Moreover we let

$$\rho_{xy} := \mathbb{P}_x(T_y < \infty)$$

and we say that x communicates *with y (in symbols, $x \longrightarrow y$) if $\rho_{xy} > 0$.* [6] *Finally we denote by $N(y)$ the number of visits to y, i.e.*

$$N(y) := \sum_{n=1}^{\infty} \mathbf{1}_{(X_n = y)}.$$

[6] Most textbooks, especially the most control-theory oriented, will say that y is accessible from x if there exists $n \in \mathbb{N}, n \geq 0$, such that $p^n(x, y) > 0$ and then will say that x and y communicate if x is accessible from y and y is accessible from x. Notice that our definition is slightly different as ρ_{xy} is defined through the time of first return, as opposed to being defined through the hitting time.

We notice without proof that

$$\mathbb{P}_x(T_y^k < \infty) = \rho_{xy}\rho_{yy}^{k-1}. \tag{17}$$

Now a few definitions regarding the classification of the states of the chain.

Definition 3.12. *We say that the state $y \in S$ is*

- Recurrent *if* $\rho_{yy} = 1$,
- Transient *if* $\rho_{yy} < 1$,
- Positive recurrent *if* $\rho_{yy} = 1$ *and* $\mathbb{E}_y(T_y) < \infty$,
- Null recurrent *if* $\rho_{yy} = 1$ *and* $\mathbb{E}_y(T_y) = \infty$.

A state $y \in S$ is absorbent *if* $\mathbb{P}_y(X_1 = y) = 1$.

Because for any given state y, we can only either have $\rho_{yy} = 1$ or $\rho_{yy} < 1$, the state of a chain can only be either recurrent or transient. In the first case we know from (17) that $\mathbb{P}_y(T_y^k < \infty) = 1$ for all $k \geq 1$. Therefore the chain will return to y infinitely many times (more precisely, $\mathbb{P}(X_n = y \text{ } i.o.) = 1$). If instead y is transient then, on average, the chain will only return to y a finite number of times.

Theorem 3.13. *A state y is recurrent if and only if $\mathbb{E}_y(N(y)) = \infty$ and it is transient if and only if $\mathbb{E}_y(N(y)) < \infty$.*

Proof. Recalling the definition of $N(y)$, Definition 3.11, we have

$$\begin{aligned}\mathbb{E}_y(N(y)) &= \sum_{k=0}^{\infty} \mathbb{P}_y(N(y) \geq k) = \sum_{k=0}^{\infty} \mathbb{P}_y(T_y^k < \infty) \\ &\overset{(17)}{=} \sum_{k=0}^{\infty} \rho_{yy}^k = \begin{cases} \infty & \text{iff } \rho_{yy} = 1 \\ \frac{1}{1-\rho_{yy}} & \text{iff } \rho_{yy} < 1. \end{cases}\end{aligned} \tag{18}$$

□

Theorem 3.14. *If x is recurrent and communicates with y, i.e. $\rho_{xy} > 0$, then y is recurrent and $\rho_{yx} = 1$.*

Proof. Let us start with proving that if x is recurrent and $\rho_{xy} > 0$ then $\rho_{yx} = 1$. Recall that x recurrent means that $\mathbb{P}_x(T_x < \infty) = 1$ i.e. $\mathbb{P}_x(T_x = \infty) = 0$. We will show that if $\rho_{xy} > 0$ and $\rho_{yx} < 1$ then x cannot be recurrent. If $\rho_{xy} > 0$ then there exists $h > 0$ such that $p^h(x, y) > 0$. Let $\bar{h}$

be the smallest h for which $p^h(x,y) > 0$. Then there exist $z_1, \ldots, z_{\bar{h}-1} \in S$ such that

$$p(x,z_1)p(z_1,z_2)\cdot\ldots\cdot p(z_{\bar{h}-1},y) > 0$$

and $z_j \neq x$ for all j (otherwise $\bar{h}$ wouldn't be the smallest h for which $p^h(x,y) > 0$). With this in mind, if $\rho_{yx} < 1$ we have

$$\begin{aligned}\mathbb{P}_x(T_x = \infty) &= p(x,z_1)\cdot\ldots\cdot p(z_{\bar{h}-1},y)\mathbb{P}_y(T_x = \infty)\\ &= p(x,z_1)\cdot\ldots\cdot p(z_{\bar{h}-1},y)(1-\rho_{yx}) > 0,\end{aligned}$$

so it has to be $\rho_{yx} = 1$. Now let us prove that y is recurrent. To do so, we will show that $\mathbb{E}_y(N(y)) = \infty$. Because $\rho_{yx} > 0$, there exists $\ell > 0$ s.t. $p^\ell(y,x) > 0$ and also recall that $p^{\bar{h}}(x,y) > 0$. From the Chapman-Kolmogorov equation we have that

$$p^{\ell+n+\bar{h}}(y,y) \geq p^\ell(y,x)p^n(x,x)p^{\bar{h}}(x,y),$$

so that summing over n on both sides we get $\mathbb{E}_y(N(y)) = \infty$ as $\mathbb{E}_x(N(x)) = \sum_{n\geq 1} p^n(x,x) = \infty$ because x is recurrent. □

Now we need another couple of definitions.

Definition 3.15. *A set $C \subset S$ is* closed *if*

$$x \in C \text{ and } \rho_{xy} > 0 \;\Rightarrow\; y \in C.$$

A set $F \subset S$ is irreducible *if*

$$x, y \in F \;\Rightarrow\; \rho_{xy} > 0.$$

A chain is irreducible if the whole state space is irreducible.

These two definitions might look similar but they are actually quite different. In the case of a closed set, if we start in C then we remain in C i.e. if $x \in C$ then $\mathbb{P}_x(X_n \in C) = 1$ for all $n \geq 1$. Also, strictly speaking, if we have a closed set and we consider the set $C' = C \cup \{z\}$, where $p(c,z) = 0$ for all $c \in C$, then C' is still closed. For an irreducible set this cannot happen as every element has to communicate with each other. Moreover, in an irreducible set, if $\rho_{xy} > 0$ then also $\rho_{yx} > 0$ and this is not true for a closed set.

With the above definition, the following corollary is a straightforward consequence of Theorem 3.14.

Corollary 3.16. *If a chain is irreducible then either all the states are recurrent or all the states are transient.*

What the following Theorem 3.17 and Corollary 3.18 prove is that in the case of a chain with finite state space, there is a way to decide which of the two cases occurs: if S is finite, closed and irreducible then all the states are recurrent.

Theorem 3.17. *If C is a finite closed set then C contains at least one recurrent state.*

Corollary 3.18. *If C is a finite set which is closed and irreducible then every state in C is recurrent.*

Proof of Theorem 3.17. By contradiction, suppose all the states in C are transient. If this is the case then, acting as in (18) we have

$$\mathbb{E}_x(N(y)) = \rho_{xy}/(1-\rho_{yy}) < \infty. \tag{19}$$

So if we pick $x \in C$, we have

$$\infty > \sum_{y\in C} \mathbb{E}_x(N(y)) = \sum_{n=1}^{\infty} \sum_{y\in C} p^n(x,y) = \infty,$$

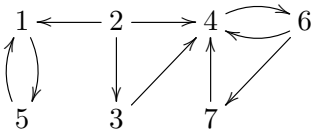
where the first inequality on the left follows from the finiteness of C and the last equality follows from the fact that $\sum_{y\in C} p^n(x,y) = 1$ as C is closed. □

Sometimes (and when I say sometimes I mean when the state space is finite and with reasonable cardinality), it can be useful to "draw" a Markov chain, which can be done as follows.

Example 3.19. When the state space is finite the transition probabilities can be organized in a *transition matrix*, $P = (p(x,y))_{x,y\in S}$. Consider the transition matrix

	1	**2**	**3**	**4**	**5**	**6**	**7**
1	0.3	0	0	0	0.7	0	0
2	0.1	0.2	0.3	0.4	0	0	0
3	0	0	0.5	0.5	0	0	0
4	0	0	0	0.5	0	0.5	0
5	0.6	0	0	0	0.4	0	0
6	0	0	0	0	0	0.2	0.8
7	0	0	0	1	0	0	0

We draw a graph without indicating the self-loops:



The sets $\{1, 5\}$ and $\{4, 6, 7\}$ are irreducible closed sets so every state in these sets is recurrent.

You might think that the fact that all the recurrent states are contained in one of the closed and irreducible sets of the chain is specific to the above example. The following theorem shows that this is not the case.

Theorem 3.20 (Decomposition Theorem)**.** *Let $R := \{r \in S : r \text{ is recurrent}\}$ be the set of all recurrent states of the chain. Then R can be written as the disjoint union of closed and irreducible sets, i.e.*

$$R = \cup_i C_i, \quad C_i \text{ closed and irreducible for any } i.$$

Proof. (Almost all of it). Let x be a recurrent state and define $C_x := \{y \in S : \rho_{xy} > 0\}$. From Theorem 3.14 we know that $C_x \subset R$ and that $\rho_{yx} > 0$. Therefore either $C_x \cap C_y = \emptyset$ or $C_x = C_y$. □

Now we want to introduce a very important notion in the theory of stochastic processes, the notion of *stationary measure*. This will lead us to the definition of *ergodicity*. Technicalities are of fundamental importance but sometimes, if not driven by the right intuition, they can obfuscate the idea underlying them. So let us first explain in words what we are trying to do, starting with a very simple example. Consider a pendulum, oscillating around the vertical position. The only (stable) stationary state for this system is the vertical position. Such a state is stationary in the sense that if we start the motion at that position, the system is going to remain in that state. In other words, such a state is an equilibrium for the system. If we start the motion out of equilibrium, in the long run we will end up at equilibrium again. However, the one that we have described is a deterministic system, whereas we deal with stochastic systems so we don't look at the deterministic evolution but rather at the evolution of probability measures, as the state of the system at time n is described by a probability measure (think about the Ehrenfest chain for example, where the state of the system is described once we exhibit the probability to have k molecules in the left chamber at time n, for all k and n). Therefore in our context we don't talk

about stationary states but we rather consider stationary measures and we say that a measure μ is stationary for the process X_n if $X_0 \sim \mu \Rightarrow X_n \sim \mu$ for all $n \geq 0$. As in the case of the pendulum, stationary measures are potential candidates to be the equilibrium state of the chain and therefore they describe the long time behaviour of the process. Obviously there can be more than one stationary measure. This leads us to the concept of ergodicity, which is a property regarding the long time behaviour of the process: a process is ergodic if it admits a unique stationary measure. We will come back to this definition later and we will state it in more detail but for the moment, roughly speaking, in order for a process to be ergodic, it has to (in the long run)

- explore the whole state space
- explore it in an "homogeneous way", i.e. we want the time spent in a given area of the state space to be proportional to how "big" that area is. Indeed the physicists' definition of ergodicity is "space averages equal time averages"
- we want all the above to happen independent of the initial condition.

Loosely speaking, if the process is ergodic, it will converge to the stationary measure. Let us now start with the proper maths, hoping to give more intuition along the way. We recall that in all that follows we are still referring to time-homogeneous Markov chains with discrete state space.

Definition 3.21 (Stationary measure)**.** *A measure μ on S is* stationary *for the Markov chain X_n with transition matrix p if*

$$\sum_x \mu(x)p(x,y) = \mu(y). \tag{20}$$

If μ is a probability measure then we call it a stationary distribution*.*

A short notation for (20) is $\mu\, p = \mu$. If μ is the initial distribution of the chain then the LHS of (20) is $\mathbb{P}_\mu(X_1 = y)$. So, if $X_0 \sim \mu$ then also $X_1 \sim \mu$ and $X_n \sim \mu$ for all $n \geq 1$ (check) and this is the reason why these measures are also called *invariant*. In particular check that if (20) holds then

$$\sum_x \mu(x)p^n(x,y) = \mu(y) \quad \text{for all } \ n \geq 1.$$

Example 3.22. *Consider the simple random walk from the exercise sheet. Then $\mu(k) \equiv 1$ is a stationary measure for the chain, as*

$$\sum_k \mu(k)p(k,j) = \mu(j-1)p(j-1,j)+\mu(j+1)p(j+1,j) = 1-p+p = 1 = \mu(j).$$

Definition 3.23 (Reversibility)**.** *If a measure μ satisfies*

$$\mu(x)p(x,y) = \mu(y)p(y,x) \tag{21}$$

then μ is a reversible measure*.*

The equality (21) is also called *detailed balance condition* (DB) and it is of fundamental importance in the context of Metropolis-Hastings algorithms.

Theorem 3.24. *If a measure μ satisfies the DB condition* (21) *then it is stationary.*

Proof. Just sum over x on both sides of (21) and get

$$\sum_{x\in S} \mu(x)p(x,y) = \mu(y)\sum_{x\in S} p(y,x) = \mu(y),$$

where the last equality follows from the fact that p is a stochastic matrix. □

Remark 3.25. Suppose that the Markov chain X_n with transition matrix p admits a stationary measure μ and that $X_0 \sim \mu$, i.e. the chain is started in stationarity. For every fixed $n \in \mathbb{N}$ we can consider the process $\{Y_m\}_{0\le m\le n}$ defined as $Y_m^n := X_{n-m}$, i.e. Y_m^n is the "time-reversed" X_n. Then for every $n \in \mathbb{N}$, Y_m^n is a time-homogeneous Markov chain with $Y_0 \sim \mu$. To calculate the transition probabilities $q(x,y)$ of Y_m^n we use Bayes' formula:

$$\begin{aligned} q(x,y) &= \mathbb{P}(Y_1 = y|Y_0 = x) = \mathbb{P}(X_{n-1} = y|X_n = x) \\ &= \mathbb{P}(X_n = x|X_{n-1} = y)\frac{\mu(y)}{\mu(x)} = \frac{p(y,x)\mu(y)}{\mu(x)}. \end{aligned}$$

If the DB condition holds, then $q(x,y) = p(x,y)$ for all x, y and in this case X_n is called *time-reversible*.

Now a very important definition, which holds for chains as well as for continuous time processes.

Definition 3.26. *A Markov chain is said to be* ergodic *if it admits a unique stationary probability distribution. In this case, the invariant distribution is said to be the ergodic measure for the chain.*

I would like to stress that, depending on the book you open, you might find slightly different definitions of ergodicity. However they all aim at describing the same intuitive idea. The next theorem says that as soon as we have a recurrent state, we can construct a stationary measure $\mu(y)$ looking at the expected number of visits to y, i.e. the expected time spent by the chain in y.

Theorem 3.27. *Recall that $T_x := \inf\{n \geq 1 : X_n = x\}$. If x is a recurrent state then the measure*

$$\mu_x(y) := \mathbb{E}_x \left[\sum_{n=0}^{T_x - 1} \mathbf{1}_{(X_n = y)} \right]$$

is stationary.

Proof of Theorem 3.27. We need to prove that

$$\sum_{y \in S} \mu_x(y) p(y, z) = \mu_x(z).$$

Let us study two cases separately, the case $x \neq z$ and the case $x = z$.

- Case $x \neq z$: To this let us rewrite $\mu_x(y) = \sum_{n=0}^{\infty} \mathbb{P}_x(X_n = y, T_x > n)$, so we have:

$$\begin{aligned} \sum_{y \in S} \mu_x(y) p(y, z) &= \sum_{n=0}^{\infty} \sum_{y \in S} \mathbb{P}_x(X_n = y, T_x > n) p(y, z) \\ &= \sum_{n=0}^{\infty} \mathbb{P}_x(X_{n+1} = z, T_x > n) \\ &= \sum_{n=0}^{\infty} \mathbb{P}_x(X_{n+1} = z, T_x > n + 1). \end{aligned}$$

  On the other hand,

$$\begin{aligned} \mu_x(z) &= \sum_{n=0}^{\infty} \mathbb{P}_x(X_n = z, T_x > n) \\ &= \sum_{n=1}^{\infty} \mathbb{P}_x(X_n = z, T_x > n) \\ &= \sum_{n=0}^{\infty} \mathbb{P}_x(X_{n+1} = z, T_x > n + 1), \end{aligned}$$

  proving the claim if $x \neq z$.

- Case $x = z$ then, since $\mu_x(x) = 1$ by definition, we need to show that

$$\begin{aligned}\sum_{y\in S}\mu_x(y)p(y,x) &= \sum_{n=0}^{\infty}\sum_{y\in S}\mathbb{P}_x(X_n = y, T_x > n)p(y,x)\\ &= \sum_{n=0}^{\infty}\mathbb{P}_x(T_x = n+1) = 1,\end{aligned}$$

  where in the last equality we used the fact that x is recurrent.

□

Theorem 3.28. *If the chain is irreducible and recurrent then there exists a unique stationary measure, up to constant multiples.*

Idea of Proof. Let R be the set defined in Theorem 3.20. The moral is the following: if we pick $x \in C_i$ then the measure μ_x defined in Theorem 3.27 is a stationary measure. If we pick any other $\bar{x}$ in the same set C_i, the corresponding measure $\mu_{\bar{x}}$ is only a constant multiple of μ_x. Under our assumptions, there is only one big recurrent irreducible set, the state space S. Therefore we can pick any $x \in S$ and construct an invariant measure μ_x. At this point one proves that any other invariant measure say μ will be a constant multiple of μ_x, i.e. there exists a constant $K > 0$ such that

$$\mu_x(y) = K\mu(y) \quad \text{for all } y \in S.$$

□

So far we have a way to establish existence of the stationary measure, Theorem 3.27, and a result to establish uniqueness, Theorem 3.28. But what we are looking for is a distribution.

Theorem 3.29. *Suppose the chain is irreducible. Then the following statements are equivalent:*

1. *the chain is positive recurrent (i.e. all the states are positive recurrent)*
2. *at least one state is positive recurrent*
3. *the chain admits a unique stationary distribution.*

In order to prove the above result we need the following two technical lemmata.

Lemma 3.30. *Suppose π is a stationary distribution for the chain. Then*

$$\pi(x) > 0 \;\Rightarrow\; x \text{ is recurrent.}$$

Proof of Lemma 3.30. Exercise.

□

Lemma 3.31. *Suppose the chain has a stationary distribution π. If the chain is also irreducible then we can express π as*

$$\pi(y) = 1/\mathbb{E}_y T_y\,.$$

Proof of Lemma 3.31. Irreducibility $\Rightarrow \pi(z) > 0$ for all $z \in S \Rightarrow$ the chain is recurrent. Irreducible recurrent chains have a unique stationary measure up to constant multiples. and we know that objects of the form $\mu_x(y) = \sum_{n=0}^{T_x-1} p^n(x,y)$ are stationary measures. The normalization factor for μ_x is

$$\begin{aligned}\sum_y \mu_x(y) &= \sum_{n=0}^{T_x-1}\sum_y p^n(x,y) = \sum_{n=0}^{\infty}\sum_y \mathbb{P}_x(X_n = y, T_x > n)\\ &= \sum_{n=0}^{\infty}\mathbb{P}_x(T_x > n) = \mathbb{E}_x T_x.\end{aligned}$$

Now it is clear that our stationary distribution is

$$\pi(z) = \frac{\mu_x(z)}{\sum_z \mu_x(z)} = \frac{C\mu_z(z)}{C\sum_y \mu_z(y)} = \frac{1}{\mathbb{E}_z T_z},$$

where in the above C is a constant and we used $\mu_z(z) = 1$.

□

Proof of Theorem 3.29. 1 $\Rightarrow$ 2 is obvious.
2 $\Rightarrow$ 3. Suppose x is positive recurrent, so $\mathbb{E}_x T_x < \infty$. Then from the proof of Lemma 3.31 and from Theorem 3.27 we know that

$$\pi(y) := \mu_x(y)/\mathbb{E}_x T_x\,.$$

3 $\Rightarrow$ 1. If the chain is irreducible then $\pi(x) > 0$ for all x and the stationary distribution can be expressed as

$$\pi(x) = \frac{1}{\mathbb{E}_x T_x} > 0,$$

hence $\mathbb{E}_x T_x < \infty$ for all x, i.e. the chain is positive recurrent. □

This completes the picture. Roughly speaking we have shown that

- recurrence $\rightsquigarrow$ existence of invariant measure
- irreducibility $\rightsquigarrow$ uniqueness of invariant measure
- positive recurrence $\rightsquigarrow$ finiteness of the stationary measure, i.e. we have a stationary distribution.

Now that we have a unique candidate for the equilibrium state of the chain, i.e. for the asymptotic behaviour of the process, we need to answer the last question: when does the chain converge to equilibrium? In other words, we want to understand the behaviour of $p^n(x, y)$ for large n. Before answering the question we need another definition.

Definition 3.32. *Let ν_x be the largest common divisor of the integers in the set $\{n \geq 1 : p^n(x, x) > 0\}$. ν_x is the* period *of x. If $\nu_x = 1$ then x is* aperiodic*. A chain is aperiodic if all the states are aperiodic.*

As you can imagine,

Lemma 3.33. *If $\rho_{xy} > 0$ then $\nu_x = \nu_y$.*

Now let us answer the question.

Theorem 3.34. *Suppose the chain is irreducible and positive recurrent. Then we know there exists a unique stationary distribution π. If the chain is also aperiodic then $p^n(x, y) \longrightarrow \pi(y)$, and the convergence is in total variation i.e.*

$$\sum_{y \in S} |p^n(x, y) - \pi(y)| \to 0. \tag{22}$$

Comment. It is clear now that $\pi(y)$ represents the probability for the chain to be in y as $n \to \infty$. It is very important that convergence to π happens irrespective of the initial datum that we pick, i.e. in the limit the initial condition is forgotten. We will not prove the above theorem as we think that the meaning of the statement is already transparent enough: if there exists a unique stationary distribution, which follows from irreducibility and positive recurrence, the only thing that can prevent the process from converging to its unique candidate limit is periodicity. Once we rule that possibility out, the asymptotic behaviour of the chain can only be described by π. Another crucial observation to bear in mind is that (22) implies $p^n(x, y) \overset{n \to \infty}{\longrightarrow} \pi(y)$ for all y. Because $\pi(y) = 1/\mathbb{E}_y T_y$, this result is quite intuitive: the probability of being in y is asymptotically inversely proportional to the expected value

of the time of first return to y. And this brings us to the last result, which is very much along the lines of "time averages equal space averages".

Define $N_n(y)$ to be the number of visits to y in the first n steps, i.e.

$$N_n(y) = \sum_{j=1}^{n} \mathbf{1}_{(X_j=y)}.$$

Theorem 3.35. *Suppose y is a recurrent state of the chain. Then for every initial state of the chain $x \in S$, we have*

$$\frac{N_n(y)}{n} \longrightarrow \frac{1}{\mathbb{E}_y T_y} \mathbf{1}_{(T_y<\infty)}, \quad \mathbb{P}_x \text{ a.s.}$$

Remark 3.36 (On the nomenclature)**.** In all of the above we have referred all our definitions to the chain X_n rather than to its transition probabilities, i.e. for example we said that the cain is recurrent or irreducible etc. However, one can equivalently refer all these definitions to the transition probabilities p and, in order to say e.g. that the chain is aperiodic we can equivalently say that p is aperiodic.

Example 3.37. This example shows why aperiodicity is a much needed assumption in Theorem 3.34. Consider the Markov Chain on state space $S = \{a, b\}$ with transition matrix

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

This chain is clearly positive recurrent and irreducible, hence ergodic with invariant measure is $\pi(a) = \pi(b) = 1/2$. However it is also periodic with period 2. It is easy to check that the result of Theorem 3.34 does not hold for this chain, indeed $p^n(a, b)$ is one if n is odd and 0 if n is even (and analogously for $p^n(b, a)$). Therefore p^n doesn't converge at all.

4 Markov Chain Monte Carlo Methods

For the material of this section we refer to [74, 67, 2, 52, 30, 10].

In the previous section we have studied the basic theory of Markov chains on finite or countable state space. Most of this theory can be translated to general state space, but this is not what we want to do now. In this section we want to look at one of the main applications of the theory that we have seen so far: Markov Chain Monte Carlo methods (MCMC). The main idea

and purpose of MCMC is readily explained: suppose we want to sample from a given probability distribution $\pi(x)$ on a state space S. We know from Theorem 3.34 that if X_n is an ergodic (and aperiodic) Markov chain with invariant distribution π then "for n large enough" $p^n(x,y) \sim \pi(y)$ i.e. the outcomes of the chain are distributed according to π. Therefore one way of sampling from π is constructing a MC which has π as limiting distribution. If we run the simulation of the chain "long enough", the elements X_n will be the desired samples from π[7]. For simulation purposes, a very important question is "how big" n needs to be; and, on top of that, how long it takes to the chain, once stationarity is reached, to explore the state space. We shall not address these points here but bear in mind that this is a very important factor to optimize.

An easier reach question at this point is...why do we need to sample from probability distributions? Well, the main uses of Monte Carlo[8] and Markov Chain Monte Carlo are in integration and optimization problems:

i) To calculate multidimensional integrals: we shall see in Section 5 that, roughly speaking, if a chain is ergodic and stationary with invariant probability π then

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} f(X_j) = \mathbb{E}_\pi(f) = \int f(x)\pi(x)dx, \text{[9]} \tag{23}$$

for every π- integrable function f (i.e. for every f such that $\int |f(x)|\ \pi(x)$ is finite). We will defer to Section 5 a more thorough discussion of the limit (23), usually known as Ergodic Theorem. For the time being the important things to notice are: (i) the limit (23) is a strong law of large numbers - see Theorem 1.10; (ii) if we take f to be the indicator function of a measurable set then (23) says precisely that, in the limit, "time averages equal space averages"; most importantly to our purposes (iii) for n large enough, the quantity on the LHS is a good approximation of the integral on the right- see Exercise 2.

[7] However MCMC should only be used when other direct analytical methods are not applicable to the situation at hand.

[8] On an hystorical note, Monte Carlo (or also Ordinary Monte Carlo (OMC)) came before MCMC. OMC was pretty much a matter of statistics: if you could simulate from a given distribution, then you could simulate a sequence of i.i.d. random variables with that distribution and then use the Law of Large numbers to approximate integrals. When the method was introduced, the sequence of i.i.d. was not necessarily created as a stationary Markov chain.

[9] Notice that as usual we are assuming that π has a density, which we keep denoting by π.

ii) To calculate the minima (or maxima) of functions, see Section 4.1.

The following example contains the main features of the sampling methods that we will discuss.

Example 4.1 (Landscape painting attempt)**.** *** Let $q(x, y)$ be a transition probability on a finite state space S . Suppose the transition matrix $Q = (q(x, y))$ is symmetric and irreducible. Given such a Q and a probability distribution $\pi(x)$ on S such that $\pi(x) > 0$ for all $x \in S$, let us now construct a new transition matrix $P = (p(x, y))$ as follows :

$$p(x,y) = \begin{cases} q(x,y) & \text{if } \pi(y) \geq \pi(x) \text{ and } x \neq y \\ q(x,y)\frac{\pi(y)}{\pi(x)} & \text{if } \pi(y) < \pi(x) \text{ and } x \neq y \\ 1 - \sum_{x \neq y} p(x,y) & \text{otherwise.} \end{cases} \tag{24}$$

First of all $P = (p(x, y))$ is a transition matrix, indeed from the definition $\sum_y p(x, y) = p(x, x) + \sum_{y \neq x} p(x, y) = 1$. Moreover, P is irreducible[10] by construction because Q is; being the state space finite, this also implies that P is recurrent and that there exists a unique stationary distribution. We can easily show that such an invariant distribution is exactly π as P is reversible with respect to π. To prove π- reversibility of P we need to show that $\pi(x)p(x, y) = p(y, x)\pi(y)$. This is obviously true when $x = y$. So suppose $x \neq y$:

i) if $\pi(y) \geq \pi(x)$: $\pi(x)p(x, y) = \pi(x)q(x, y)$ but also $\pi(y)p(y, x) = q(y, x)\frac{\pi(x)}{\pi(y)}\pi(y)$ so that using the symmetry of q we get $\pi(y)p(y, x) = q(x, y)\pi(x)$ and we are done.

ii) if $\pi(y) < \pi(x)$: clearly same as above with roles of x and y reversed.

We are in good shape but we haven't yet proved convergence to π. This is left as an exercise, see Exercise 17. It turns out that convergence happens unless π is the uniform distribution on S.
Let us pause everything for a moment to think about what we have done: starting from a (pretty much) arbitrary transition kernel q, we have constructed a chain with transition kernel p that has π as limiting distribution. We will see that the "(pretty much)" can be (pretty much) removed. Now there is only one point left to address: how do we sample from the chain that has P as transition matrix? By using the following algorithm

[10] I.e. the whole state space is irreducible under P; this implies that the state space is also closed under P.

Algorithm 4.2. Given $X_n = x_n$,

1. generate $y_{n+1} \sim q(x_n, \cdot)$;
2. if $\pi(y_{n+1}) \geq \pi(x_n)$ then $X_{n+1} = y_{n+1}$;
   if $\pi(y_{n+1}) < \pi(x_n)$ then $X_{n+1} = y_{n+1}$ with probability $\pi(y_{n+1})/\pi(x_n)$
   otherwise $X_{n+1} = X_n$ with probability $1 - \pi(y_{n+1})/\pi(x_n)$.

In words, given the state of the chain at time n, we pick the *proposal* $y_{n+1} \sim q(x_n, \cdot)$. Then the proposed move is accepted with probability $\alpha(x_n, y_{n+1}) := \min\{1, \pi(y_{n+1})/\pi(x_n)\}$. If it is rejected, the chain remains where it was. $\alpha(x, y)$ is called the *acceptance probability*.

Algorithm 4.2 is a first example of a *Metropolis-Hastings algorithm*. A naïve explanation of the reason why we always accept moves towards points with higher probability comes from the Ergodic Theorem, limit (23), reread in the case in which the state space is finite (so that the integral on the RHS is just a sum). If we want to construct an ergodic chain with invariant probability π then the time spent by the chain in each point y of S equals, in the long run, the probability assigned by π to y, i.e. $\pi(y)$.

Remark 4.3. Notice that by using the acceptance probability, the kernel (24) can be rewritten in a slightly more compact form:

$$p(x, y) = q(x, y)\alpha(x, y) + \delta_x(y) \sum_{w \in S} (1 - \alpha(x, w))q(x, w),$$

where $\delta_x(y)$ is the Kronecker delta. In view of Section 4.3 it is also useful to note that Algorithm 4.2 can be equivalently expressed as follows: given $X_n = x_n$,

1. generate $y_{n+1} \sim q(x_n, \cdot)$;
2. set $X_{n+1} = \begin{cases} y_{n+1} & \text{with probability } \alpha(x_n, y_{n+1}) \\ x_n & \text{otherwise.} \end{cases}$

4.1 Simulated Annealing

We now address the point **ii)** listed at the beginning of this section.

Suppose again we are in the case in which the state space S is finite and again in the setting of Example 4.1. Let us now apply the reasoning (24) to the case when the target distribution is

$$\pi_\epsilon(x) = \frac{e^{-H(x)/\epsilon}}{\mathcal{Z}_\epsilon}, \tag{25}$$

where $\epsilon > 0$ is a positive parameter, $H(x)$ is a function on S and $\mathcal{Z}_\epsilon$ is a normalization constant, i.e.

$$\mathcal{Z}_\epsilon = \sum_{x \in S} e^{-H(x)/\epsilon}, \quad \text{so that } \sum_{x \in S} \pi_\epsilon(x) = 1.$$

If $Q = (q(x,y))$ is a symmetric and irreducible transition matrix, the prescription (24) now becomes

$$p_\epsilon(x,y) = \begin{cases} q(x,y) & \text{if } H(y) \leq H(x) \text{ and } x \neq y \\ q(x,y) e^{(H(x)-H(y))/\epsilon} & \text{if } H(y) > H(x) \text{ and } x \neq y \\ 1 - \sum_{x \neq y} p_\epsilon(x,y) & \text{otherwise.} \end{cases}$$

Notice that to simulate the Markov Chain with transition probability $p_\epsilon(x,y)$ we don't need to know a priori the value of the normalization constant $\mathcal{Z}_\epsilon$. We now know that for n large enough, $p_\epsilon^n \sim \pi_\epsilon$.

Now observe that the definition of π_ϵ remains unaltered if the function H is modified through an additive constant c, i.e. if instead of H we consider $\tilde{H}(x) = H(x) + c$ for some constant c then

$$\pi_\epsilon(x) = \frac{e^{-c/\epsilon} e^{-H(x)/\epsilon}}{e^{-c/\epsilon} \sum_{x \in S} e^{-H(x)/\epsilon}} = \frac{e^{-H(x)/\epsilon}}{\mathcal{Z}_\epsilon}.$$

Therefore we can assume without loss of generality that the minimum of H is zero. We are after the K (possibly $K > 1$) points of S, $x_1, \ldots, x_K$, such that $H(x_j) = 0$. In order to find such a set of points where the minimum is reached we observe that as ϵ goes to zero π_ϵ tends to the uniform distribution on $x_1, \ldots, x_K$, indeed

$$e^{-H(x)/\epsilon} \xrightarrow{\epsilon \to 0} \begin{cases} 1 & \text{if } x = x_j \text{ for some } j = 1, \ldots, K \\ 0 & \text{otherwise,} \end{cases}$$

hence $\mathcal{Z}_\epsilon \xrightarrow{\epsilon \to 0} K$ and

$$\pi_\epsilon(x) \xrightarrow{\epsilon \to 0} \begin{cases} 1/K & \text{if } x = x_j \text{ for some } j = 1, \ldots, K \\ 0 & \text{otherwise.} \end{cases}$$

In this way, if ϵ is small enough, for large n the chain will be in one of the minima with very high probability.

Instead of fixing a small ϵ, we can decrease ϵ at each step. This way we obtain a non-homogeneous Markov Chain, which we haven't discussed. However, when ϵ is interpreted as the temperature of the system, such a

procedure of decreasing ϵ is what gives the name "simulated annealing" to the technique that we have just presented. Such a name is borrowed from the field of metallurgy as it comes from the heat treatment that some metals are subject to, when they are left to cool off in a controlled way.

In the context of equilibrium statistical mechanics a measure of the form (25) is commonly called *Gibbs measure* at inverse temperature β, where $\beta = 1/\epsilon$. Such measures are of great importance as they represent the equilibrium measure for a Hamiltonian system with Hamiltonian H. Therefore the simulated annealing optimization method can be used to locate the global minima of the energy of the system.

4.2 Accept-Reject method

Example 4.1 and Algorithm 4.2 give an idea of what we are aiming for. However let us take a step back and start again where we began. The aim of the game is sampling from a given probability distribution. So suppose we want to produce sample outcomes of a real valued random variable X with density function $\pi(x)$[11]. The idea is to use an auxiliary density function $\nu(x)$ which we know how to sample from.

The method is as follows: we start with producing two samples, independent of each other, $Y \sim \nu$ and $U \sim \mathcal{U}_{[0,1]}$, where $\mathcal{U}_{[0,1]}$ denotes the uniform distribution on $[0,1]$. If $U \leq \frac{\pi(Y)}{M\,\nu(Y)}$ then we accept the sample and set $X = Y$ otherwise we start again. The algorithm is as follows

Algorithm 4.4 (Accept-Reject algorithm)**.**

1. Generate $Y \sim \nu$ and, independently, $U \sim \mathcal{U}_{[0,1]}$;
2. if $U \leq \frac{\pi(Y)}{M\,\nu(Y)}$ then set $X = Y$ otherwise go back to step one.

It is clear from the above that there are only two constraints on $\pi(x)$ and $\nu(x)$ in order for this method to be applicable :

- the auxiliary density ν is such that $\nu(x) > 0$ if $\pi(x) > 0$;
- there exists a constant $M > 0$ such that $\pi(x)/\nu(x) \leq M$ for all x.

Now we need to check that the samples generated this way are actually distributed according to π. Because

$$\mathbb{P}(X \leq a) = \mathbb{P}\left(Y \leq a \,\middle|\, U \leq \frac{\pi(Y)}{M\,\nu(Y)}\right) \quad \text{for all } a \in \mathbb{R},$$

[11] This sampling method works in higher dimensions as well, here we present it in one dimension just for ease of notation.

what we need to show is that

$$\mathbb{P}\left(Y \leq a \,\middle|\, U \leq \frac{\pi(Y)}{M\,\nu(Y)}\right) = \frac{\int_{-\infty}^{a} dy\,\pi(y)}{\int_{-\infty}^{\infty} dy\pi(y)} \quad \text{for all } a \in \mathbb{R}.$$

This is a simple calculation, once we recall that for any two real valued random variables, W and Z, with joint probability density $f_{W,Z}(w,z)$, we have

$$\mathbb{P}(W \in A | Z \in B) = \frac{\int_B \int_A f_{W,Z}(w,z)dw\,dz}{\int_B \int_{\mathbb{R}} f_{W,Z}(w,z)dw\,dz}.$$

In our case it is clearly $f_{Y,U} = \nu(y)\cdot 1$, as Y and U are independent, so using the above equality we obtain

$$\begin{aligned}\mathbb{P}\left(Y \leq a \,\middle|\, U \leq \frac{\pi(Y)}{M\,\nu(Y)}\right) &= \frac{\int_{-\infty}^{a} dy \int_0^{\pi(y)/M\nu(y)} \nu(y)\,du}{\int_{-\infty}^{\infty} dy \int_0^{\pi(y)/M\nu(y)} \nu(y)\,du} \\ &= \frac{\int_{-\infty}^{a} dy\,\pi(y)}{\int_{-\infty}^{\infty} dy\pi(y)} = \mathbb{P}(X \leq a).\end{aligned}$$

Remark 4.5. Notice that to use the accept-reject method we don't need to know the normalization constant for the density function π i.e. we don't need to know $\mathcal{Z} = \int_{-\infty}^{\infty} \pi(x)dx$. However we need to know the constant M. Indeed, even if M cancels in the calculation above – so that it can be arbitrary – we still need to know its value in order to decide whether to accept or reject the sample $Y \sim \nu$ (see step 2 of Algorithm 4.4). Moreover, M is a measure of the efficiency of the algorithm. Indeed, the probability of acceptance at each iteration is exactly $1/M$ – see Exercise 19. Therefore in principle we would like to choose a distribution ν that makes M as small as possible.

One last, probably superfluous, observation: this method has nothing to do with Markov chains. The reason why we presented this algorithm should be clear in view of Example 4.1. The next section is actually about sampling methods that exploit Markov Chains.

4.3 Metropolis-Hastings algorithm

Before reading this section, read again Example 4.1. However this time our state space is $\mathbb{R}^N$. [12].

[12] This algorithm can be presented and studied on a general metric space, see [74]

A Metropolis-Hastings (M-H) algorithm is a method of constructing a time-homogeneous Markov chain or, equivalently, a transition kernel $p(x, y)$, that is reversible with respect to a given target distribution $\pi(x)$. To construct the π-invariant chain X_n we make use of a proposal kernel $q(x, y)$ which we know how to sample from and of an accept/reject mechanism with acceptance probability

$$\alpha(x, y) = \min\left\{1, \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}\right\}. \tag{26}$$

We require that $\pi(y)q(y, x) > 0$ and $\pi(x)q(x, y) > 0$. The M-H algorithm consists of two steps:

Algorithm 4.6 (Metropolis-Hastings algorithm)**.** Given $X_n = x_n$,

1. generate $y_{n+1} \sim q(x_n, \cdot)$;
2. set $X_{n+1} = \begin{cases} y_{n+1} & \text{with probability } \alpha(x_n, y_{n+1}) \\ x_n & \text{otherwise.} \end{cases}$

Lemma 4.7. *If α is the acceptance probability* (26)*, (and assuming $\pi(y)q(y, x) > 0$ and $\pi(x)q(x, y) > 0$) the Metropolis-Hastings algorithm, Algorithm 4.6, produces a π- invariant time-homogeneous Markov chain.* [13]

Proof. This is left as an easy exercise, see Exercise 18. □

Notice that the samples produced with this MCMC method are correlated, as opposed to those produced with the simple accept/reject method, which are i.i.d. The M-H samples are correlated for two reasons: because the proposed move y_{n+1} depends on x_n and because the acceptance probability depends on x_n.

Remark 4.8. In order to implement Algorithm 4.6 we don't need to know the normalizing constant for π, as it gets canceled in the ratio (26). However we do need to know the normalizing constant for q: q is a transition probability so by definition for every fixed x the function $y \to q(x, y)$ is a probability density i.e. it integrates to one. However the normalizing constant of $q(x, \cdot)$ can, and in general will, depend on x. In other words, $q(x, y)$

[13] On a technical note, the Lemma 4.7 can be made a bit more general (see [74]): first recall that we are assuming that $q(x, \cdot)$ and $\pi(\cdot)$ are probability densities with respect to the Lebesgue measure; having said that we can consider the set $R = \{x, y \in S : \pi(y)q(y, x) > 0 \text{ and } \pi(x)q(x, y) > 0\}$ and prove that the chain produced by Algorithm 4.6 satisfies the detailed balance condition if and only if the $\alpha(x, y) = 0$ on R^c.

will in general be of the form $q(x,y) = \mathcal{Z}_x^{-1}\tilde{q}(x,y)$, with $\int dy\, \tilde{q}(x,y) = \mathcal{Z}_x$ so that the ratio in the acceptance probability (26) can be more explicitly written as $\frac{\pi(y)\mathcal{Z}_x\tilde{q}(y,x)}{\pi(x)\mathcal{Z}_y\tilde{q}(x,y)}$.

Clearly the choice of the proposal q is crucial in order to improve the efficiency of the algorithm. We will make more remarks on this point later on. For the moment we want to explore some special cases of M-H, namely

1 when $q(x,y)$ depends on y only, i.e. $q(x,y) = q(y)$, in which case the corresponding M-H algorithm is called *independence sampler*;

2 when q is symmetric, i.e. $q(x,y) = q(y,x)$ and in particular when q is the transition density of a random walk, so that $q(x,y) = q(|x-y|)$ (see Example 3.8); in this case the resulting M-H algorithm is better known as *Random Walk Metropolis* (RWM);

Let us start with the independence sampler.

1 Independence Sampler. If the proposal kernel is independent of the current state of the chain, i.e. $q(x,y) = q(y)$ – with $q(y)$ a probability density – then the acceptance probability looks like

$$\alpha(x,y) = \min\left\{1, \frac{\pi(y)q(x)}{\pi(x)q(y)}\right\},$$

so in this case we don't need to know neither the normalization constant for the target π nor the one for the proposal q (see Remark 4.8). Algorithm 4.6 becomes

Algorithm 4.9 (Independence Sampler)**.** Given $X_n = x_n$,

1. generate $y_{n+1} \sim q(\cdot)$;
2. set $X_{n+1} = \begin{cases} y_{n+1} & \text{with probability } \alpha(x_n, y_{n+1}) \\ x_n & \text{otherwise.} \end{cases}$

If in general it makes sense to compare the M-H algorithm with the accept/reject method, it makes even more sense to compare the independence sampler with Algorithm 4.4, as they look very much alike. Let us spend a couple of words to stress the differences between these two: first of all bear in mind that the independence sampler produces a chain X_n, the outcomes of which are π-distributed only for large n, while with accept/reject each sample is π-distributed. In both cases the proposal y_{n+1} is generated independently of the value that had been produced at the previous iteration.

However the independence sampler still produces correlated samples – the acceptance probability still depends on x_n – as opposed to accept-reject, which produces i.i.d. outcomes (see also the comment before Remark 4.8). On the other hand, if we use accept/reject the upper bound on $\pi(x)/q(x)$ needs to be available a priori, whereas in the independence sampler the constant M does not need to be known.

2 The symmetric case and Random Walk Metropolis. If $q(x,y) = q(y,x)$ the acceptance probability reads

$$\alpha(x,y) = \min\left\{1, \frac{\pi(y)}{\pi(x)}\right\} \tag{27}$$

and therefore it doesn't depend on q at all. It is clear that in M-H, as much as in accept/reject, the choice of q, which can in principle be arbitrary, is of paramount importance in order to optimize the algorithm. The choice of the proposal q affects the efficiency of M-H at least on two levels as both the speed of convergence and the acceptance probability depend on q. In the symmetric case the acceptance probability (27) does not depend on the proposal kernel, however the *average acceptance rate* ρ does,

$$\rho = \iint \alpha(x,y)\pi(x)q(x,y)\,dx\,dy,$$

so it is wrong to conclude from (27) that in the symmetric case the only factor to optimize in the choice of the proposal is the speed of convergence.

A very popular M-H method is the so called *Random Walk Metropolis*, where the proposal y_{n+1} is of the form

$$y_{n+1} = x_n + \xi_{n+1},$$

where ξ_{n+1} is whatever noise independent of x_n. In RWM, the outcomes $\xi_1, \xi_2, \ldots, \xi_n, \ldots$ are i.i.d. with common density $g(x)$, which is assumed to be symmetric with respect to the origin, i.e. $g(x) = g(|x|)$. In this way $q(x,y) = g(y-x) = g(|y-x|)$ so $q(x,y) = q(|y-x|)$. For example, if $\xi \sim \mathcal{N}(0,\sigma^2)$ then $q(x,y) \sim \mathcal{N}(x,\sigma^2)$. The case in which the noise ξ is gaussian has been extensively studied in the literature., for target measures defined on $\mathbb{R}^N$. As you can imagine the efficiency of the algorithm decreases as the dimension N of the state space increases. This is a well known phenomenon commonly referred to as *curse of dimensionality*. Therefore, choosing the proposal variance becomes a more and more delicate matter as N grows. In $\mathbb{R}^N$ it is customary to consider $\sigma^2 = cN^{-\gamma}$ where $c, \gamma > 0$ are

two appropriate parameters, the most interesting of the two being γ. If γ is too large then σ^2 is too small, so the proposed moves tend to stay close to the current value of the chain and the state space is explored very slowly. If instead γ is too small, more precisely smaller than a critical value γ_c, it was shown in [3, 6, 7] that the average acceptance rate decreases very rapidly to zero as N tends to infinity. Such a critical value is, for RMW, equal to one. If we choose $\gamma = 1$ then the acceptance probability does not depend on N.

Remark 4.10. Let us repeat that M-H is a method to generate a π-reversible time-homogeneous Markov chain. As we have already noticed, the fact that the chain is π-reversible does not imply that π is the only invariant distribution for the chain or even less that the chain converges to π. In these lecture notes we shall not be concerned with the matter of convergence of the chain constructed via M-H, which is probably better studied case by case (see for example Exercise 17). However, for the sake of completeness, let us mention the two following results:

- If the target density $\pi(x)$ goes to zero exponentially fast as $|x| \to \infty$ then the M-H algorithm is ergodic and converges to π.
- If $\sup_y \pi(y)/q(y)$ is bounded, the independence sampler is ergodic and converges to the invariant distribution.

The precise statement and proof of these results can be found in [53, Chapter 20] and references therein.

4.4 The Gibbs Sampler

The Gibbs sampler is a method of sampling from a distribution which is the marginal of a joint probability. To be more clear, suppose we have two random variables X and Y with joint distribution $f_{X,Y}(x, y)$ and suppose we know that our target distribution π is precisely the marginal

$$\pi(x) = f_X(x) = \int f_{X,Y}(x, y)dy\,.$$

If we can sample from the conditional distributions $f_{X|Y}$ and $f_{Y|X}$, then the *two-stage Gibbs sampler* works as follows:

Algorithm 4.11 (Two-stage Gibbs sampler)**.** *Set* $Y_0 = y_0$.
Then, for $n = 0, 1, 2, \ldots$

1. $X_n \sim f(x|Y_n = y_n)$

2. $Y_{n+1} \sim f(y|X_n = x_n)$.

Yes I know at first sight it seems impossible that this algorithm does anything at all. Let us explain what it does and why it works. First of all, Algorithm 4.11 produces two Markov chains, X_n and Y_n. We are interested in sampling from $f_X(x)$ so the chain of interest to us is the chain X_n. One can prove that the chain X_n has f_X as unique invariant distribution and that, under appropriate assumptions on the conditional distributions $f_{X|Y}$ and $f_{Y|X}$, the chain converges to f_X. We shall illustrate this fact on a relatively simple but still meaningful example.

Example 4.12. Suppose X and Y are marginally Bernoulli random variables, so that we work in finite state space $S = \{0, 1\}$. Suppose the joint distribution of X and Y is assigned as follows:

$$\begin{bmatrix} f_{X,Y}(0,0) & f_{X,Y}(1,0) \\ f_{X,Y}(0,1) & f_{X,Y}(1,1) \end{bmatrix} = \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix}, \quad a_1 + a_2 + a_3 + a_4 = 1 .$$

The marginal distribution of X is then

$$f_X = [f_X(0)\ f_X(1)] = [a_1 + a_3\ a_2 + a_4] . \tag{28}$$

Consider the matrices containing the conditional distributions:

$$F_{Y|X} = \begin{bmatrix} \frac{a_1}{a_1+a_3} & \frac{a_3}{a_1+a_3} \\ \frac{a_2}{a_2+a_4} & \frac{a_4}{a_2+a_4} \end{bmatrix} \quad \text{and } F_{X|Y} = \begin{bmatrix} \frac{a_1}{a_1+a_2} & \frac{a_2}{a_1+a_2} \\ \frac{a_3}{a_4+a_3} & \frac{a_4}{a_4+a_3} \end{bmatrix},$$

i.e. $(F_{Y|X})_{ij} = \mathbb{P}(Y = j|X = i)$, $i, j \in \{0, 1\}$. Starting from the above matrices we can construct the matrix

$$F_{X|X} = F_{Y|X} F_{X|Y} .$$

Such a matrix is precisely the transition matrix for the Markov chain X_n. Indeed by construction, the sequence $X_0 \to Y_1 \to X_1$ is needed to construct the first step of the chain X_n, i.e. the step $X_0 \to X_1$. This means that the transition probabilities of X_n are precisely

$$\mathbb{P}(X_1 = x_1|X_0 = x_0) = \sum_y \mathbb{P}(X_1 = x_1|Y_1 = y)\mathbb{P}(Y_1 = y|X_0 = x_0) .$$

Therefore the entries of the matrix $(F_{X|X})^n$ are the transition probabilities $p^n(x_0, x) = \mathbb{P}(X_n = x|X_0 = x_0)$. Let the distribution of X_n be represented by the row vector

$$f^n = [f^n(0)\ f^n(1)] = [\mathbb{P}(X_n = 0)\ \mathbb{P}(X_n = 1)],$$

so that we have

$$f^n = f^0 \, F_{X|X}^n = f^{n-1} F_{X|X} \quad \forall n \geq 1 \,. \tag{29}$$

If all the entries of the matrix $F_{X|X}$ are strictly positive then the chain X_n is irreducible and, because the state space is finite, there exists a unique invariant distribution, which for the moment we call $\pi = [\pi(0) \; \pi(1)]$. However if all the entries of the transition matrix are strictly positive then the chain is also *regular* (see Exercise 16) hence f^n does converge to the distribution π as $n \to \infty$ [14], irrespective of the choice of the initial distribution for the chain X_n (which, in the case of the algorithm at hand, means irrespective of the choice of y_0). Taking the limit as $n \to \infty$ on both sides of (29) then gives

$$\pi = \pi F_{X|X} .$$

Because there is only one invariant distribution, the solution of the above equation is unique. It is easy to check that the marginal f_X defined in (28) satisfies the relation $f_X = f_X F_{X|X}$ (check it, by simply using the explicit expressions for $f_X, F_{Y|X}$ and $f_{X|Y}$) and hence f_X must be the invariant distribution.

On a very formal level, one could understand Gibbs sampling as a Metropolis-Hastings algorithm with acceptance probability constantly equal to 1.

A generalization of what we have done with two variables is the following. Suppose we have $m \geq 2$ variables $X_1, \ldots, X_m$ with joint distribution $f_{X_1,\ldots,X_m}(x_1, \ldots, x_m)$ and suppose we want to sample from the marginal

$$f_1(x_1) = \int f_{X_1,\ldots,X_m}(x_1, \ldots, x_m) dx_2 \ldots dx_m \,.$$

With the notation

$$f_{1|2,\ldots,m}(x_1|x_2, \ldots, x_m) = f_{X_1|X_2,\ldots,X_m}(x_1|x_2, \ldots, x_m) \,,$$

and similarly for $f_{j|1,\ldots,j-1,j+1,\ldots,m}$, the *multi step Gibbs sampler* is as follows.

Algorithm 4.13 (Multi stage Gibbs sampler)**.** *Set* $X_2 = x_2^{(0)}, \ldots, X_m = x_m^{(0)}$*. Then, for* $n = 0, 1, 2, \ldots$*, generate*

[14] From (22) we know that $p^n(x, y) \to \pi(y)$ for all $x \in S$. Because the chain X_n is started with $X_0 \sim f^0(x) = f_{X|Y}(x|Y = y_0)$, this implies that $\lim_{n\to\infty} f^n(x) = \sum_{z \in S} (\lim_{n\to\infty} p^n(z, x)) \, f_{X|Y}(z|Y = y_0) = \pi(x)$.

1. $X_1^{(n)} \sim f_{1|2,\ldots,m}(x_1|x_2^{(n)},\ldots,x_m^{(n)})$

2. $X_2^{(n+1)} \sim f_{2|1,3,\ldots,m}(x_2|x_1^{(n)},x_3^{(n)},\ldots,x_m^{(n)})$

3. $X_3^{(n+1)} \sim f_{3|1,2,4,\ldots,m}(x_3|x_1^{(n)},x_2^{(n+1)},x_4^{(n)},\ldots,x_m^{(n)})$

$\vdots$

m. $X_m^{(n+1)} \sim f_{m|1,\ldots,m-1}(x_m|x_1^{(n)},x_2^{(n+1)},\ldots,x_{m-1}^{(n+1)})$.

The principle behind Algorithm 4.13 is the same that we have illustrated for Algorithm 4.11.

5 The Ergodic Theorem

More details about the material of this section can be found in [15, 26, 13]. In this section we want to give some more details about the limit (23). In particular, we want to link the definition of ergodicity that we have given for Markov processes with the definition that you might have already seen in the context of dynamical system theory.

5.1 Dynamical Systems

Throughout this section $(\Omega,\mathcal{F})$ will denote a measurable space, as usual.

Definition 5.1 (Invariant measure and measure preserving map)**.** *Let* $\varphi : (\Omega,\mathcal{F}) \longrightarrow (\Omega,\mathcal{F})$ *be a measurable map and* μ *be a probability measure on* $(\Omega,\mathcal{F})$*. We define* $\varphi^*(\mu)$ *to be the* pushforward of the measure μ under φ*:*

$$\varphi^*(\mu)(A) := \mu(\varphi^{-1}(A)), \quad \forall\, A \in \mathcal{F}.$$

If $\varphi^*(\mu) = \mu$*, i.e. if*

$$\mu(\varphi^{-1}(A)) = \mu(A), \quad \forall\, A \in \mathcal{F} \tag{30}$$

then we say that the measure μ *is* invariant under φ *and that* φ *is* measure preserving *and preserves the measure* μ*.*

For a given map φ there will be in general more than one measure which is preserved by φ.

Definition 5.2 (Invariant set)**.** *A set $A \in \mathcal{F}$ is* invariant *under φ if $\varphi^{-1}(A) = A$. We denote by $\mathcal{J}_\varphi$ the set of invariant sets under φ:*

$$\mathcal{J}_\varphi := \{A \in \mathcal{F} : \varphi^{-1}(A) = A\}.$$

When it is clear from the context which map we are referring to, we will drop the subscript φ and simply write $\mathcal{J}$.

Definition 5.3 (Ergodic measure and ergodic map)**.** *A probability measure μ which is invariant under φ is said to be* ergodic *if the set $\mathcal{J}_\varphi$ is trivial i.e. if $\mu(A)$ is equal to either 0 or 1 for all $A \in \mathcal{J}_\varphi$. If this is the case then the map φ on $(\Omega, \mathcal{F}, \mu)$ is said to be* ergodic*.*

Let me stress again that for a given map φ one can in general find more than one measure that is invariant under φ and, among these, more than one that is ergodic.

Notice that in Definition 5.1 and Definition 5.3 the measure μ doesn't need to be a probability measure in the sense that we can still start with a finite measure and then normalize it. The above definitions assume that this normalization process has already been performed on the measures at hand. A very important fact about ergodic measures is the following.

Lemma 5.4. *Given a map φ, let $\mathcal{M}_\varphi$ be the set of φ-invariant measures. Then:*

- *Two ergodic measures in $\mathcal{M}_\varphi$ either coincide or they are mutually singular.*[15]
- *$\mathcal{M}_\varphi$ is a convex set. A measure μ is ergodic if and only if it is an extreme point of such a set i.e. if μ cannot be written as $t\mu_1 + (1-t)\mu_2$ for $t \in (0,1)$, $\mu_1, \mu_2 \in \mathcal{M}_\varphi$.*
- *As a consequence of the above, if φ admits only one invariant measure, that measure must be ergodic.*

Now we state the main theorem of this section. To this end we recall that the notation

$$L^1(\mathbb{P}) := \{f : (\Omega, \mathcal{F}, \mathbb{P}) \to S \text{ s.t. } \int_\Omega |f| \, d\mathbb{P} < \infty\},$$

where S in the above is any Polish space.

[15] Two finite measures μ and ν on the same measurable space $(E, \mathcal{E})$ are said to be mutually singular if there exist two disjoint sets $A, B \in \mathcal{E}$ such that $A \cup B = E$ and $\mu(A) = \nu(B) = 0$.

Theorem 5.5 (Ergodic Theorem)**.** *Let $X : (\Omega, \mathcal{F}, \mathbb{P}) \to \mathbb{R}^n$ be integrable, i.e. $X \in L^1(\mathbb{P})$ and φ be a measure preserving transformation that preserves $\mathbb{P}$. Then*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} X(\varphi^k(\omega)) = \mathbb{E}(X|\mathcal{J}).$$

The above limit holds a.s. and in L^1.

The immediate consequence of such a Theorem is the following corollary.

Corollary 5.6. *With the setting of Theorem 5.5, if the measure $\mathbb{P}$ is ergodic for φ, then*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} X(\varphi^k(\omega)) = \mathbb{E}(X).$$

5.2 Stationary Markov Chains and Canonical Dynamical Systems

Suppose we have a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a map $\varphi : \Omega \to \Omega$ that preserves the measure $\mathbb{P}$ and a r.v. $X : \Omega \to \mathbb{R}$. Let φ^m denote the m-th iterate of φ, i.e. $\varphi^0(\omega) = \omega$ and $\varphi^m(\omega) = \varphi(\varphi^{m-1})(\omega)$, for all $m \geq 1$ and define a sequence of r.v. $X_n(\omega)$ as follows

$$X_0(\omega) = X(\omega),\ X_1(\omega) = X(\varphi(\omega)), \ldots, X_n(\omega) = X(\varphi^n(\omega)), \ldots \qquad (31)$$

It is easy to check that the sequence of r.v. X_n constructed as above is a stationary sequence. Indeed let A be the set $A := \{\omega : (X_0(\omega), \ldots, X_m(\omega)) \in B\}$, for some arbitrary Borel set B in $\mathbb{R}^{m+1}$. Then for all $k \geq 0$

$$\begin{aligned}\mathbb{P}\{(X_k(\omega), \ldots, X_{m+k}(\omega)) \in B\} &= \mathbb{P}(\omega : \varphi^k(\omega) \in A) = \mathbb{P}(\varphi^{-k}(A)) \\ &= \mathbb{P}(A) = \mathbb{P}\{(X_0(\omega), \ldots, X_m(\omega)) \in B\}.\end{aligned}$$

We have hence just proved the following lemma.

Lemma 5.7. *Let $\mathbb{P}$ be a measure on $(\Omega, \mathcal{F})$, φ a map that preserves $\mathbb{P}$ and X a $\mathcal{F}$-measurable, $\mathbb{R}$-valued random variable. Then the sequence* (31) *is a stationary sequence.*

Comment. Lemma 5.7 is important because, at the cost of changing measure space, any stationary sequence can be represented in the form (31). Let us be more clear about this: we have already seen – see Remark 3.6 – that any Markov Chain, and hence in particular any stationary Markov Chain, can be represented as the coordinate map in the sequence space $\mathbb{R}^{\mathbb{N}}$ endowed

with the σ-algebra generated by cylinder sets and with the measure P constructed by using the Kolmogorov extension Theorem. So suppose that X_n is a stationary MC, $X_n : \Omega \to \mathbb{R}$. The corresponding process on $\mathbb{R}^{\mathbb{N}}$ which has the same distribution as X_n is the process $Y_n : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}$ such that $Y_n(\omega) = \omega_n$, $\omega = (\omega_0, \omega_1, \dots) \in \mathbb{R}^{\mathbb{N}}$. This means that Y_n will be stationary on $(\mathbb{R}^{\mathbb{N}}, \mathcal{R}^{\mathbb{N}}, P)$. If we consider the r.v. $Y : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}$ defined as $Y(\omega) = \omega_0$ and the shift map $\varphi : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}^{\mathbb{N}}$ defined as $\varphi(\omega_0, \omega_1, \dots) = (\omega_1, \omega_2, \dots)$, then one can prove that φ preserves P (you can check it by yourself, using the fact that X_n is stationary, see Exercise **??**) and indeed the sequence $Y_n(\omega)$ is nothing but $Y(\varphi^n(\omega))$. Clearly everything we have said still holds if we substitute $\mathbb{R}$ with any Polish space, i.e. if the random variable X takes values in a Polish space or in a subset of a Polish space.

Suppose from now on that the chain X_n takes values in a discrete space S. If X_n admits an invariant distribution (i.e. a stationary distribution in the sense of Definition 3.21), π, and we start the chain with $X_0 \sim \pi$ then the chain is stationary, i.e. $X_n \sim \pi$ for all $n \geq 0$. Denote by P_π the corresponding measure on path space obtained via the Kolmogorov's Extension Theorem. Then, by the comment above, P_π is invariant (in the sense of Definition 5.1[16]) for the shift map φ. The pair (φ, P_π), together with the space $(S^{\mathbb{N}}, \mathcal{S}^{\mathbb{N}})$ is called the *canonical dynamical system* associated with the Markov chain X_n which has π as invariant distribution and that is started at π, i.e. $X_0 \sim \pi$.

For a given initial stationary distribution π, the measure P_π is the only φ-invariant measure such that

$$P_\pi(\omega \in S^{\mathbb{N}} : \omega_i \in A_i, i = 1, \dots, m) = \mathbb{P}_\pi(X_1 \in A_1, \dots, X_m \in A_m).$$

Now suppose the invariant distribution π is such that $\pi(x) > 0$ for all x (which is in no way restrictive since any state with $\pi(x) = 0$ is irrelevant to the process as it cannot be visited). Under this assumption it is possible to show that the shift map φ is ergodic if and only if the chain X_n is irreducible (see for example [9, Section 3] for a proof in finite state space). Therefore there exists a unique invariant distribution for X_n, i.e. X_n is ergodic (in the sense of Definition 3.26).

Theorem 5.8 (Ergodic Theorem for stationary Markov Chains)**.** *Let X_n be a stationary Markov chain. Then for any integrable function, the limit*

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

[16] Another good reason why π is called invariant.

exists, a.s. and in L^1. If the chain is also ergodic then

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} f(X_k) = \mathbb{E}(f(X_0)).$$

Take $f = \mathbf{1}_A$ for some $A \in \mathcal{F}$. Then the above limit says exactly that asymptotically, space averages equal time averages.

We shall talk more about ergodicity and the link between Markov Processes and dynamical systems in the context of diffusion theory. For the time being I would like to mention the following result, which will be useful later on.

Lemma 5.9. *A measure preserving map on a probability space, $\varphi : (\Omega, \mathcal{F}, \mathbb{P}) \to (\Omega, \mathcal{F}, \mathbb{P})$ is ergodic if and only if*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \mathbb{P}((\varphi^{-k}A) \cap B) = \mathbb{P}(A)\mathbb{P}(B), \quad \textit{for all} \;\; A, B \in \mathcal{F}.$$

That is, ergodicity is equivalent to asymptotic average independence. This motivates the following definition, which introduces a property stronger than ergodicity.

Definition 5.10. *Let φ be measure preserving map on a probability space, as in the above Lemma 5.9. φ is said to be* (strongly) mixing *if*

$$\lim_{n\to\infty} \mathbb{P}((\varphi^{-n}A) \cap B) = \mathbb{P}(A)\mathbb{P}(B), \quad \textit{for all} \;\; A, B \in \mathcal{F}.$$

Start with noticing that mixing is stronger than ergodic, i.e. mixing $\Rightarrow$ ergodic. To have a better intuition about the definition of mixing system and the difference with an ergodic one suppose for a moment that the map φ is invertible; if this is the case, the definition (30) is equivalent to

$$\mathbb{P}(\varphi(A)) = \mathbb{P}(A) \quad \text{for all} \;\; A \in \mathcal{F}\,.$$

Therefore

$$\mathbb{P}((\varphi^{-n}A) \cap B) = \mathbb{P}(\varphi^n((\varphi^{-n}A) \cap B)) = \mathbb{P}(A \cap \varphi^n(B))\,.^{17}$$

Which means that if φ is invertible then the limit of Definition 5.10 is equivalent to

$$\lim_{n\to\infty} \frac{\mathbb{P}(A \cap (\varphi^n(B)))}{\mathbb{P}(A)} = \mathbb{P}(B), \quad \text{for all} \;\; A, B \in \mathcal{F}. \tag{32}$$

[17] It is true in general that $f(A \cap f^{-1}B) = f(A) \cap B$ but it is not always true that $f^{-1}(A \cap f(B))$ is equal to $f^{-1}(A) \cap B$.

Now let us think of an example that will be recurrent in the following: milk and coffe. So suppose we have a mug of coffe and we add some milk. At the beginning the milk is all in the region B. To fix ideas suppose $\mathbb{P}(B) = 1/3$ (which means that, after adding the milk, my mug will contain $1/3$ of milk and $2/3$ of coffee). The stain of milk will spread around and at time n it will be identified by $\varphi^n(B)$. The property (32) then means that asymptotically (i.e., in the long run) the proportion of milk that I will find in whatever subset A of the mug is precisely $1/3$. The limit of Lemma 5.9 says instead that such a property is only true asymptotically on average.

6 Continuous time Markov processes

The definition of Markovianity for a continuous time stochastic process is only formally different than the analogous definition for chains, Definition 3.1.

Definition 6.1. *Let $\{X_t\}_{t\in\mathbb{R}_+}$ be a continuous time stochastic process taking values on a general state space $(S, \mathcal{S})$ and let $\mathcal{F}_t$ be the filtration generated by X_t, $\mathcal{F}_t := \sigma(X_s, 0 \le s \le t)$. If*

$$\mathbb{P}(X_t \in B|\mathcal{F}_s) = \mathbb{P}(X_t \in B|X_s) \quad \textit{for all } 0 \le s \le t \textit{ and } B \in \mathcal{S}\,, \tag{33}$$

then X_t is a continuous time Markov process.

It is self-evident that (33) is just (11) in continuous time. In this context, time-homogeneity can be defined as follows.

Definition 6.2. *The Markov process X_t is* time-homogeneous *if*

$$\mathbb{P}(X_t \in B|X_s) = \mathbb{P}(X_{t-s} \in B|X_0), \quad \forall B \in \mathcal{S} \textit{ and } t \ge s \ge 0\,.$$

In discrete time, the transition functions needed to be assigned only for one time step so the transition probabilities were a function of two arguments only; in the present case the transition probabilities will depend on time as well.

Definition 6.3. *A map $p_t(x, B) := p(t, x, B) : \mathbb{R}_+ \times S \times \mathcal{S} \to [0, 1]$ is a* transition function *if*

1. *for fixed t and x, $p(t, x, \cdot)$ is a probability measure, i.e. $p(t, x, S) = 1$,*
2. *for fixed t and B, $p(t, \cdot, B)$ is a $\mathcal{S}$-measurable function.*

Given a transition function p, if

$$\mathbb{P}(X_t \in B | X_0 = x) = p_t(x, B), \quad \textit{for all } t \geq 0, \tag{34}$$

for some time-homogeneous Markov process X_t, then p is the transition function of the process X_t.

Notice that in order for the transition probabilities to satisfy (34) for $t = 0$ one needs to have

$$p_0(x, B) = \delta_x(B) = \begin{cases} 1 & \text{if } x \in B \\ 0 & \text{if } x \notin B \end{cases}.$$

In particular, for $t = 0$ the transition function can't have a density.

As in the case of discrete time, once an initial datum (or an initial distribution) for the process is chosen, the transition probabilities uniquely determine the Markov process. We now prove the Chapman-Kolmogorov relation.

Theorem 6.4 (Chapman-Kolmogorov)**.** *Let X_t be a time-homogeneous Markov process. Then, for all $0 \leq s < u < t$ and $B \in \mathcal{S}$, we have*

$$\mathbb{P}(X_t \in B | X_s) = \int_S \mathbb{P}(X_u \in dy | X_s) \mathbb{P}(X_t \in B | X_u = y) \tag{35}$$

$$= \int_S \mathbb{P}(X_{u-s} \in dy | X_0) \mathbb{P}(X_{t-u} \in B | X_0 = y). \tag{36}$$

We specify that the first equality is true for any Markov process, the second is due to time-homogeneity.

Proof of Theorem 6.4. Since (35) holds for any Markov process, we only need to prove (35), as (36) follows from (36), when X_t is time-homogeneous. Let $\mathcal{F}_t$ be the filtration generated by X_t. If $s < u < t$ then $\mathcal{F}_s \subset \mathcal{F}_u \subset \mathcal{F}_t$. We will use this fact, together with Markovianity and the properties of conditional expectation in order to prove (35):

$$\begin{aligned} \mathbb{P}(X_t \in B | X_s) &= \mathbb{E}[\mathbf{1}_{\{X_t \in B\}} | \mathcal{F}_s] \\ &= \mathbb{E}[\mathbb{E}[\mathbf{1}_{\{X_t \in B\}} | \mathcal{F}_u] | \mathcal{F}_s] = \mathbb{E}[\mathbb{P}(X_t \in B | X_u) | X_s] \end{aligned}$$

Now because for every measurable function φ, we can express the conditional expectation $\mathbb{E}(\varphi(X_t)|X_s) = \int \varphi(y) \mathbb{P}(X_t \in dy | X_s)$, we obtain the desired result. □

If we want the Chapman-Kolmogorov equation (C-K) to hold in a stronger sense, i.e. if we want

$$\mathbb{P}(X_t \in B|X_s = x) = \int_S \mathbb{P}(X_u \in dy|X_s = x)\mathbb{P}(X_t \in B|X_u = y) \qquad (37)$$

for all $x \in S$ and $0 \leq s \leq u \leq t$, then this is something that we need to require as an assumption from the transition probability themselves, as it doesn't follow from the Markov property alone. For this technical detail see for example [22, Chapter 3]. However it is always possible to modify the transition probabilities so that (37) holds, i.e. so that (35) holds pointwise (see [1, Chapter 2]). From now on we will always assume that this has been done and that the C-K equation holds in its stronger form (37). If the process is time-homogeneous, (37) can be rewritten as

$$\mathbb{P}(X_t \in B|X_s = x) = \int_S \mathbb{P}(X_{u-s} \in dy|X_0 = x)\mathbb{P}(X_{t-u} \in B|X_0 = y)\,. \quad (38)$$

Using the transition probabilities, the above can also be expressed as

$$p_{t+s}(x, B) = \int_S p_s(x, dy)p_t(y, B)\,.$$

If the transition functions have a density, i.e. if $p_s(x, dy) = p_s(x, y)dy$ for all $x \in S$, we can also write

$$p_{t+s}(x, z) = \int_S p_s(x, y)\, p_t(y, z)\, dy\,. \qquad (39)$$

Continuous time Markov processes arise mainly as solutions of SDEs. We will see in Section 8.2.3 that, under some conditions on the coefficients of the SDE, the solution of the equation enjoys the Markov property. Therefore, all the examples of Section 8.2.2 are examples of continuous time Markov processes. The next section is about the "most important" Markov process.

7 Brownian Motion

Brownian Motion was observed in 1827 by the Botanist Robert Brown, who first noticed the peculiar properties of the motion of a pollen grain suspended in water. What Brown saw looked something like the path in Figure 1 below. In 1900 Louis Bachelier modeled such a motion using the theory of stochastic processes and obtaining results that where rediscovered, although in a

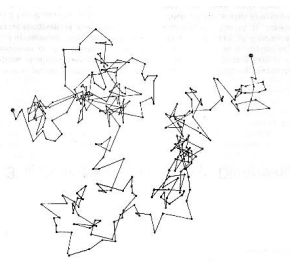


Figure 1: Brownian motion

different context, by A. Einstein in 1905. The understanding provided by Bachelier and Einstein was then put on firm mathematical basis by Norbert Wiener in the 1920's.

We have given a formal definition of BM in Example 2.7, which we now would like to justify. First of all, what is BM? In a container there are many "small" particles moving around - in Brown's case, water particles. Now suppose we put a "bigger" particle (the pollen grain) in our container. At each instant in time the small particles, which move in all possible directions, will kick our pollen grain and the result of this bombardment is an erratic motion called Brownian Motion. If $B(t)$ is the position of the pollen grain at time t, then it should be clear by our naive description that $B(t) - B(s)$ is independent from $B(s) - B(u)$ for any $u \leq s \leq t$, as the kicks that the grain receives are independent. If we put two pollen grains in the water then the two motions are independent. Also, the trajectory of the particle is so irregular that it is (a.s.) nowhere differentiable.

Exercise 7.1. Try and draw a continuous path that is nowhere differentiable. Managed it? Well look better, I am pretty sure that what you have drawn is a.s. differentiable.

7.1 Heuristic motivation for the definition of BM.

Consider a particle that moves along the x axis by jumping a distance ν every τ seconds. Every time the particle can jump left or right with equal probability. Suppose that the particle is at the origin at time 0. If we denote by $p(x,t)$ the probability of finding the particle in $x = \nu\, k$ ($k \in \mathbb{Z}$) at time

$t = n\,\tau$, we have

$$p^{\nu,\tau}(x, t+\tau) = \frac{1}{2}[p^{\nu,\tau}(x-\nu, t) + p^{\nu,\tau}(x+\nu, t)]\,^{18}.$$

If τ and ν are small then by expanding both sides of the above equation we get

$$p^{\nu,\tau}(x,t) + \frac{\partial p^{\nu,\tau}(x,t)}{\partial t}\tau = p^{\nu,\tau}(x,t) + \frac{1}{2}\frac{\partial^2 p^{\nu,\tau}(x,t)}{\partial x^2}\nu^2\,.$$

Now let $\tau, \nu \to 0$ in such a way that

$$\frac{\nu^2}{\tau} \to D,$$

where $D > 0$ is a constant, which we will later call the *diffusion constant*. Then $p^{\nu,\tau}(x,t) \to p(x,t)$, where $p(x,t)$ satisfies the *heat equation* (or *diffusion equation*):

$$\partial_t p(x,t) = \frac{1}{2} D\,\partial_{xx} p(x,t). \tag{40}$$

$p(x,t)$ is the probability density of finding the particle in x at time t. The fundamental solution of (40) is precisely the probability density of a Gaussian with mean zero and variance t, see (8).

The calculation that we have just presented shows another very important fact: pour some milk into your coffee and watch the milk particles diffuse around in your mug. We will later rigorously define the term *diffusion*; however for the time being observe that we already know that at the microscopic (probabilistic) level, the motion of the milk molecules is described by BM. The macroscopic (analytic) description is instead given by the heat equation.

7.2 Rigorous motivation for the definition of BM.

We look at the same picture as before, of a particle jumping a distance ν every τ seconds to the right or to the left with the same probability. This time, in order to rigorously derive BM from the random walk, we will use the CLT. To this end, let $B^{\nu,\tau}(t)$ denote the position of our particle at time $t = n\,\tau$ and let $\{X_i\}_i$ be i.i.d. r.v. with $\mathbb{P}(X_i = 0) = 1/2 = \mathbb{P}(X_i = 1)$. Consider $S_n = \sum_{i=1}^n X_i$, which clearly represents the number of moves to the right by time $t = n\tau$, and observe that

$$B^{\nu,\tau}(t) = S_n\,\nu + (n - S_n)(-\nu) = (2S_n - n)\,\nu.$$

[18] Later we will prove that the Wiener process is a Markov process. The Markovianity of the process is already in this equation.

Since $\mathbb{E}(X_i) = 1/2$ and $Var(X_i) = 1/4$, $\mathbb{E}(S_n) = n/2$ and $Var(S_n) = n/4$. Therefore, $\mathbb{E}(B^{\nu,\tau}(t)) = 0$ and

$$Var(B^{\nu,\tau}(t)) = 4\,\nu^2 \frac{n}{4} = \nu^2\, n = t\frac{\nu^2}{\tau}.$$

Assuming $\nu^2/\tau = D$, we can rewrite

$$B^{\nu,\tau}(t) = \frac{S_n - (n/2)}{\sqrt{n/4}}\sqrt{n}\nu = \frac{S_n - (n/2)}{\sqrt{n/4}}\sqrt{tD}.$$

Now the CLT implies

$$\begin{aligned}\lim_{\substack{n\to\infty \\ t=n\tau}} \mathbb{P}(a \le B^{\nu,\tau}(t) < b) &= \lim_{n\to\infty} \mathbb{P}\left(\frac{a}{\sqrt{tD}} \le \frac{S_n - (n/2)}{\sqrt{n/4}} \le \frac{b}{\sqrt{tD}}\right) \\ &= \frac{1}{\sqrt{2\pi}} \int_{\frac{a}{\sqrt{tD}}}^{\frac{b}{\sqrt{tD}}} e^{-\frac{x^2}{2}}\, dx;\end{aligned}$$

if $D = 1$ the RHS of the above is the same as the RHS of (8) after a change of variables.

7.3 Properties of Brownian Motion

From the description of BM as the motion of a pollen grain - and even more from its derivation as the limit of a random walk - it should be clear why the following result holds.

Theorem 7.2. *The Wiener process is a Markov process with*

$$\mathbb{P}(B(t) \in (a,b) | B(s)) = \frac{1}{\sqrt{2\pi(t-s)}} \int_a^b e^{-\frac{(x-B(s))^2}{2(t-s)}}\, dx\,. \tag{41}$$

Now an important definition.

Definition 7.3. *A $\mathbb{R}^n$-valued s.p. M_t is a* martingale *with respect to the filtration $\mathcal{E}_t$ if*

1. $\mathbb{E}\,|M_t| < \infty$ *for all* t
2. M_t *is adapted to* $\mathcal{E}_t$
3. $\mathbb{E}[M_t|\mathcal{E}_s] = M_s$ *for all* $s \le t$.

Martingales enjoy many nice properties, among which they satisfy the following inequality

Theorem 7.4 (Doob's martingale inequality)**.** *If M_t is a continuous martingale then*

$$\mathbb{P}\left(\sup_{t\in[0,T]}|M_t|\geq\lambda\right)\leq\frac{\mathbb{E}\,|M_T|^p}{\lambda^p},$$

for all $\lambda>0$, $T\geq 0$ and $p\geq 1$.

In order to understand the martingale's inequality, it is useful to compare it with the Markov inequality (4).

Theorem 7.5. *Brownian Motion is a martingale with respect to the filtration $\mathcal{F}_t=\sigma\{B(s);0\leq s\leq t\}$.*

Proof. Brownian motion is square integrable - as $\mathbb{E}(B(t)^2)=t$ from i) and ii) of Example 2.7 - and therefore integrable and it is adapted w.r.t. $\mathcal{F}_t$ by definition. Also, for any $t\geq s$,

$$\mathbb{E}[B(t)|\mathcal{F}_s]=\mathbb{E}[B(t)-B(s)+B(s)|\mathcal{F}_s]=0+B_s$$

as $B(t)-B(s)$ is independent of $\mathcal{F}_s$. □

Let us now look at the path properties of the Wiener process: we will see that while the trajectories of BM are continuous, they are a.s. nowhere differentiable. The non-differentiability is intimately related with the fact that BM has infinite total variation, which makes it impossible to define the stochastic integral w.r. to Brownian motion in the same way in which we define the Riemann integral.

Theorem 7.6. *The sample paths of BM are a.s. continuous.*

Proof. We want to use Kolmogorov's continuity Criterion, Theorem 2.5. For every $k>2$ and $0\leq s\leq t$,

$$\begin{aligned}\mathbb{E}\,|B(t)-B(s)|^k&=\frac{1}{\sqrt{2\pi(t-s)}}\int_{\mathbb{R}}|x|^k\,e^{-\frac{x^2}{2(t-s)}}\,dx\\(y=x/\sqrt{t-s})&=\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}}|y|^k\,(t-s)^{k/2}e^{-\frac{y^2}{2}}\,dy=C\,|t-s|^{k/2}\,.\end{aligned}$$

So we can apply the criterion with $\beta=k$ and $\alpha=k/2-1$, for all $k>2$ and we obtain the desired continuity. □

As a byproduct of the proof of Theorem 7.6 the Wiener process has γ-Hölder continuous paths for every exponent $0 < \gamma < \alpha/\beta = 1/2 - 1/k$, for all $k > 2$, i.e. the paths are γ-Hölder continuous for every $0 < \gamma < 1/2$. If a function is differentiable then it is Hölder continuous with exponent $\gamma = 1$. It turns out that we can prove that BM is not γ-Hölder continuous for $\gamma \geq 1/2$ and therefore it is a.s. not differentiable. But there is more to it.

Definition 7.7. *Let $\Pi = \{0 = t_0 \leq t_1 \leq \cdots \leq t_k = T\}$ be a partition of the interval $[0, T]$ and let $\|\Pi\| = \max(t_{i+1} - t_i)$ be the mesh of the partition. Let also Π_n be a sequence of refining partitions of $[0, T]$, such that $\|\Pi_n\| \to 0$. Given a continuous function g on $[0, T]$, we define the* p-th variation *of g as*

$$\lim_{n\to\infty} \sum_{t_i \in \Pi_n} |g(t_{i+1}) - g(t_i)|^p \ , \quad p > 0.$$

When $p = 1$ the 1-variation is more often called the total variation *of g.*

Continuously differentiable functions have finite total variation, i.e. if $g \in C^1[0, T]$ then the total variation of g is $\int_0^T |g'(s)|\, ds$.

Theorem 7.8. *The Wiener process $B(t)$ has a.s. infinite total variation and finite quadratic variation: if Π_n are refining sequences of the interval $[0, t]$ with mesh size tending to zero then*

1. $\lim_{n\to\infty} \sum_{t_i \in \Pi_n} |B(t_{i+1}) - B(t_i)| = \infty$ *a.s.*
2. $\lim_{n\to\infty} \sum_{t_i \in \Pi_n} |B(t_{i+1}) - B(t_i)|^2 = t$ *a.s.*

It is important to realize that property 1. in the above theorem is a consequence of the almost nowhere differentiability of BM.

Another property of BM that might be useful in the following is

$$\mathbb{E} \sup_{t\in[0,T]} |B(t)|^N \leq CT^{N/2}, \quad \text{fome some } \ C > 0 \ \text{ and for all } \ N \geq 1.$$

Now one last fact about BM. As we have seen, BM is almost surely non differentiable, so talking about its derivative

$$\text{„}\frac{dB(t)}{dt} = \xi(t)\text{”}$$

is a bit of a nonsense. However the ”derivative of Brownian Motion” is commonly referred to as *white noise*. Clearly, the process $\xi(t)$ doesn't exist,

at least not in any classical sense. It is possible to make sense of ξ as a distribution-valued process but this is not what we want to do here. For us $\xi(t)$ will be a stationary Gaussian process with autocovariance function

$$\mathbb{E}(\xi(t)\xi(s)) = \delta(t-s),$$

where $\delta = \delta_0$ is the delta function with mass at 0. The reason why ξ is called white noise is the following: if $c(t)$ is the covariance function of a stationary process then

$$f(\lambda) = \int_{\mathbb{R}} e^{-i\lambda t} c(t)dt, \quad \lambda \in \mathbb{R},$$

is the *spectral density* of the process. In the case of white noise,

$$f(\lambda) = \int_{\mathbb{R}} e^{-i\lambda t} \delta_0(t)dt = 1.$$

Inverting the Fourier transform, this means that all the frequencies contribute equally to the covariance function, in the same way in which all colours are equally present in the white colour.

8 Elements of stochastic calculus and SDEs theory

All of you know that the solution of an ODE, say

$$\frac{dX_t}{dt} = b(t, X_t), \tag{42}$$

looks something like this

Figure 2: Trajectory of the solution of an ODE

However, the paths of real life objects are more something like this

Figure 3: Trajectory of the solution of an SDE

This is due to at least two main reasons: i) the motion of the object that we are observing is subject to many small disturbances and ii) the graph of the path we are interested in is the result of our measurements, and measurements are subject to many small errors. This is why, when collecting data, it is rare to see a smooth curve. What we will more realistically see is the result of "smooth curve+ erratic effects". These random effects are commonly called *noise*. For this reason it is of great importance in the applied sciences to consider equations of the type

$$\frac{dX_t}{dt} = b(t, X_t) + "noise". \tag{43}$$

The first term on the RHS of this equation gives the average (smooth) behaviour, the second is responsible for the fluctuations that turn Figure 2 into Figure 3. Bear in mind that the solution to (42) is a function, the solution to (43) – whatever this equation might mean – is a stochastic process.

Depending on the particular situation that we want to model we can consider all sorts of noise. In this course we will be interested in the case of white noise ξ_t, which has been the first to be studied thanks to the good properties that it enjoys:

i) being mean zero, it doesn't alter the average behaviour, i.e. the drift is only due to $b(t, x)$;

ii) ξ_t is independent of ξ_s if $s \neq t$ (recall that $\mathbb{E}(\xi_t \xi_s) = \delta_0(t - s)$);

iii) it is stationary, so that its law is constant in time, i.e. the noise acting on the object we are observing is qualitatively the same at each moment in time.

Now just a small problem: we said that we can't make sense of ξ_t in a classic way as ξ_t is the derivative of Brownian motion, which is not differentiable.[19]

[19] Another, possibly better, way of looking at this matter is as follows: for practical

So, how do we make sense of (43)? Well, if "$\xi_t = dW_t/dt$" then, at least formally, (43) can be rewritten as

$$dX_t = b(t, X_t)dt + dW.$$

More in general, we might want to consider the equation

$$dX_t = b(t, X_t)dt + \sigma(t, X_t)dW. \tag{44}$$

In this way, we have that, so far still formally,

$$X_t = X_0 + \int_0^t b(s, X_s)ds + \int_0^t \sigma(s, W_s)dW_s\,. \tag{45}$$

Assuming that $\int_0^t |b(s, X_s)|\, ds < \infty$ with probability one, we know how to make sense out of the first integral on the RHS. The real problem is how to make sense of the second integral on the RHS. However, I hope we all agree that if we can rigorously define the integral

$$\int_0^T f(t, \omega)dW_t \tag{46}$$

for a stochastic process $f_t = f(t, \omega)$ [20] belonging to some appropriate class, then we are done making rigorous sense of (43). Equation (45) is a general *stochastic differential equation* (SDE), driven by BM.

8.1 Stochastic Integrals.

Unless otherwise specified, throughout this section we will be referring to real-valued stochastic processes.

Let us start by very briefly (and roughly) recalling what happens in the case of the Riemann integral: we want to define $\int_0^T g(t)dt$ where g is some deterministic *continuous* function. In order to do so, we start with defining the integral of *step functions*, i.e. functions of the form $\bar{g}(t) = \sum_{j=0}^K c_j\, \chi_{[t_j, t_{j+1})}$, where the c_j's are constants and $\{0 = t_0, t_1, \ldots, t_K = T\}$ is some partition of the interval $[0, T]$. For step functions we define

$$\int_0^T g(t)dt = \sum_{j=0}^K c_j\, (t_{j+1} - t_j)\ .$$

reasons, suggested for example by engineering problems, we want to consider a noise that satisfies the properties i), ii) and iii) listed above. However one can prove that there is no such a process on the path space $\mathbb{R}^{[0,\infty)}$. In any event, the intuition about ξ_t suggested by the these three properties, leads to think of ξ_t as the derivative of BM. More detail on this point are contained in Appendix A.

[20] In the integral (46) we wanted to stress the dependence on both t and ω.

At this point we consider a partition of $[0,T]$ of size 2^{-n} – i.e. given by $t_j = j/2^n$ (the notation for t_j should be $t_j^{(n)}$ but we drop the dependence on n to streamline the notation) for all j's such that $t_j < T$ and $t_j = T$ if $j/2^n \geq T$ – and observe that $\bar{g}_n(t) = \sum_{j\geq 0} g(t_j)\chi_{[t_j,t_{j+1})}$ is, for each $n \in \mathbb{N}$, a step function that approximates $g(t)$. If

$$\lim_{n\to\infty} \sum_{j\geq 0} g(t_j)(t_{j+1} - t_j) \quad \text{converges} \tag{47}$$

then we define

$$\int_0^T g(t)dt := \lim_{n\to\infty} \int_0^T \bar{g}_n(t)dt = \lim_{n\to\infty} \sum_{j\geq 0} g(t_j)(t_{j+1} - t_j). \tag{48}$$

I want to stress that if the integrand is continuous, the limit on the right hand side of (48) does not change if we evaluate g at any other point $t_* \in [t_j, t_{j+1}]$, i.e. for continuous functions we have:

$$\int_0^T g(t)dt = \lim_{n\to\infty} \sum_{j\geq 0} g(t_j)(t_{j+1} - t_j) = \lim_{n\to\infty} \sum_{j\geq 0} g(t_*)(t_{j+1} - t_j),$$

for every $t_* \in [t_j, t_{j+1}]$.

Inspired by this procedure we want to do something similar to define the stochastic integral. In doing so we need to bear in mind that we are attempting to integrate a stochastic process - as opposed to a function - with respect to another stochastic process and that while the integral of a function is, for each fixed T, a number, the integral (46) is, for every fixed T, a random variable. The analogous of step functions are the *elementary* or *simple processes* on $[0,T]$, i.e. processes of the form

$$\varphi(t,\omega) = \sum_{j=0}^{K} \phi_j(\omega)\chi_{[t_j,t_{j+1})}, \tag{49}$$

where $\{0 = t_0, t_1, \ldots, t_K = T\}$ is again some partition of the interval $[0,T]$ and the $\phi_j' s$ are random variables, (a.s.) uniformly bounded in j and ω. For processes of this form, it seems reasonable to define

$$\int_0^T \varphi(t,\omega)dW_t := \sum_{j=0}^{K} \phi_j(\omega)(W(t_{j+1}) - W(t_j)). \tag{50}$$

Then what we want to do is to find a sequence φ_n of elementary processes that approximate the process $f(t,\omega)$ and finally define the stochastic integral

(46) as the limit of the stochastic integrals of φ_n. The plan sounds good, except there are a couple of problems and points to clarify.

First of all, assuming the procedure of taking the limit (in some sense) of the integral of elementary process does work, we would want the limit to be independent of the point $t_* \in [t_j, t_{j+1}]$ that we choose to approximate the integrand. It turns out that this is not the case, not even if the integrand is continuous. We show this fact with an example.

Example 8.1. Suppose we want to calculate $\int_0^T W_t \, dW_t$. We consider the partition with mesh size 2^{-n} and approximate the integrand with the elementary process $\varphi_n = \sum_{j\geq 0} W(t_j)\chi_{[t_j,t_{j+1})}$. Then, for every $n \in \mathbb{N}$,

$$\mathbb{E}\sum_{j\geq 0} W(t_j)[W(t_{j+1}) - W(t_j)] = 0,$$

by using the fact that $W(t_{j+1}) - W(t_j)$ is independent of $W(t_j)$. If instead we approximate the integrand with the process $\tilde{\varphi}_n = \sum_{j\geq 0} W(t_{j+1})\chi_{[t_j,t_{j+1})}$, then

$$\begin{aligned}&\mathbb{E}\sum_{j\geq 0} W(t_{j+1})[W(t_{j+1}) - W(t_j)]\\ =&\mathbb{E}\sum_{j\geq 0}[W(t_{j+1}) - W(t_j)][W(t_{j+1}) - W(t_j)] \overset{n\to\infty}{\longrightarrow} T.\end{aligned}$$

Brownian motion is continuous so, what is going on? Both φ_n and $\tilde{\varphi}_n$ seem perfectly reasonable approximating sequences, so why are we obtaining different results? The problem is that Brownian motion, being a.s. non differentiable, simply "varies too much" in the interval $t_* \in [t_j, t_{j+1}]$ and this leads to the phenomenon illustrated by Example 8.1. There is no way of "solving" this pickle, it is simply a fact of life that different choices of $t_* \in [t_j, t_{j+1}]$ lead to different definitions of the stochastic integral. The most popular choices are

- $t_* = t_j$, which gives the *Itô integral*, and
- $t_* = (t_j + t_{j+1})/2$, which gives the *Stratonovich integral*.

We will discuss the differences between these two stochastic integrals later on. For the moment let us stick to the choice $t_* = t_j$ and talk about the Itô interpretation of (46).

8.1.1 Stochastic integral in the Itô sense.

In these notes we will not go through all the details of the proofs of the construction of the Itô integral but we simply provide the main steps of such a construction.

As we have already said, we want to find a sequence φ_n of elementary processes that approximate the process $f(t,\omega)$ and finally define the Itô integral as the limit of the stochastic integrals of φ_n. We still need to specify in what sense we will take the limit (and in what sense the φ_n's approximate f). In the case of (48), $\{\int_0^T \bar{g}_n(t)dt\}_n$ is simply a sequence of real numbers but in our case we are dealing with a sequence of random variables, so we need to specify in what sense they converge. It is clear from Theorem 7.8 that trying with L^1-convergence is recipe for disaster, as BM has infinite first variation. However, it has finite second variation, so we can try with L^2- convergence, which is what we are going to do. As you might expect, the procedure that we are going to sketch does not work for any integrand – in the same way in which not any function is Riemann integrable – but it is successful for stochastic processes $f(t,\omega) : \mathbb{R}_+ \times \Omega \to \mathbb{R}$ enjoying the following properties:

(a) $f(t,\omega)$ is $\mathcal{B} \times \mathcal{F}$- measurable;

(b) f_t is $\mathcal{F}_t$-adapted, for all $t \in \mathbb{R}_+$, where $\mathcal{F}_t$ is the natural filtration associated with the BM W_t ;

(c) $\mathbb{E}\int_0^T f_t^2\, dt < \infty$.

Definition 8.2. *We denote by $\mathfrak{I}(0,T)$, or simply $\mathfrak{I}$ when the extrema of integration are clear from the context, the class of stochastic processes $f(t,\omega) : \mathbb{R}_+ \times \Omega \to \mathbb{R}$ for which the above three properties hold.*

The reason why we impose this set of conditions on the integrands will be more clear once you read the construction procedure. However, property (b) comes from the fact that we are choosing the left hand point to approximate our integral, property (c) comes from the Itô isometry (see below). And now, finally, the main steps of the construction of the Itô integral are as follows:

1. Consider the set of elementary processes (49), where we require that the random variables ϕ_j are square integrable and $\mathcal{F}_{t_j}$-measurable [21] (with $\mathcal{F}_t$ the filtration generated by W_t); for these processes the

[21] This is because we are choosing $t^* = t_j$. Notice that this condition does not hold for the $\tilde{\varphi}_n$'s of Example 8.1, as $W_{t_{j+1}}$ is not $\mathcal{F}_{t_j}$-measurable.

stochastic integral is defined by (50). If $\varphi(t,x)$ belongs to such a set, then one can prove the following fundamental equality

$$\mathbb{E}\left(\int_0^T \varphi(t,\omega)\,dW_t\right)^2 = \mathbb{E}\int_0^T \varphi^2(t)\,dt. \tag{51}$$

The above equality is the *Itô isometry* for simple processes; it indeed establishes an isometry between $L^2(\Omega)$ and $L^2([0,T]\times\Omega)$, at least for simple processes at the moment. It is clear that (51) follows from point 2. of Theorem 7.8.

2. For any process $f(t,\omega) \in \mathfrak{I}$, there exists a sequence of elementary processes φ_n such that

$$\mathbb{E}\int_0^T |f(t)-\varphi_n(t)|^2\,dt \;\longrightarrow 0.$$

Therefore φ_n is a Cauchy sequence in $L^2([0,T]\times\Omega)$ (as it converges in this space), but also in $L^2(\Omega)$, thanks to (51). From the completeness of L^2, this means that $\int_0^T \varphi_t\,dW_t$ has a limit, which belongs to L^2. Such a limit is precisely the stochastic integral in the Itô sense of f_t.

Remark 8.3. The proof of the above goes as follows: once we define the stochastic integral for elementary processes, we can define it for bounded and continuous integrands $f_t \in \mathfrak{I}$. For such integrands the approximating sequence is precisely $\varphi_n(t) = \sum_j f(t_j)\chi_{[t_j,t_{j+1}]}$ (good to know for practical purposes). Notice that f_{t_j} is $\mathcal{F}_{t_j}$-measurable if $f_t \in \mathfrak{I}$. The next steps are proving that there is a sequence of approximants - with the required properties - for $f \in \mathfrak{I}$ which are bounded and then for any $f \in \mathfrak{I}$.

Now that we know what is a stochastic integral, we know how to interpret (45), at least when $\int_0^t \sigma\,dW$ is intended in the Itô sense. We will often use the shorter notation (44), which is to be interpreted to mean (45). When we want to refer to the Stratonovich definition, we use, as customary, the notation

$$\int_0^t \sigma(s,W_s)\circ dW_s\,.$$

Let us now list all the properties that help in the calculation of the stochastic integral.

Theorem 8.4 (Properties of the Itô integral)**.** *For every $f,g \in \mathfrak{I}(0,T)$, $t,s \in [0,T]$ and for every $\alpha,\beta \in \mathbb{R}$,*

(i) *Additivity:* $\int_0^T f\,dW = \int_0^t f\,dW + \int_t^T f\,dW$*;*

(ii) *Linearity:* $\int_0^T (\alpha f + \beta g)dW = \alpha \int_0^T f\,dW + \beta \int_0^T g\,dW$*;*

(iii) $\mathbb{E}\int_0^T f\,dW = 0$

(iv) $I_T := \int_0^T f\,dW$ *is* $\mathcal{F}_T$*-measurable;*

(v) I_t *is an* $\mathcal{F}_t$*-martingale (hence it satisfies Doob's inequality);*

(vi) I_t *is a continuous process (or better, it admits a continuous version)*

(vii) *Itô isometry:*

$$\mathbb{E}\left(\int_0^T f dW\right)^2 = \mathbb{E}\int_0^T f^2 dt,$$

or, more in general,

$$\mathbb{E}\left(\int_0^t f(u)dW_u \int_0^s g(u)dW_u\right) = \mathbb{E}\int_0^{t\wedge s} f(u)g(u)\,du\,.$$

(viii) *If* $f(t)$ *is a deterministic function then* I_t *is Gaussian with mean zero and variance* $\int_0^t f^2(s)ds$*.*

You surely remember that when you were taught the Riemann integral you were first required to calculate the integral from the definition and then, once you were provided with integration rules, to calculate more complicated integrals by using such rules. This is exactly what we will do here. The only difference is that in the Riemann case the integration rules follow from the differentiation rules. In this case we only have a stochastic integral calculus (see (52) and (55)) without a corresponding stochastic differential calculus.

Example 8.5. Calculate $\int_0^t W_s\,dW_s$, using the definition. The approximating sequence that we will use is the sequence φ_n of Example 8.1. If you think that $\int_0^t W_s\,dW_s = W_t^2/2$ you are in for a surprise. So we want to calculate the following limit in L^2

$$\lim_{n\to\infty}\int_0^t \varphi_n(s)dW_s = \lim_{n\to\infty}\sum_j W_{t_j}(W_{t_{j+1}} - W_{t_j}) = \lim_{n\to\infty}\sum_j W_{t_j}(\Delta W_{t_j}).$$

In order to do so, observe that $W_{t_j}(\Delta W_{t_j}) = \frac{1}{2}[\Delta(W_{t_j}^2) - (\Delta W_{t_j})^2]$. Therefore

$$\begin{aligned}\lim_{n\to\infty} \sum_j W_{t_j}(\Delta W_{t_j}) &= \frac{1}{2} \lim_{n\to\infty} \sum_j \Delta(W_{t_j}^2) - \frac{1}{2} \lim_{n\to\infty} \sum_j (\Delta W_{t_j})^2 \\ &= \frac{1}{2} W_t^2 - \frac{1}{2} t^2,\end{aligned}$$

where we have used the fact that the sum $\sum_j \Delta(W_{t_j}^2)$ is a telescopic sum and Theorem 7.8.

At this point you should be convinced that the Itô integral doesn't follow the usual integration rules. And you might wonder, what about the chain rule and integration by parts? Well I am afraid they don't stay the same either.

- Itô chain rule: suppose X_t satisfies the equation

  $$dX = b(t, X_t)dt + \sigma(t, X_t)dW_t\,.$$

  Then, if $g = g(t, x)$ is a $C^{1,2}([0, \infty) \times \mathbb{R})$ function,

  $$dg(t, X_t) = \frac{\partial g}{\partial t}(t, X_t)\, dt + \frac{\partial g}{\partial x}(t, X_t)\, dX_t + \frac{1}{2}\frac{\partial^2 g}{\partial x^2}(t, X_t)\sigma^2(t, X_t)dt\,. \tag{52}$$

- Product rule: if $X_i(t)$, $i = 1, 2$ satisfy the equation

  $$dX_i(t) = b_i(t, X_i(t))dt + \sigma_i(t, X_i(t))dW_t\,, \tag{53}$$

  respectively, then

  $$d(X_1X_2) = X_1dX_2 + X_2dX_1 + dX_1dX_2\,. \tag{54}$$

- Integration by parts, which follows from the product rule: with the same notation as in the product rule,

  $$\int_0^t X_1dX_2 = X_1(t)X_2(t) - X_1(0)X_2(0) - \int_0^t X_2dX_1 - \int_0^t dX_1dX_2\,. \tag{55}$$

How do we calculate the term $dX_1 dX_2$? I am afraid I will adopt a practical approach, so the answer is...by using this multiplication table:

$$dt \cdot dt = dW \cdot dt = dt \cdot dW = 0$$

while

$$dW \cdot dW = dt \quad \text{and} \quad dW \cdot dB = 0$$

if W and B are two independent Brownian motions. Just to be clear, in the case of X_1 and X_2 of equation (53), because the BM driving the equation for X_1 is the same as the one driving the equation for X_2, we have $dX_1 dX_2 = \sigma_1 \sigma_2 \, dt$. In this way if X_1 is a deterministic, say continuous function, (55) coincides with the usual integration by parts for the Riemann integral.

I would like to stress that this rule of thumb of the multiplication table can be rigorously justified (think of Theorem 7.8), but this is beyond the scope of this course.

Example 8.6. Let us calculate $\int_0^t W^3 dW$. We can use the integration by parts formula with $X_1 = W^3$ and $X_2 = W$. To use (55) we need to calculate dX_1 first, which we can do by applying (52) with $g(t,x) = g(x) = x^3$. So we get

$$dX_1 = 3W^2 dW + 3W dt,$$

hence $dX_1 dX_2 = 3W^2 dt$ and

$$\int_0^t W^3 dW = W^4 - \int_0^t W(3W^2 dW + 3W dt) - 3\int_0^t W^2 dt\,;$$

rearranging,

$$\int_0^t W^3 dW = \frac{W^4}{4} - \frac{3}{2}\int_0^t W^2 dt\,.$$

Before concluding this subsection, we would like to explicitly write down the multidimensional version of the main expressions that we have presented so far: suppose $X_t : [0,\infty) \times \Omega \to \mathbb{R}^d$ satisfies the following SDE

$$dX_t = b(t, X_t)dt + \sigma(t, X_t)dW_t$$

where $b(t,x) : [0,\infty) \times \mathbb{R}^d \to \mathbb{R}^d$, $\sigma(t,x) : [0,\infty) \times \mathbb{R}^d \to \mathbb{R}^d \times \mathbb{R}^m$, i.e. it is a $d \times m$ matrix, and W_t is a m-dimensional BM. In this case $\int_0^t \sigma(s, X_s) dW_s$ is simply a d-vector of Itô integrals, i.e. the i-th component of such a vector is

$$\left[\int_0^t \sigma \, dW_s\right]^i = \int_0^t \sum_{j=1}^m \sigma^{ij} dW^j\,.$$

If $g = g(t,x) : [0,\infty) \times \mathbb{R}^d \to \mathbb{R}$ and $Y = g(t, X_t)$, the multidimensional Itô formula reads:

$$dY(t,X_t) = \frac{\partial g}{\partial t}dt + \sum_{i=1}^{d} \frac{\partial g}{\partial x_i} dX^i + \frac{1}{2}\sum_{i,j=1}^{d} \frac{\partial^2 g}{\partial x_i \partial x_j} \sum_{l=1}^{m} \sigma^{il}\sigma^{jl} dt.$$

The same formula holds for each component of g, if $g = g(t,x) : [0,\infty) \times \mathbb{R}^d \to \mathbb{R}^p$: in this case for each $1 \le k \le p$,

$$dY^k(t,X_t) = \frac{\partial g^k}{\partial t}dt + \sum_{i=1}^{d} \frac{\partial g^k}{\partial x_i} dX^i + \frac{1}{2}\sum_{i,j=1}^{d} \frac{\partial^2 g^k}{\partial x_i \partial x_j} dX^i dX^j \,.$$

8.1.2 Stochastic integral in the Stratonovich sense.

The construction of the integral in the Stratonovich sense is very similar to the one of the Itô integral, so we won't repeat it.

We stress, once again, that the result of the Itô integration and the result of the Stratonovich integration do not in general coincide: $\int_0^t \sigma\, dW$ is not the same as $\int_0^t \sigma \circ dW$. However there is a useful conversion formula to write an Itô SDE in Stratonovich form and viceversa:

$$\begin{aligned} X_t &= X_0 + \int_0^t b(s,X_s)ds + \int_0^t \sigma(s,X_s) \circ dW_s \qquad (56)\\ &= X_0 + \int_0^t b(s,X_s)ds + \frac{1}{2}\int_0^t \sigma'(s,X_s)\sigma(s,X_s)ds + \int_0^t \sigma(s,X_s)dW_s \,. \end{aligned}$$

where $'$ denotes derivative with respect to x. The conversion formula (56) indicates that in going from the Stratonovich to the Itô formulation, the only part of the SDE that changes is the drift term, the diffusion coefficient remains unchanged. Moreover, the two equations coincide if the drift does not depend on x.

One point in favour of the Stratonovich formulation is that it obeys the ordinary chain rule: if

$$dX_t = b(t,X_t)dt + \sigma(t,X_t) \circ dW_t$$

then, for every smooth function (actually differentiable with continuous first derivatives is enough) $g = g(t,x)$, we have that the process $Y_t = g(t,X_t)$ solves the SDE

$$\begin{aligned} dY_t &= \frac{\partial g}{\partial t}(t,X_t)dt + \frac{\partial g}{\partial x}(t,X_t) \circ dX_t \\ &= \frac{\partial g}{\partial t}(t,X_t)dt + \frac{\partial g}{\partial x}(t,X_t)b(t,X_t)dt + \frac{\partial g}{\partial x}(t,X_t)\sigma(t,X_t) \circ dW_t \,. \end{aligned}$$

This is an important advantage of the Stratonovich integral, which turns out to be useful for example in the context of stochastic calculus on manifolds. However, $\int_0^t \sigma \circ dW$ is not a martingale while the Itô integral is a martingale, as we have seen. Moreover the Itô integral has the property of "not looking into the future" that makes it so popular in applications to finance – remember that the integrand needs to be only $\mathcal{F}_t$-measurable, for every t, so at time t the only information that we need is the one available up until that time. It is not the case that one of the two integrals is "better" than the other, but deciding whether we should look at

$$dX_t = b(t, X_t)dt + \sigma(t, X_t) \circ dW_t \tag{57}$$

rather than

$$dX_t = b(t, X_t)dt + \sigma(t, X_t)dW_t, \tag{58}$$

is clearly relevant, as the solution to (57) is not the same as the solution to (58); and this is not just a technical matter, it is rather a problem of modelling, and the answer depends on the specific application that we are considering.

By the modelling point of view, the Stratonovich integral has the big advantage of being stable under perturbations of the noise. Let us explain what we mean.

1. Assuming that (57) and (58) admit only one solution, let X_t be the solution of (58) and $\tilde{X}_t$ the solution of (57). In general $X_t \neq \tilde{X}_t$, as we have explained.

2. Now consider the equation

$$dX_t^k = b(t, X_t^k)dt + \sigma(t, X_t^k)\xi_t^k,$$

   where ξ^k is a sequence of smooth random variables that converge to white noise ξ as $k \to \infty$. Notice that because the ξ_k's are smooth, such an equation is, for each ω, a simple ODE, so we don't need to decide what we mean by it because we know it already.

3. If $k \to \infty$, you would expect that $X^k \to X$. Well this is not the case. Indeed, $X^k \to \tilde{X}$.

This is only the tip of the iceberg of the more general landscape described by the Wong-Zakai approximation theory, see for example [76, 41] and references therein. The fact that Itô SDEs are not stable under perturbations of the noise can be interpreted to mean, by the modelling point of view,

that unless we are absolutely sure that the noise to which our experiment is subject is actually white noise, the phenomenon at hand is probably better described by (57). On this topic, see for example [75]. After having studied Section 8.2.2 you will be able to solve Exercise 30, which gives a practical example of the procedure described above.

8.2 SDEs

From now on we focus on Itô SDEs, unless otherwise stated.

8.2.1 Existence and uniqueness.

We have so far handled equation (58) a bit clumsily, as in all fairness we haven't yet defined what we mean by a solution to (58). We will at first work with real valued processes but everything we say can be easily rewritten in higher dimension. To define what we mean by a solution to (58) we need some notation first. Let

- W_t be a one dimensional standard BM and $\mathcal{F}_t^W$ be the filtration generated by W_t;
- η be a random variable, independent of W_t, for all t;
- $\mathcal{G}_t$ be the filtration generated by η and W_t, i.e. for all t, $\mathcal{G}_t := \sigma(\eta, W_s; 0 \leq s \leq t)$;
- $\mathcal{N}$ be the family of null sets (i.e. sets of $\mathbb{P}$-measure zero) of the underlying probability space Ω;
- $\mathcal{F}_t$ be the σ-algebra generated by $\mathcal{N}$ and $\mathcal{G}_t$.

With this notation, let us give the following definition:

Definition 8.7 (Strong Solution)**.** *Given a Brownian motion W_t and an initial datum η, a* strong solution *to* (58) *is a continuous stochastic process X_t such that*

i) *X_t is $\mathcal{F}_t$-adapted, for all $t > 0$;*

ii) $\mathbb{P}(X_0 = \eta) = 1$*;*

iii) *for every $t \geq 0$, $\int_0^t (|b(s, X_s)| + |\sigma(s, X_s)|^2) ds < \infty$, $\mathbb{P}$ a.s.;*

iv) (58)*, or better* (45)*, holds almost surely.*

Such a solution is unique if, whenever $\hat{X}_t$ is another strong solution, $\mathbb{P}(X_t \neq \hat{X}_t) = 0$, for all $t \geq 0$.

Clearly the above definition can be extended to higher dimension if we take W_t to be a m-dimensional standard BM and condition iii) becomes

$$\int_0^t (\left|b^i(s, X_s)\right| + \left|\sigma^{ij}(s, X_s)\right|^2) ds < \infty, \quad \mathbb{P} - a.s.,$$

for all $t \geq 0$ and $1 \leq i \leq d, 1 \leq j \leq m$. Notice that in this definition the BM is assigned a priori, i.e. we know b, σ and W_t and we are looking for X_t. This might look like a silly remark, but it is not, as this is no longer true when we talk about *weak solutions*.

As for ODEs, we want to establish some general conditions under which the strong solution exists and is unique. We will state such a result directly for d-dimensional SDEs, so d and m are as at the end of Section 8.1.1. Also, for any $n \times \ell$ matrix A, we denote the Frobenius norm of A as follows:

$$\|A\|_F^2 := \sum_{i=1}^{n} \sum_{j=1}^{\ell} \left|a^{ij}\right|^2 .$$

Theorem 8.8. *Suppose the coefficients b and σ are globally Lipshitz and grow at most linearly, i.e. for all $t \geq 0$ and for all $x, y \in \mathbb{R}^d$ there exists a constant $C > 0$ (independent of x and y) such that*

$$\|b(t, x) - b(t, y)\| + \|\sigma(t, x) - \sigma(t, y)\|_F \leq C\|x - y\| \tag{59}$$

$$\|b(t, x)\|^2 + \|\sigma(t, x)\|_F^2 \leq C^2(1 + \|x\|^2). \tag{60}$$

Suppose also that there exists a r.v. η, independent of the m-dimensional BM and with finite second moment, i.e.

$$\mathbb{E}\|\eta\|^2 < \infty .$$

Then there exists a unique strong solution to equation (58), with initial condition $X(0) = \eta$. Such a solution is continuous and has bounded second moment. In particular, for every $T > 0$ there exists a constant $K > 0$, depending only on C and T, such that

$$\mathbb{E}\|X_t\|^2 \leq K(1 + \mathbb{E}\|X_0\|^2)e^{KT}, \quad \textit{for all } \ 0 \leq t \leq T .$$

Comment. You might be thinking that all this fuss of Itô and Stratonovich

is a bit useless as we could also look at an SDE as being an ODE, for each fixed ω. This is the approach taken in [73]. However bear in mind that because W_t is continuous but not C^1, you would end up with an ODE with non C^1 coefficients. In [73] is it shown that one can still makes sense of such an ODE by using a variational approach and that the solution found by means of the variational method coincides with the Stratonovich solution of the SDE. The conditions of the existence and uniqueness theorem, Theorem 8.8, can be somewhat weakened.

Theorem 8.9. *Suppose all the assumptions of Theorem 8.8 hold, but replace condition* (59) *with the following: for each N there exists a constant C_N such that*

$$\|b(t,x)-b(t,y)\|+\|\sigma(t,x)-\sigma(t,y)\|_F \leq C_N\|x-y\| \quad \textit{when } \|x\|,\|y\| \leq N\,. \tag{61}$$

Then there exists a unique strong solution to equation (58).

In other words, under an assumption of local lipshitzianity for the coefficients, the solution still exists and is unique. The linear growth condition instead is there to guarantee that the solution does not blow up. This is clear also from the thoery of ordinary ODEs. Consider for example the ODE

$$dX_t = X_t^2 dt, \quad X_0 = x.$$

Then the solution to this equation exists and is unique but it blows up in finite time:

$$X_t = \begin{cases} 0 & \text{if } x = 0 \\ \frac{x}{1-tx} & \text{if } x \neq 0\,. \end{cases}$$

8.2.2 Examples and methods of solution.

SDEs can be solved in the same way in which we solve ODEs, just taking into account the Itô chain rule.

Example 8.10 (Ornstein-Uhlenbeck process)**.** Consider the one dimensional equation

$$dX_t = \alpha X_t dt + \sigma dW_t \tag{62}$$

where $\alpha \in \mathbb{R}$ and $\sigma > 0$ are constants. The solution to this equation (which exists and is unique by Theorem 8.8) is the Ornstein-Uhlenbeck process (OU), which is a model of great importance in physics as well as in finance. X_t is also called *coloured noise* and, among other things, is used to describe the velocity of a Brownian particle. By the point of view of non-equilibrium

statistical mechanics, (62) is the simplest possible *Langevin equation*, but we will talk about this later. Now let us solve (62). As we would do for an ODE, let us write

$$dX_t - \alpha X_t dt = \sigma dW_t$$

and multiply both sides by $e^{-\alpha t}$:

$$e^{-\alpha t}(dX_t - \alpha X_t dt) = e^{-\alpha t}\sigma dW_t\,.$$

If we were dealing with an ODE we would now write

$$e^{-\alpha t}(dX_t - \alpha X_t dt) = d(e^{-\alpha t}X_t) \tag{63}$$

and integrate both sides. However we are dealing with an SDE so we need to use the chain rule of Itô calculus, equation (54). In this particular case, because the second derivative with respect to x of $e^{-\alpha t}x$ is zero, applying (54) leads precisely to (63) (check) – however I want to stress that this will not happen in general, as we will see from the next example. So now we can simply integrate both sides and find

$$X(t) = e^{\alpha t}X_0 + \int_0^t e^{\alpha(t-s)}\sigma dW_s\,.$$

Taking expectation on both sides and using property (iii) of Theorem 8.4 we get

$$\mathbb{E}(X_t) = e^{\alpha t}\,\mathbb{E}(X_0)\,,$$

so that $\mathbb{E}(X_t) \to 0$ if $\alpha < 0$ and $\mathbb{E}(X_t) \to +\infty$ if $\alpha > 0$. If X_0 is Gaussian or deterministic then the OU process is Gaussian, as a consequence of Theorem 8.4, property (vii). Moreover, it is a Markov process. The proof of the latter fact is a corollary of the results of the next section.

Example 8.11 (Geometric Brownian Motion)**.** The SDE

$$dX_t = rX_t\,dt + \sigma X_t\,dW_t\,,$$

is a simple population growth model (but it is quite popular in finance as well), where X_t represents the population size and $\tilde{r} = r + \sigma dW$ is the relative rate of growth. To solve this SDE we proceed as follows:

$$\frac{dX_t}{X_t} = r\,dt + \sigma\,dW_t\,. \tag{64}$$

At the risk of being boring, we can't say that $d(\log X_t) = dX_t/X_t$, because we are deling with the stochastic chain rule. So, applying the Itô chain rule, we find

$$d(\log X_t) = \frac{dX_t}{X_t} - \frac{1}{2}\sigma^2 dt. \tag{65}$$

From (64) and (65) we then have

$$d(\log X_t) = \left(r - \frac{1}{2}\sigma^2\right) dt + \sigma dW_t\,,$$

which gives

$$X_t = X_0\, e^{\left(r-\frac{1}{2}\sigma^2\right)t+\sigma W_t}\,.$$

Example 8.12. Also for SDEs it makes sense to try and look for a solution in product form. For example, consider the SDE

$$dX_t = d(t)X_t\, dt + f(t)X_t\, dW_t\,, \quad X(0) = X_0\,. \tag{66}$$

If $d(t)$ and $f(t)$ are uniformly bounded on compacts then the assumptions of the existence and uniqueness theorem are satisfied and we can look for the solution. So assuming for example that $d(t)$ and $f(t)$ are continuous will do. We look for a solution in the form

$$X(t) = Y(t)\, Z(t),$$

where $Y(t)$ and $Z(t)$ solve, respectively,

$$\begin{cases} dY = f\, Y\, dW \\ Y(0) = X_0 \end{cases} \qquad \text{and} \qquad \begin{cases} dZ = A(t)\, dt + B(t) dW \\ Z(0) = 1\,. \end{cases}$$

Applying the Itô product rule,

$$d(YZ) = (fX + BY)dW + Y(A + fB)dt\,.$$

We now compare the above expression with (66); in order for the two expressions to be equal, we need to have

$$fX + BY = fX \quad \text{and} \quad d(t)X_t = Y(A + fB);$$

the first equation gives $B \equiv 0$ so that the second implies $A(t) = d(t)Z_t$. The equation for Z becomes therefore deterministic:

$$dZ = d(t)Z dt\,, \quad Z(0) = 1 \;\Rightarrow\; Z_t = \exp\left(\int_0^t d(s)ds\right).$$

The equation for Y can be solved similarly to what we have done for geometric Brownian motion:

$$\begin{aligned} dY &= f(t)Y dW\,, \quad Y(0) = X_0 \\ \Rightarrow \frac{dY}{Y} &= f(t)dW\,. \end{aligned}$$

Because $d(\log Y) = dY/Y - f^2 dt/2$, we get

$$d(\log Y) = f dW - \frac{1}{2} f^2 dt \;\Rightarrow\; \log \frac{Y_t}{Y_0} = \int_0^t f dW - \frac{1}{2}\int_0^t f^2 dt$$
$$\Rightarrow Y(t) = X_0 \exp\left(\int_0^t f dW - \frac{1}{2}\int_0^t f^2 ds\right).$$

Putting everything together, finally

$$X(t) = X_0 \exp\left(\int_0^t f dW - \frac{1}{2}\int_0^t f^2 ds + \int_0^t d(s) ds\right). \tag{67}$$

8.2.3 Solutions of SDEs are Markov Processes

The solution of an Itô SDE is also called an *Itô process*. In this section we will always work under the assumptions of the existence and uniqueness theorem and we will prove two very important facts about the solutions of Itô SDEs: first of all, the solution of an Itô SDE is a continuous time Markov process. Secondly, if the coefficients of the equation do not depend on time, the solution is a time-homogeneous Markov process.

Theorem 8.13. *Suppose the assumptions of Theorem 8.8 hold. Then the solution of the SDE*

$$dX_s = b(s, X_s)\, ds + \sigma(s, X_s)\, dW_s, \quad X_0 = x, \tag{68}$$

is a Markov process.

The proof of this theorem is not particularly instructive, as the technicalities obfuscate the intuition behind them. The reason why the above theorem holds true is morally simple: we know by the existence and uniqueness theorem that, once X_s is known, the solution of (68) is uniquely determined for every $t \geq s$. Theorem 8.8 doesn't require any information about X_u for $u < s$ in order to determine X_u for $u > s$, once X_s is known. Which is why the above statement holds under the same assumptions as Theorem 8.8.

Theorem 8.14. *Suppose the assumptions of Theorem 8.8 hold. If the coefficients of the Itô SDE* (58) *do not depend on time then the Itô process is time-homogeneous. I.e., the solution of the equation*

$$dX_s = b(X_s)\,ds + \sigma(X_s)\,dW_s, \quad X_0 = x,$$

is a time-homogeneous Markov process.

As we have already said, a time-homogeneous process is a process describing a phenomenon that happens in conditions that do not change in time. Because the coefficients b and σ describe precisely the conditions in which the phenomenon happens, if such coefficients are time independent it is almost tautological that the solution of the equation should be time-homogeneous.

Proof of Theorem 8.14. The Markovianity is a consequence of Theorem 8.13, so we only need to prove time-homogeneity, which means that we need to prove that

$$\mathbb{P}(X_u \in B | X_0 = x) = \mathbb{P}(X_{u+t} \in B | X_t = x), \quad \forall B \in \mathcal{S},\, x \in S, t \geq 0\,.$$

Denote by $X^{x,t}(u+t)$ the solution of the equation

$$\beta(t+u) = x + \int_t^{t+u} b(\beta(s))ds + \int_t^{t+u} \sigma(\beta(s))dW_s \tag{69}$$

and by $X^{x,0}(u)$ the solution of the equation

$$\beta(u) = x + \int_0^{u} b(\beta(s))ds + \int_0^{u} \sigma(\beta(s))dW_s\,. \tag{70}$$

What we want to prove is that $X^{x,t}(u+t)$ has the same distribution as $X^{x,0}(u)$. To this end, notice that if W_v is a standard BM then $W_{t+v} - W_t$ is a standard BM as well (check), i.e. $\tilde{W}_v = W_{t+v} - W_t$ has the same distribution as W_v. With this in mind, a simple change of variable concludes the proof. Indeed, the RHS of (140) can be rewritten as

$$\beta(t+u) = x + \int_0^{u} b(\beta(v+t))dv + \int_0^{u} \sigma(\beta(v+t))dW_{t+v}\,.$$

At this point, because $\tilde{W}$ and W have the same distribution, and the above is nothing but (141), when we replace W with $\tilde{W}$, it is clear by the uniqueness of the solution that $\beta(t+u)$ has the same distribution as $\beta(u)$, which is, $X^{x,t}(u+t)$ has the same distribution as $X^{x,0}(u)$. □

8.3 Langevin and Generalized Langevin equation

A very interesting example is provided by the so called *Langevin equation*

$$\ddot{q}(t) = -\partial_q V(q) - \gamma \dot{q}(t) + f(t). \tag{71}$$

In the above equation $\gamma > 0$ is a constant, $V(q)$ is a confining[22] potential and $f(t)$ is white noise. Therefore equation (71) describes the position $q(t) \in \mathbb{R}$ of a particle subject to Newton's equation of motion plus dumping and noise. The Langevin equation was introduced as a stochastic model for chemical reactions, in which a particle, held by intermolecular forces, undergoes the reaction when activated by random molecular collisions. In this framework, the term $-\gamma\dot{q}$ expresses the rate at which the reaction slows down due to such random interactions. Notice that equation (71) can be rewritten in a form more familiar to you, i.e. as a system of first order SDEs, by just enlarging the state space with the introduction of the momentum variable $p(t)$:

$$\begin{aligned} \dot{q}(t) &= p(t) \\ \dot{p}(t) &= -\partial_q V(q) - \gamma p(t) + f(t)\,. \end{aligned} \tag{72}$$

If $V(q) = q^2/2$, by setting

$$B = \begin{bmatrix} 0 & 1 \\ -1 & -\gamma \end{bmatrix} \quad \text{and} \quad \Sigma = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \tag{73}$$

you can observe that (72) is nothing but an O-U process, this time in two dimensions:

$$dX_t = BX_t + \Sigma dW_t, \quad \text{where} \quad X_t = \begin{bmatrix} q(t) \\ p(t) \end{bmatrix},$$

and $W_t = (W_t^1, W_t^2)$ is a two-dimensional standard BM, i.e. W^1 and W^2 are independent one dimensional standard BMs. However notice that the *diffusion matrix*[23] Σ is degenerate (in the sense that it has zero determinant); for this reason this is a *degenerate* process (in particular one can prove that this is an *hypoelliptic* process). We will comment again on this later on, as the degeneracy of the diffusion matrix adds big difficulties to the analysis of such a process.

[22] A function $g(x) : \mathbb{R}^n \to \mathbb{R}$ is said to be confining if $g(x) \to \infty$ when $\|x\| \to \infty$.

[23] We shall see in the next section why this object is called *diffusion matrix*.

The model (72) assumes the collisions between molecules to occur instantaneously, hence it is not valid in physical situations in which such approximation cannot be made. In these cases a better description is given by the *Generalized Langevin Equation* GLE :

$$\ddot{q}(t) = -\partial_q V(q) - \int_0^t ds\, \gamma(t-s)\dot{q}(s) + F(t). \tag{74}$$

Here $\gamma(t)$ is a smooth function of t and $F(t)$ is a mean zero stationary Gaussian noise. $F(t)$ and the memory kernel $\gamma(t)$ are related through the following *fluctuation – dissipation relation*

$$\mathbb{E}(F(t)F(s)) = \beta^{-1}\gamma(t-s), \tag{75}$$

where $\beta > 0$ is a constant, representing the inverse temperature of the system. The GLE together with the fluctuation–dissipation theorem principle (75) appears in various applications such as surface diffusion and polymer dynamics. It also serves as one of the standard models of nonequilibirum statistical mechanics, describing the dynamics of a small Hamiltonian system (a distinguished particle) coupled to a heat bath at inverse temperature β. Just to provide some context: given a system in equilibrium, we can drive it away from its stationary state by either coupling it to one or more large Hamiltonian systems or by using non-Hamiltonian forces. In the Hamiltonian approach, which is often referred to as the *open systems theory* [66], the system we are interested in is coupled to one or more heat reservoirs. Multiple reservoirs and the non-Hamiltonian methods are used to study *non equilibrium steady states* (for example if we consider two heat baths a different temperatures, you can imagine that there will be a constant heat flux from the warmer to the colder reservoir; such a state is called a non-equilibrium steady state), see [68, 28]; coupling with a single heat bath is used to study return to equilibrium.[24] The GLE serves precisely this purpose. Let us now say a couple of words about the derivation of the GLE. In order to derive such an equation – which, I would like to point out, is a stochastic integro-differential equation – we think of the system "particle + bath" as a mechanical system in which a distinguished particle interacts with n heat bath molecules of mass $\{m_j\}_{1\le j\le n}$, through linear springs with random stiffness parameter $\{k_j\}_{1\le j\le n}$; the Hamiltonian of the system is

[24] Quoting D.Ruelle: *The purpose of nonequilibrium statistical mechanics is to explain irreversibility on the basis of microscopic dynamics, and to give quantitative predictions for dissipative phenomena.*

then :

$$H(q_n, p_n, Q_1, ..., Q_n, P_1, ..., P_n) = V(q_n) + \frac{p_n^2}{2} + \frac{1}{2}\sum_{i=1}^{n} \frac{P_i^2}{m_i} + \frac{1}{2}\sum_{i=1}^{n} k_i (Q_i - q_n)^2, \tag{76}$$

where (q_n, p_n) and $(Q_1, ..., Q_n, P_1, ..., P_n)$ are the positions and momenta of the tagged particle and of the heat bath molecules, respectively (the notation (q_n, p_n) is to stress that the position and momentum of the particle depend on the number of molecules it is coupled to).

Excursus. The theory of Equilibrium statistical mecanics is involved with the study of many particle systems in their equilibirum state; such a theory has been put on firm ground by Boltzman and Gibbs in the second half of the 19th century. Consider an Hamiltonian system with Hamiltonian $H(q, p)$ (for example our gas particles in a box). The formalism of equilibrium statistical mechanics allows to calculate macroscopic properties of the system starting from the laws that govern the motion of each particle (good reading material are the papers [46, 45, 68]). The *Boltzmann-Gibbs prescription* states the following: the equilibrium value of any quantity at inverse temperature β is the average of that quantity in the canonical ensamble, i.e. it is the average with respect to the *Gibbs measure* $\rho_\beta(q, p) = \mathcal{Z}^{-1} \exp\{-\beta H(q, p)\} dqdp$. The reason why this measure plays such an important role is a consequence of the fact that the Hamiltonian flow preserves functions of the Hamiltonian as well as the volume element $dqdp$.

Going back to where we were, we can write down the equations of motion of the system with Hamiltonian (76):

$$\begin{cases} \dot{q}_n &= p_n \\ \dot{p}_n &= -\partial_q V(q_n) + \sum_{i=1}^{n} k_i (Q_i - q_n) \\ \dot{Q}_i &= P_i / m_i \qquad\qquad 1 \le i \le n \\ \dot{P}_i &= -k_i (Q_i - q_n) \qquad 1 \le i \le n. \end{cases}$$

The initial conditions for the distinguished particle are assumed to be deterministic, namely $q_n(0) = q_0$ and $p_n(0) = p_0$; those for the heat bath are randomly drawn from a *Gibbs distribution at temperature* β^{-1}, namely

$$\rho_\beta := \frac{1}{\mathcal{Z}} e^{-\beta H}, \quad \mathcal{Z} \text{ normalizing constant},$$

where the Hamiltonian H has been defined in (76). Integrating out the heat bath variables we obtain a closed equation for q_n, of the form (74). In the

thermodynamic limit as $n \to \infty$ we recover the GLE. Under the assumption that at time $t = 0$ the heat bath is in equilibrium at inverse temperature β, we obtain the fluctuation dissipation relation (75) as well. The form of the memory kernel $\gamma(t)$ depends on the choice of the distribution of the spring constants of the harmonic oscillators in the heat bath [23].

It is important to point out that, for a general memory kernel $\gamma(t)$, the GLE is non Markovian and this feature makes it not very amenable to analysis (even though it is worth noticing that significant progress has been mad in the study of non-Markovian processes [27]). One way to go about this problem is trying to approximate the non Markovian dynamics (74) with a Markovian one. As noticed in [43], the problem can be recast as follows: we want to approximate the non Markovian process (74) with the Markovian dynamics given by the system of ODEs:

$$\begin{aligned} dq &= p\,dt && \text{(77a)} \\ dp &= -\partial_q V(q)\,dt + g \cdot u\,dt && \text{(77b)} \\ du &= (-p\,g - \mathcal{A}u)\,dt + C\,dW(t)\,, && \text{(77c)} \end{aligned}$$

where $(q, p) \in \mathbb{R}^2$, u and g are column vectors of $\mathbb{R}^d$, $\cdot$ denotes Euclidean scalar product, $W(t) = (W_1(t), \ldots, W_d(t))$ is a d-dimensional Brownian motion, $V(q)$ is a potential and $\mathcal{A}$ and C are constant coefficients $d \times d$ matrices, related through the fluctuation dissipation principle:

$$\mathcal{A} + \mathcal{A}^T = CC^T. \tag{78}$$

We also recall that the noise in (74) is Gaussian, stationary and mean zero. Because the memory kernel and the noise in (74) are related through the fluctuation-dissipation relation, the rough idea is that we might try to either approximate the noise and hence obtain the corresponding memory kernel or, the other way around, we could approximate the correlation function and read off the noise. The latter is the approach that we shall follow in this section. As a motivation, we would like to notice that for some specific choices of the kernel, equation (74) is *equivalent* to a finite dimensional Markovian system in an extended state space. If, for example, we choose $\gamma(t) = \lambda^2 e^{-t}$, $t > 0$, then (74) becomes

$$\begin{cases} \dot{q} &= p \\ \dot{p} &= -\partial_q V(q) - \lambda^2 \int_0^t e^{-(t-s)} p(s) ds + F(t); \end{cases} \tag{79}$$

the fluctuation dissipation theorem (with $\beta = 1$) yields

$$E(F(t+s)F(t)) = \lambda^2 e^{-|s|}. \tag{80}$$

Since we are requiring $F(t)$ to be stationary and Gaussian, (80) implies that $F(t)$ is the Ornstein-Uhlenbeck process.[25] If we write $F(t) = \lambda v(t)$, with $v(t)$ satisfying the equation $\dot{v} = -v + \sqrt{2}\dot{W}$, and we define the new process

$$z(t) = -\lambda \int_0^t e^{-(t-s)} p(s) ds + v(t), \tag{81}$$

then (79) becomes

$$\begin{cases} \dot{q} &= p \\ \dot{p} &= -\partial_q V + \lambda z \\ \dot{z} &= -\lambda p - z + \sqrt{2}\dot{W}, \end{cases}$$

which is precisely system (77) with $d = 1$, $\mathcal{A} = 1$ and $g = \lambda$. The "Markovianization" of (74) was first done by Mori [54] by first approximating the Laplace transform of the memory kernel $\gamma(t)$, $\tilde{\gamma}(\xi)$, by a rational function (if and when this is possible) and then imposing the fluctuation relation, which gives the matrices $\mathcal{A}$ and C as well as the vector g. If $\gamma(t)$ itself is a sum of exponentials, $\gamma_d(t) = \sum_{i=1}^d \lambda_i^2 e^{-\alpha_i t}$, then $\tilde{\gamma}_d = \sum_{i=1}^d \lambda_i^2/(\xi + \alpha_i)$, so the procedure indicated by Mori is clearly successful and it corresponds to the case in which $\mathcal{A} = \mathrm{diag}\{\alpha_1, \ldots, \alpha_d\}$ and $g = (\lambda_1, \ldots, \lambda_d)^T$. Another typical situation is when the Laplace transform of γ has a continued fraction representation

$$\tilde{\gamma}(\xi) = \cfrac{\epsilon_1^2}{\xi + \theta_1 + \cfrac{\epsilon_2^2}{\xi + \theta_2 + \cfrac{\epsilon_3^2}{\xi + \theta_3 + \ddots}}}, \qquad \theta_i > 0.$$

In this case the approximation is done by truncating the fraction at step d and then reading off the corresponding Markovian system of $(d+2)$ SDEs. The matrix $\mathcal{A}$ is then tridiagonal,

$$\mathcal{A} = \begin{vmatrix} \theta_1 & -\epsilon_2 & & & \\ \epsilon_2 & \theta_2 & -\epsilon_3 & & \\ & \epsilon_3 & \theta_3 & \ddots & \\ & & \ddots & \ddots & \\ & & & & \theta_d \end{vmatrix}$$

and $g = (\epsilon_1, 0, \ldots, 0)^T$.

[25] It is possible to show that the only mean zero stationary Gaussian process with autocorrelation function e^{-t} is the stationary O-U process.

9 Markov Semigroups and their Generator

The properties of Markov processes can be studied by means of functional analytic tools, which we come to introduce in this section. The (very little) background on functional analysis that you need to study this section is in the appendix. The framework we describe will refer to a continuous-time setting; therefore it will be applied to continuous time Markov processes and, in the next section, to diffusion processes (which are a special subclass of the continuous time Markov processes). However some of the definitions and results presented in the following, can also be formulated in discrete-time.

For the content of this section we refer to [25, 13, 62]. In all that follows $(\mathfrak{B}, \|\cdot\|)$ will denote a real Banach space.

Definition 9.1 (Markov Semigroup)**.** *A one parameter family of linear bounded operators over a Banach space $\{\mathcal{P}_t\}_{t\in\mathbb{R}_+}$, $\mathcal{P}_t : \mathfrak{B} \to \mathfrak{B}$ for all $t \geq 0$, is a* semigroup *if*

1. *$\mathcal{P}_0 = I$, where I denotes the identity on $\mathfrak{B}$;*
2. *$\mathcal{P}_{t+s} = \mathcal{P}_t\mathcal{P}_s = \mathcal{P}_s\mathcal{P}_t$, for all $t, s \in \mathbb{R}_+$.*

If the map $\mathbb{R}_+ \ni t \longrightarrow \mathcal{P}_t f \in \mathfrak{B}$ is continuous for all $f \in \mathfrak{B}$, then the semigroup is said to be strongly continuous*. A strongly continuous semigroup of bounded linear operators is also called a $\mathcal{C}_0$-*semigroup*. A semigroup is a* Markov semigroup *if*

i) *it* preserves constants*: $\mathcal{P}_t 1 = 1$ for all $t \in \mathbb{R}_+$ (here 1 is the constant function equal to one);*

ii) *it is* positivity preserving*: $f \geq 0 \Rightarrow \mathcal{P}_t f \geq 0$ for all $t \in \mathbb{R}_+$.*

Definition 9.2. *With the notation of the previous definition, the semigroup P_t is contractive if*

$$\|\mathcal{P}_t f\| \leq \|f\|, \quad \text{for all } t \in \mathbb{R}_+ \text{ and } f \in \mathfrak{B}.$$

Comment. *i*) You are likely to have already seen the definition of semigroup when talking about the solution of differential equations, for example linear equations of the type

$$\partial_t f = Lf,$$

where, for every fixed t, $f(t, x)$ belongs to some Banach space $\mathfrak{B}$ and L is a differential operator on $\mathfrak{B}$. Indeed properties 1. and 2. simply say that the solution of the equation is unique. Strong continuity is used to make sense

out of the notion of initial datum for the equation.
ii) Let $p_t(x, dy)$ be the transition function of a continuous time and time-homogeneous Markov process X_t . To fix ideas suppose X_t is real valued. Then

$$\mathcal{P}_t f(x) = \mathbb{E}\left[f(X_t)|X_0 = x\right] = \int_{\mathbb{R}} f(y) p_t(x, dy), \tag{82}$$

defines a positivity preserving semigroup with $\mathcal{P}_t 1 = 1$ (check) on the space $\mathcal{B}_m$ of bounded and measurable functions from $\mathbb{R}$ to $\mathbb{R}$. Whether the semigroup is strongly continuous or not, this depends on the process X_t and on the function space that we restrict to. The Banach spaces that we will consider will be function spaces and we assume that such spaces will contain at least the space $\mathcal{B}_m$. Another popular space for us will be the space $\mathcal{C}_b$ of bounded and continuous functions from $\mathbb{R}$ to $\mathbb{R}$ (more in general one can consider functions defined on a Polish space).

The semigroup (82) is the *Markov semigroup associated with the process* X_t. Such a semigroup will be the main focus of the remainder of this set of lecture notes. The aim of the game will be deriving properties of the Markov process X_t from properties of the semigroup $\mathcal{P}_t$. The fascinating bit here is that by doing so, we are trying to solve a stochastic problem by purely functional analytic means.

Example 9.3. Let $(\Omega, \mathcal{F}, \mu)$ be a probability space, $\mathcal{C}_b(\Omega)$ be the space of continuous and bounded functions on the space Ω (endowed with the uniform norm) and $m > 0$ a constant. Then

$$\mathcal{P}_t f(x) := e^{-mt} f(x) + (1 - e^{-mt})\mathbb{E}(f),$$

where $\mathbb{E}(f) := \int_\Omega f(x) d\mu(x)$*, defines a strongly continuous Markov semigroup.*

Example 9.4. The relation

$$\mathcal{P}_t f(x) := \frac{1}{\sqrt{2\pi t}} \int_{\mathbb{R}} f(y) e^{-\frac{|x-y|^2}{2t}} dy, \tag{83}$$

defines a strongly continuous semigroup on the space of bounded and uniformly continuous functions from $\mathbb{R}$ to $\mathbb{R}$. This is precisely the semigroup associated with Brownian motion through (82). In other words, because the transition probabilities of BM are given by (41), we can rewrite the expression (83) as

$$\mathcal{P}_t f(x) = \mathbb{E}[f(B_t)|B_0 = x]\,.$$

Given a semigroup, we can associate to it an operator, the *generator* of the semigroup.

Definition 9.5. *Given a $\mathcal{C}_0$-semigroup, P_t, we define the* infinitesimal generator *of the semigroup $\mathcal{P}_t$ to be the operator*

$$\mathcal{L}f := \lim_{t\to 0^+} \frac{\mathcal{P}_t f - f}{t}, \tag{84}$$

for all $f \in \mathcal{D}(\mathcal{L}) := \{f \in \mathfrak{B} :$ the limit on the RHs of (84) exists in $\mathfrak{B}\}$. If $\mathcal{P}_t$ is a strongly continuous Markov semigroup, then the operator $\mathcal{L}$ is a Markov generator*. The generator of the semigroup defined in (82), is often referred to as the* generator of the Markov process X_t*.*

From now on we will always work with strongly continuous semigroups, unless otherwise stated. The following properties hold.

Theorem 9.6. *With the notation and nomenclature introduced so far, let $\mathcal{P}_t$ be a strongly continuous semigroup. Then we have:*

1. *if $f \in \mathcal{D}(\mathcal{L})$ then also $\mathcal{P}_t f \in \mathcal{D}(\mathcal{L})$, for all $t \geq 0$;*
2. *the semigroup and its generator commute: $\mathcal{L}\mathcal{P}_t f = \mathcal{P}_t \mathcal{L} f$ for all $f \in \mathcal{D}(\mathcal{L})$;*
3. *$\partial_t(\mathcal{P}_t f) = \mathcal{L}\mathcal{P}_t f$.*

Notice that point 3. of Theorem 9.6 means that if we define a function $g_t(x) = g(t,x) := \mathcal{P}_t f(x)$ then such a function satisfies, for all $t \in \mathbb{R}_+$, the equation $\partial_t g = \mathcal{L} g$.

Proof of Theorem 9.6. 1. and 2. can be proved together: for any $t, s \geq 0$, we can write

$$\frac{\mathcal{P}_s \mathcal{P}_t f - \mathcal{P}_t f}{s} = \mathcal{P}_t \frac{\mathcal{P}_s f - f}{s}.$$

No we can take the limit as $s \to 0^+$ on both sides (the limit of the RHS makes sense and therefore also the one on the LHS does, which means that 1. is proved) and obtain

$$\mathcal{L}\mathcal{P}_t f = \lim_{s\to 0^+} \frac{\mathcal{P}_s \mathcal{P}_t f - \mathcal{P}_t f}{s} = \mathcal{P}_t \lim_{s\to 0^+} \frac{\mathcal{P}_s f - f}{s} = \mathcal{P}_t \mathcal{L} f\,.$$

Now we come to proving 3. In the remainder of this proof $h > 0$; let us look at the right limit first:

$$\lim_{h\to 0^+} \frac{\mathcal{P}_{t+h}f - \mathcal{P}_t f}{h} = \mathcal{P}_t \lim_{h\to 0^+} \frac{\mathcal{P}_h f - f}{h} = \mathcal{P}_t \mathcal{L} f \overset{\text{by 2.}}{=} \mathcal{L}\mathcal{P}_t f\,.$$

Therefore the right derivative is equal to $\mathcal{L}\mathcal{P}_t f$. Now we need to show the same for the left derivative as well:

$$\begin{aligned}\lim_{h\to 0}\left[\frac{\mathcal{P}_t f - \mathcal{P}_{t-h} f}{h} - \mathcal{L}\mathcal{P}_t f\right] &= \lim_{h\to 0} \mathcal{P}_{t-h}\frac{\mathcal{P}_h f - f}{h} - \mathcal{P}_t\mathcal{L} f \\ &\le \lim_{h\to 0} \mathcal{P}_{t-h}\left[\frac{\mathcal{P}_h f - f}{h} - \mathcal{L} f\right] \\ &\quad + \lim_{h\to 0}\left[\mathcal{P}_{t-h}\mathcal{L} f - \mathcal{P}_t\mathcal{L} f\right]\,.\end{aligned}$$

The second limit is equal to zero by strong continuity. Using Exercise 31, also the first limit is equal to zero, indeed

$$\lim_{h\to 0} \|\mathcal{P}_{t-h}\left[\frac{\mathcal{P}_h f - f}{h} - \mathcal{L} f\right]\| \le \lim_{h\to 0} c(t)\|\frac{\mathcal{P}_h f - f}{h} - \mathcal{L} f\|,$$

where $c(t)$ is a constant depending on t, not on h. Now the limit on the RHS of the above is clearly equal to zero, by the definition of $\mathcal{L}$. □

The following very famous result gives a necessary and sufficient condition in order for an operator to be the generator of a Markov semigroup.

Theorem 9.7 (Hille-Yosida Theorem for Markov semigroups)**.** *A linear operator $\mathcal{L}$ is the generator of a Markov semigroup $\{\mathcal{P}_t\}_{t\in\mathbb{R}_+}$ on the Banach space $\mathfrak{B}$ if and only if*

1. *$1 \in \mathcal{D}(\mathcal{L})$ and $\mathcal{L}1 = 0$;*
2. *$\mathcal{D}(\mathcal{L})$ is dense in $\mathfrak{B}$;*
3. *$\mathcal{L}$ is closed;*
4. *for any $\lambda > 0$, the operator $\lambda I - \mathcal{L}$ is invertible; its inverse is bounded,*

$$\sup_{\|f\|\le 1} \|(\lambda I - \mathcal{L})^{-1} f\| \le \lambda^{-1},$$

and positivity preserving.

Example 9.8. *[Generator of Brownian motion]*From the discussion of Section 7 we expect the generator of BM, i.e. the generator of the semigroup (83), to be $\Delta/2$, where Δ denotes the Laplacian. We work in one dimension so we expect to obtain $\partial^2_{xx}/2$. Indeed, let f be a bounded and continuous function, with bounded and continuous first and second derivatives. Then, if $\mathcal{P}_t$ denotes the Brownian semigroup (83), we have

$$\begin{aligned}\frac{\mathcal{P}_t f(x) - f(x)}{t} &= \frac{1}{t}\left[\frac{1}{\sqrt{2\pi t}}\int_{\mathbb{R}}(f(y)-f(x))e^{-\frac{|x-y|^2}{2t}}dy\right]\\ &= \frac{1}{t}\left[\frac{1}{\sqrt{2\pi t}}\int_{\mathbb{R}}(f(x+z)-f(x))e^{-z^2/2t}dz\right]\\ &= \frac{1}{t}\left[\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}}(f(x+w\sqrt{t})-f(x))e^{-w^2/2}dw\right]\\ &= \frac{1}{t}\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}}\left[f'(x)w\sqrt{t}+\frac{1}{2}f''(x+\theta w\sqrt{t})tw^2\right]e^{-w^2/2}dw,\end{aligned}$$

where the last inequality holds, by Taylor's Theorem, for some $\theta \in [0,1]$. From the continuity of f'' and observing that

$$\int_{\mathbb{R}} f'(x)w\,e^{-w^2/2}dw = f'(x)\int_{\mathbb{R}} w\,e^{-w^2/2}dw = 0,$$

we have

$$\lim_{t\to 0^+}\frac{\mathcal{P}_t f(x)-f(x)}{t} = \frac{f''(x)}{2}.$$

Definition 9.9. *Given a time-homogeneous Markov process $X_t : (\Omega, \mathcal{F}, \mathbb{P}) \to \mathbb{R}$, we have defined the associated Markov semigroup $\mathcal{P}_t$ by* (82)*. Denoting by $p_t(x, dy)$ the transition functions of the process, we can also define the* dual semigroup *associated to X_t and acting on the space of probability measures on $\mathbb{R}$ (denoted by $\mathcal{M}_1(\mathbb{R})$):*

$$(\mathcal{P}^*_t\mu)(B) = \int_{\mathbb{R}} p_t(x,B)\mu(dx), \quad t \geq 0, B \in \mathcal{B}(\mathbb{R}), \mu \in \mathcal{M}_1(\mathbb{R}). \tag{85}$$

Remark 9.10. We recall that the dual of $C([0,T],\mathbb{R})$ is the space of measures μ on $[0,T]$ with bounded total variation on $[0,T]$ (see [80], page 119); the total variation on $[0,T]$ is defined as

$$\|\mu\|_{TV} := \sup_{\substack{f\in C([0,T])\\ \|f\|\leq 1}}\int_{[0,T]} f(s)\mu(ds).$$

Also, the dual of the space $L^\infty(S, \mathcal{S}, \lambda)$, where $(S, \mathcal{S}, \lambda)$ is a measure space with $\lambda(S) < \infty$, is the space of finitely additive measures μ absolutely continuous with respect to λ and such that $\sup_B |\mu(B)| < \infty$ (see again same reference).

Notice that the Markov semigroup $\mathcal{P}_t$ acts on functions, its dual acts on probability measures. Moreover, $\mathcal{P}_t f$ is a function, while $\mathcal{P}_t^* \mu$ is a probability measure on $\mathbb{R}$. (Useless to say that in all of the above $\mathbb{R}$ can be replaced by any Polish space S and everything works anyway.) The reason why $\mathcal{P}_t^*$ is called the dual semigroup is the following: with Remark 9.10 in mind [26], if we use the notation

$$\langle \mathcal{P}_t f, \mu \rangle = \int_{\mathbb{R}} (\mathcal{P}_t f)(x) \mu(dx),$$

then for any say bounded and measurable function f and for any probability measure μ we have

$$\begin{aligned} \langle \mathcal{P}_t f, \mu \rangle &= \int_{\mathbb{R}} \left[\int_{\mathbb{R}} p_t(x, dy) f(y) \right] \mu(dx) \\ &= \int_{\mathbb{R}} f(y) \int_{\mathbb{R}} p_t(x, dy) \mu(dx) \\ &= \int_{\mathbb{R}} f(y) (\mathcal{P}_t^* \mu)(dy) = \langle f, \mathcal{P}_t^* \mu \rangle. \end{aligned}$$

So, at least formally, $\mathcal{P}_t^*$ is the dual (or adjoint) operator of $\mathcal{P}_t$.

9.1 Ergodicity for continuous time Markov processes.

Definition 9.11. *Let $\mathcal{P}_t$ be a Markov semigroup. A probability measure μ is said to be invariant for $\mathcal{P}_t$ if for any bounded and measurable function φ,*

$$\int_{\mathbb{R}} (\mathcal{P}_t \varphi)(x) \mu(dx) = \int_{\mathbb{R}} \varphi(x) \mu(dx). \tag{86}$$

Given a time-homogeneous Markov process, we say that a measure μ is invariant for X_t if it is invariant for the associated Markov semigroup.

[26] In view of Remark 9.10, this is just a duality relation, i.e. it expresses the action of the dual of a space on the space itself, see the Appendix; in the case of a Hilbert space this would be just a scalar product, which is why we denote duality relations and scalar products in the same way.

With the notation introduced at the end of the previous section, (86) can be rewritten as

$$\langle \mathcal{P}_t \varphi, \mu \rangle = \langle \varphi, \mu \rangle. \tag{87}$$

Therefore, by the point of view of the dual (or adjoint) semigroup, we can also say that the measure μ is invariant if

$$\mathcal{P}_t^* \mu = \mu, \quad \forall t \geq 0.^{27} \tag{88}$$

Comment. I suppose that you would like to reconcile Definition 9.11 with Definition 3.21. If $\mathcal{P}_t$ is the Markov semigroup associated with a given process X_t and we take $\varphi(x) = \mathbf{1}_B(x)$, for some measurable set B, then

$$\begin{aligned} \int \mathcal{P}_t \varphi(x) \mu(dx) &= \int \left(\int \mathbf{1}_B(y) p_t(x, y) dy \right) \mu(dx) \\ &= \int p_t(x, B) \mu(dx). \end{aligned}$$

Therefore equality (86) implies

$$\int p_t(x, B) \mu(dx) = \int \mathbf{1}_B(x) d\mu(x) = \mu(B).$$

Observe, as we have done for the time discrete case, that if μ is invariant for X_t and $X_0 \sim \mu$, then X_t is stationary, indeed in this case

$$\mathbb{P}_\mu(X_t \in B) = \int_{\mathbb{R}} p_t(x, B) \mu(dx) = \mathcal{P}_t^* \mu(B) = \mu(B). \tag{89}$$

Comment. The stationarity of μ is used in (89) only for the last equality. All the other equalities are true simply by definition of $\mathcal{P}_t^*$. This means that if we start the process with a certain distribution ν, then *the semigroup $\mathcal{P}_t^*$ describes the evolution of the law of the process*:

$$(\mathcal{P}_t^* \nu)(B) = \mathbb{P}_\nu(X_t \in B),$$

for every measurable set B. On the other hand, from the definition (82) of the semigroup $\mathcal{P}_t$, it is clear that $\mathcal{P}_t$ *describes the evolution of the* so called *observables*.

Again analogously to the time-discrete case, we give the following definition.

[27] If (87) holds for all φ buonded and measurable then take φ to be the indicator function of a measurable set and you obtain (88). On the other hand, if (88) is true then $\langle \mathcal{P}_t \varphi, \mu \rangle = \langle \varphi, \mathcal{P}_t^* \mu \rangle = \langle \varphi, \mu \rangle$, for all φ bounded and measurable.

Definition 9.12. *A time-homogeneous (continuous time) Markov process is* ergodic *if the associate semigroup (i.e. the semigroup defined through the relation* (82)*) admits a unique invariant measure.*

Before making the comment below, we would like to point out a technical fact.

Lemma 9.13. *Let $\mathcal{P}_t$ be a Markov semigroup associated with some real valued process X_t. If μ is an invariant measure for the semigroup, then $\mathcal{P}_t$ can be extended to a strongly continuous semigroup on $L^p(\mathbb{R}, \mu)$, for every $p \geq 1$.*

Notation: if S is any Polish space, $L^p(S, \mu)$ is a *weighted L^p space* (of $\mathbb{R}$ or $\mathbb{R}^n$-valued functions) over S, that is

$$L^p(S, \mu) := \{\text{functions } f : S \to \mathbb{R} \text{ such that } \int_S |f|^p (x) d\mu(x) < \infty\}.$$

From now on, we shall refer to the non-weighted L^p space as to the *flat L^p*. Thanks to the above Lemma, we can work in a Hilbert space setting as soon as an invariant measure exists.

Comment. One can prove that the set of all invariant probability measures for $\mathcal{P}_t$ is convex and an invariant probability measure is *ergodic* if and only if it is an extremal point of such set (compare with Lemma 5.4). Hence, if a semigroup has a unique invariant measure that measure is ergodic. This justifies the definition of ergodic process that we have just given. It can be shown (see [13]) that, given a C_0-Markov semigroup, the following statements are equivalent:

(i) μ is ergodic.

(ii) $\varphi \in L^2(S, \mu)$ and $\mathcal{P}_t \varphi = \varphi \quad \forall t > 0$, μ a.s. $\Rightarrow \varphi$ is a constant (μ a.s.).

(iii) for any $\varphi \in L^2(S, \mu)$,

$$\lim_{t \to \infty} \frac{1}{t} \int_0^t (\mathcal{P}_s \varphi)(x)\, ds = \int_S \varphi(x) \mu(dx), \quad \text{in } L^2(S, \mu).$$

Because the Markov semigroup preserves constants, condition (ii), read at the level of the generator of the semigroup, says that 0 is a simple eigenvalue of $\mathcal{L}$, if and only if $\mathcal{P}_t \varphi = \varphi \quad \forall t > 0$, μ a.s. $\Rightarrow \varphi$ is a constant. In other words, $\mathcal{L}\varphi = 0 \Leftrightarrow \varphi$ is constant (for $\varphi \in L^2(S, \mu) \cap \mathcal{D}(\mathcal{L})$). As for (iii),

this is precisely the ergodic theorem for continuous time Markov processes. Notice that the time average on the LHS depends on x (the initial value of the process) while the RHS does not: again, the limiting behaviour does not depend on the initial conditions.

Remark 9.14. When the semigroup has an invariant measure μ, we shall always assume that the generator of such a semigroup is densely defined in $L^2(\mathbb{R}, \mu)$. This assumption is not always trivial to check in practical cases.

For the moment only formally, the generator of $\mathcal{P}_t^*$ is just $\mathcal{L}^*$, the flat L^2-adjoint of the generator of $\mathcal{P}_t$, $\mathcal{L}$. From (88), a measure is invariant if

$$\mathcal{L}^* \mu = 0. \tag{90}$$

If $\mathcal{P}_t$ is the semigroup associated with some time homogeneous Markov process then the process is ergodic if equation (90) admits a unique (normalized) solution. (Notice indeed that $\mathcal{L}$ and $\mathcal{L}^*$ will be differential operators, so the solution to (90) cannot be unique.)

10 Diffusion Processes

For the material contained in this section we refer the reader to [1, 25, 22, 41].

In this section we want to give a mathematical description of the physical phenomenon called *diffusion*. Simply put, we want to mathematically describe the way milk diffuses into coffee.

It seems only right and proper to start this section by reporting the way in which Maxwell described the process of diffusion in his *Encyclopedia Britannica* article:

"*When two fluids are capable of being mixed, they cannot remain in equilibrium with each other; if they are placed in contact with each other the process of mixture begins of itself, and goes on till the state of equilibrium is attained, which, in the case of fluids which mix in all proportions, is a state of uniform mixture. This process of mixture is called diffusion. It may be easily observed by taking a glass jar half full of water and pouring a strong solution of a coloured salt, such as sulphate of copper, through a long-stemmed funnel, so as to occupy the lower part of the jar. If the jar is not disturbed we may trace the process of diffusion for weeks, months, or years, by the gradual rise of the colour into the upper part of the jar, and the weakening of the colour in the lower part. This, however, is not a method capable of giving accurate measurements of the composition of the liquid at different depths in the vessel. ...*

If we observe the process of diffusion with our most powerful microscopes, we cannot follow the motion of any individual portions of the fluids. We cannot point out one place in which the lower fluid is ascending, and another in which the upper fluid is descending. There are no currents visible to us, and the motion of the material substances goes on as imperceptibly as the conduction of heat or electricity. Hence the motion which constitutes diffusion must be distinguished from those motions of fluids which we can trace by means of floating motes. It may be described as a motion of the fluids, not in mass but by molecules. ..."

By a phenomenological point of view, the word diffusion denotes a transport phenomenon which happens without bulk motion and results in "mixing" or "spreading". [28] The two typical examples that you should bear in mind: the way milk diffuses into coffee and the way gas molecules confined to one side of a box divided in two compartments spread to the whole container as soon as the division is lifted. In the first case, if you wait long enough, you will obtain a well stirred coffee with milk; in the second case you will see that the gas molecules reach an equilibrium state described by the Gibbs distribution. [29]

10.1 Definition of diffusion process

Athough we will mostly deal with time-homogeneous examples, we give the definition below for the general context of non necessarily time-homogeneous Markov processes. We therefore use the notation introduced in Exercise 21.

Definition 10.1. *A real valued, continuous time Markov process,* $\{X_t\}_{t\in[0,T]}$*, is a* diffusion process *if its transition probabilities* $p(s,x,t,A)$*, satisfy the following conditions:*

1. *For any* $\epsilon > 0$*,*

$$\lim_{\delta\to 0}\frac{1}{\delta}\int_{|x-y|>\epsilon} p(t,x,t+\delta,dy) = 0, \quad \textit{for all } t\in[0,T] \textit{ and } x\in\mathbb{R};$$

2. *There exists a real valued function* $b(t,x)$ *such that for all* $\epsilon > 0$

$$\frac{1}{\delta}\lim_{\delta\to 0}\int_{|x-y|\le\epsilon}(y-x)p(t,x,t+\delta,dy) = b(t,x), \quad \textit{for all } t\in[0,T] \textit{ and } x\in\mathbb{R};$$

[28] Indeed, the word diffusion comes from the latin verb "diffundere", which means "to spread out".

[29] At this point one should mention the apparent paradox created by the Poincaré recurrence Theorem, but we will refrain from getting into the details of this long diatribe.

3. *There exists a real valued function* $D(t,x)$ *such that for all* $\epsilon > 0$

$$\frac{1}{\delta}\lim_{\delta\to 0}\int_{|x-y|\le\epsilon}(y-x)^2 p(t,x,t+\delta,dy) = D(t,x), \quad \textit{for all } t\in[0,T] \textit{ and } x\in\mathbb{R}.$$

The function $b(t,x)$ *is often called the* drift *while the function* $D(t,x)$ *is the* diffusion coefficient.

Before explaining the meaning of Definition 10.1, let us Remark that the same definition can be given for $\mathbb{R}^n$ valued processes.

Remark 10.2. For ease of notation, Definition 10.1 has been stated for one-dimensional processes. The exact same definition carries to the case in which X_t takes values in $\mathbb{R}^n$, $n \ge 1$. In this case $b(t,x)$ will be a $\mathbb{R}^n$-valued function and $D(t,x)$ will be a $n\times n$ matrix such that for any $\epsilon > 0$,

$$\lim_{\delta\to 0}\int_{|x-y|\le\epsilon}(y-x)(y-x)^T p(t,x,t+\delta,dy) = D(t,x),$$

for all t and x. In the above $(y-x)$ is a column vector and $(y-x)^T$ denotes its transpose.

Now let us see what the three conditions of Definition 10.1 actually mean. Condition 1. simply says that diffusion processes do not jump. In conditions 2. and 3. we had to cut the space domain to $|x-y| < \epsilon$ because strictly speaking at the moment we don't know yet whether the transition probabilities of the process have first and second moment or not. Assuming they do, then conditions 2. and 3. say, roughly speaking, that the displacement of the process from time t to time $t+\delta$ is the sum of two terms: an average drift term, $b(t,x)$, plus random effects, say η, plus small terms, $o(\delta)$; the random effects are such that $\mathbb{E}(\eta)^2 = D(t,x)\delta + o(\delta)$.[30]

Lemma 10.3. *Suppose a real valued Markov process* X_t *on* $[0,T]$ *with transition probabilities* $p(s,x,t,A)$ *satisfies the following three conditions:*

1. *there exists* $a > 0$ *such that*

$$\lim_{\delta\to 0}\frac{1}{\delta}\int_{\mathbb{R}}|y-x|^{2+a}\,p(t,x,t+\delta,dy) = 0, \quad \textit{for all } t\in[0,T] \textit{ and } x\in\mathbb{R};$$

[30]Here and in the following the notation $o(\delta)$ denotes a term, say $f(\delta)$, which is small in δ in the sense that

$$\lim_{\delta\to 0}\frac{f(\delta)}{\delta} = 0.$$

2. *there exists a function* $b(t,x)$ *such that*

$$\lim_{\delta\to 0}\frac{1}{\delta}\int_{\mathbb{R}}(y-x)p(t,x,t+\delta,dy)=b(t,x),\quad \textit{for all } t\in[0,T] \textit{ and } x\in\mathbb{R};$$

3. *there exists a function* $D(t,x)$ *such that*

$$\lim_{\delta\to 0}\frac{1}{\delta}\int_{\mathbb{R}}(y-x)^2p(t,x,t+\delta,dy)=D(t,x),\quad \textit{for all } t\in[0,T] \textit{ and } x\in\mathbb{R}.$$

Then X_t *is a diffusion process.*

Proof. We want to show that if the above three conditions hold then the conditions of Definition 10.1 are satisfied. Let us start with checking that under our assumptions, condition 1. of Definition 10.1 holds: for any $\epsilon>0$, we have

$$\frac{1}{\delta}\int_{|y-x|>\epsilon}p(t,x,t+\delta,dy)\le\frac{1}{\delta}\int_{\mathbb{R}}\frac{|y-x|^{2+a}}{\epsilon^{2+a}}p(t,x,t+\delta,dy)\overset{\delta\to 0}{\longrightarrow}0,$$

having used the fact that $|y-x|>\epsilon\Rightarrow(|y-x|/\epsilon)>1$. To verify that condition 2. of Definition 10.1 holds, observe that for every $\epsilon>0$,

$$\begin{aligned}\frac{1}{\delta}\int_{\mathbb{R}}(y-x)p(t,x,t+\delta,dy)&=\frac{1}{\delta}\int_{|y-x|\le\epsilon}(y-x)p(t,x,t+\delta,dy)\\&\quad+\frac{1}{\delta}\int_{|y-x|\ge\epsilon}(y-x)p(t,x,t+\delta,dy).\end{aligned}$$

If we can prove that the second addend on the RHS of the above converges to 0 as $\delta\to 0$, then we are done by assumption 2. Using again $|y-x|>\epsilon\Rightarrow(|y-x|/\epsilon)>1$, we act as before and we get

$$\left|\frac{1}{\delta}\int_{|y-x|\ge\epsilon}(y-x)p(t,x,t+\delta,dy)\right|\le\frac{1}{\delta}\int_{\mathbb{R}}\frac{|y-x|^{2+a}}{\epsilon^{1+a}}p(t,x,t+\delta,dy)\overset{\delta\to 0}{\longrightarrow}0.$$

You can act analogously to show that also condition 3. of Definition 10.1 is satisfied. □

What we want to do now is showing that, under appropriate conditions on the involved coefficients, solutions of SDEs are diffusion processes. Viceversa, a diffusion process solves, under certain conditions, an appropriate SDE. Unless otherwise stated, everything we say in the following is referred to real valued processes. The multidimensional case will be commented on in separate remarks.

Theorem 10.4. *Consider the SDE*

$$dX_u = b(u,x)du + \sigma(u,x)dW_u, \quad X_t = x. \tag{91}$$

Suppose the coefficients $b(u,x)$ and $\sigma(u,x)$ are continuous in both arguments and satisfy all the assumptions of the existence and uniqueness theorem, Theorem 8.9. Then the process X_u, solution of (91), is a diffusion process with diffusion coefficient $D(u,x) = \sigma^2(u,x)$.

Before proving the above statement, let us make the following remark.

Remark 10.5. *Under the conditions of Theorem 8.9 one can prove that the solution of equation (91) is still a Markov process and the following estimate holds: there exists a constant $C > 0$ such that*

$$\mathbb{E}\,|X_u - x|^{2m} \le Cu^m e^{Cu}(|x|^{2m} + 1), \quad \textit{for all } m \ge 1\,.$$

Proof of Theorem 10.4. We already know that X_u is a Markov process. All we need to verify are the conditions listed in Definition 10.1. To this end, we shall show that the three conditions of Lemma 10.3 are satisfied.

1. The estimate of Remark 10.5, with $m = 2$, gives

$$\mathbb{E}\,|X(t+\delta) - x|^4 = \int_{\mathbb{R}} |x-y|^4\, p(t,x,t+\delta,dy) \le K\delta^2(1+|x|^4).$$

   Dividing by δ and letting δ go to zero, condition 1. is verified.

2. We want to prove that

$$\frac{\mathbb{E}[X_{t+\delta} - x]}{\delta} = \frac{1}{\delta}\int_{\mathbb{R}} (y-x)p(t,x,t+\delta,dy) \overset{\delta\to 0}{\longrightarrow} b(t,x).$$

   Using (91), we can write

$$X_{t+\delta} = x + \int_t^{t+\delta} b(s,X_s)ds + \int_t^{t+\delta} \sigma(s,X_s)dW_s\,.$$

   Therefore

$$\frac{\mathbb{E}[X_{t+\delta} - x]}{\delta} = \frac{1}{\delta}\mathbb{E}\int_t^{t+\delta} b(s,X_s)ds\,.$$

   Tha change of variables $s = t + u\delta$ then gives

$$\frac{\mathbb{E}[X_{t+\delta} - x]}{\delta} = \mathbb{E}\int_0^1 b(t+u\delta, X_{t+u\delta})du \longrightarrow b(t,x).$$

In the above, we could exchange the limit and the integral because by assumption

$$|b(t+u\delta, X_{t+u\delta})|^2 \leq C(1+|X_{t+u\delta}|^2)$$

and we know by the existence and uniqueness theorem that

$$\mathbb{E}\int_0^1 |b(t+u\delta, X_{t+u\delta})|^2 \leq C\int_0^1 (1+\mathbb{E}\,|X_{t+u\delta}|^2) < \infty.$$

3. We are left with showing that

$$\mathbb{E}\frac{(X_{t+\delta}-x)^2}{\delta} = \frac{1}{\delta}\int_{\mathbb{R}} (y-x)^2 p(t,x,t+\delta,dy) \xrightarrow{\delta\to 0} D(t,x)\,.$$

To this end, let us write

$$\mathbb{E}(X_{t+\delta}-x)^2 = \mathbb{E}(X_{t+\delta})^2 - x^2 - 2x\left[\mathbb{E}X_{t+\delta} - x\right]\,. \tag{92}$$

Using Itô formula to calculate $d(X_u)^2$, we get

$$d(X_u)^2 = 2X_u dX_u + \sigma^2(u, X_u)du\,.$$

Integrating between t and $t+\delta$ and taking expectation then gives

$$\mathbb{E}(X_{t+\delta})^2 - x^2 = \mathbb{E}\int_t^{t+\delta} 2X_s b(s, X_s)ds + \mathbb{E}\int_t^{t+\delta} \sigma^2(s, X_s)ds\,.$$

Combining with (92):

$$\begin{aligned}\mathbb{E}(X_{t+\delta}-x)^2 = \mathbb{E}\int_t^{t+\delta} \left[2X_s b(s,X_s) + \sigma^2(s,X_s)\right] ds \\ - 2x(b(t,x)\delta + o(\delta))\,.\end{aligned}$$

After dividing both sides by δ, we can, as before, make the change of variables $s = t+u\delta$, exchange limit and integral (which is justified with an argument similar to the one explained above) and conclude by sending δ to zero.

□

Now the reverse.

Theorem 10.6. *Let X_t be a diffusion process on $[0,T]$. Suppose the coefficients of the diffusion $b(t,x)$ and $D(t,x)$ satisfy the following conditions:*

- *$b(t,x)$ is continuous in both arguments and it grows linearly, i.e. there exists a constant $C>0$ such that*

$$|b(t,x)| \leq C(1+|x|);$$

- *$D(t,x)$ is continuous in both arguments and it has bounded an continuous first derivatives (i.e. the derivatives $\partial_t D$ and $\partial_x D$ exist and are bounded and continuous);*

- *$(D(t,x))^{-1}$ is bounded;*

- *there exists a function $\psi(x)$, independent of t and δ, such that*

  a) *$\psi(x) > 1+|x|$ and $\sup_{t\in[0,T]} \mathbb{E}(\psi(X_t)) < \infty$;*

  b) *$\left|\int (y-x)p(t,x,t+\delta,dy)\right|$ and $\int (y-x)^2 p(t,x,t+\delta,dy)$ are bounded above by $\psi(x)\delta$;*

  c) *$\int (|y|+y^2)p(t,x,t+\delta,dy)$ is bounded above by $\psi(x)$.*

I will not prove the above theorem. However, let me give you some heuristics to understand intuitively why, loosely speaking, a diffusion process satisfies an SDE. To this end, let X_t be a diffusion process. Then conditions 2. and 3. of Definition 10.1 imply

$$\begin{aligned}
&\mathbb{E}(X_t - X_s | X_s = x) = b(s,x)(t-s) + o(t-s)\\
&\mathbb{E}((X_t - X_s)^2 | X_s = x) = D(s,x)(t-s) + o(t-s)\,.
\end{aligned}$$

Therefore, if we take a function $\sigma(t,x)$ such that $\sigma^2(t,x) = D(t,x)$, recalling that $\mathbb{E}(W_t - W_s)^2 = (t-s)$, we can write

$$X_t - X_s = b(s,x)(t-s) + \sigma(s,x)(W_t - W_s) + o(t-s),$$

which, forgetting about the $o(t-s)$ terms, can be rewritten in differential form to be precisely equation (91).

10.2 Backward Kolmogorov and Fokker-Planck equations

For the moment we will keep working in one dimension. At the end we shall comment on the multi-dimensional extension of the following results.

Lemma 10.7. *Let $f(x)$ be twice differentiable and suppose there exist $m, C > 0$ such that*

$$|f(x)| + \left|\frac{d}{dx}f(x)\right| + \left|\frac{d^2}{dx^2}f(x)\right| \leq C(1+|x|^m), \quad x \in \mathbb{R}.$$

Assume also that $b(t,x)$ and $D(t,x)$ satisfy the assumptions of Theorem 10.4. Then

$$\lim_{\delta\to 0}\frac{\mathbb{E}f(X_t)-f(x)}{\delta}=b(t,x)\frac{d}{dx}f(x)+\frac{1}{2}\sigma^2(t,x)\frac{d^2}{dx^2}f(x),$$

where X_u is the solution of

$$dX_u=b(u,X_u)du+\sigma(u,X_u)dW_u,\quad X_{t-\delta}=x\,. \tag{93}$$

Proof. The proof of this lemma is left as an exercise, see Exercise 37. □

Theorem 10.8. *Let X_u be the solution of the SDE*

$$dX_u=b(u,X_u)du+\sigma(u,X_u)dW_u,\quad X_t=x\,, \tag{94}$$

where we assume that the coefficients of the SDE are continuous in both arguments with continuous partial derivatives $\partial_x b(u,x)$, $\partial_x\sigma(u,x)$, $\partial_{xx}b(u,x)$, $\partial_{xx}\sigma(u,x)$. Suppose $f(x)$ is a twice continuously differentiable function and that there exist $K,m>0$ such that

$$\begin{aligned}
&|b(u,x)|+|\sigma(u,x)|\le K(1+|x|)\,,\\
&|\partial_x b(u,x)|+|\partial_{xx}b(u,x)|+|\partial_x\sigma(u,x)|+|\partial_{xx}\sigma(u,x)|\le K(1+|x|^m)\,,\\
&|f(x)|+\left|\frac{d}{dx}f(x)\right|+\left|\frac{d^2}{dx^2}f(x)\right|\le K(1+|x|^m)\,.
\end{aligned}$$

Then the function $h(t,x)=\mathbb{E}\left[f(X_u)|X_t=x\right]$ satisfies the equation

$$\frac{\partial h(t,x)}{\partial t}+b(t,x)\frac{\partial h(t,x)}{\partial x}+\frac{1}{2}\sigma^2(t,x)\frac{\partial^2 h(t,x)}{\partial x^2}=0,\quad t\in(0,u) \tag{95}$$

$$\lim_{t\to u}h(t,x)=f(x)\,. \tag{96}$$

Remark 10.9. Equation (95) is called the *Backward Kolmogorov equation* (BK). The name of the equation comes from the fact that it is an equation for the "backward" variables t and x. Moreover, notice that $h(t,x)$ is the solution of a final value problem.

Comment. Let $\mathcal{L}_t$[31] be the differential operator operator

$$\mathcal{L}_t:=b(t,x)\frac{\partial}{\partial x}+\frac{1}{2}\sigma^2(t,x)\frac{\partial^2}{\partial x^2}\,.$$

[31] The notation $\mathcal{L}_t$ rather than just $\mathcal{L}$ is to emphasize that the coefficients of this operator do depend on time.

$\mathcal{L}_t$ is called the *backward operator* or also the *generator* of the (non time-homogeneous) diffusion (94). Theorem 10.8 says that the evolution equation

$$\begin{aligned}\frac{\partial h(t,x)}{\partial t} &= -\mathcal{L}_t h(t,x)\\ \lim_{t\to u} u(t,x) &= f(x)\end{aligned}$$

is a "backward" equation for the observables.

Proof of Theorem 10.8(sketch). I will not prove the differentiability properties of the function $h(t,x)$ but rather show formula (95). More details about this proof ca be found in [22, Section 11]. For the purposes of this proof we shall denote by $X^{x,t}(u)$ the solution at time u of equation (94) started in x at time t. With this notation we can simply write $h(t,x) = \mathbb{E}[f(X^{x,t}(u))]$. What we want to show is

$$\begin{aligned}\partial_t h(t,x) &= \lim_{\delta\to 0} \frac{h(t,x) - h(t-\delta,x)}{\delta}\\ &= -b(t,x)\partial_x h(t,x) - \frac{1}{2}\sigma^2 \partial_{xx} h(t,x)\,.\end{aligned}$$

Let us start with observing that

$$X^{x,t-\delta}(u) = X^{X^{x,t-\delta},t}(u).$$

Therefore

$$\begin{aligned}h(t-\delta,x) &= \mathbb{E}[\mathbb{E}(f(X^{x,t-\delta}(u))|X^{x,t-\delta}(t))]\\ &= \mathbb{E}h(t, X^{x,t-\delta}(t)).\end{aligned}$$

Using the above and Lemma 10.7 we get the result:

$$\begin{aligned}\lim_{\delta\to 0} \frac{h(t,x) - h(t-\delta,x)}{\delta} &= -\lim_{\delta\to 0} \frac{\mathbb{E}h(t, X^{x,t-\delta}(t)) - h(t,x)}{\delta}\\ &= -b(t,x)\partial_x h(t,x) - \frac{1}{2}\sigma^2 \partial_{xx} h(t,x).\end{aligned}$$

□

Theorem 10.10. *Let X_t, $t \in [0,T]$ be a diffusion process with coefficients $b(t,x)$ and $D(t,x)$ such that the limits in Definition 10.1 hold uniformly*

for $t \in [0,T]$ and $x \in \mathbb{R}$. Suppose the transition probabilities of X_t have a density [32] *$p(s,x,t,y)$ for every $t > s$. If the partial derivatives*

$$\frac{\partial p}{\partial t},\ \frac{\partial}{\partial y}(b(t,y)p),\ \frac{\partial^2}{\partial y^2}(D(t,y)p)$$

exist and are continuous, then the transition density $p(s,x,t,y)$ solves the equation

$$\frac{\partial p}{\partial t} + \frac{\partial}{\partial y}(b(t,y)p) - \frac{1}{2}\frac{\partial^2}{\partial y^2}(D(t,y)p) = 0 \tag{97}$$

$$\lim_{t\to s} p(s,x,t,y) = \delta(x-y)\,. \tag{98}$$

Equation (97) is a differential equation in the "forward variables" t and y and for this reason it is known as the *Fokker-Planck* or *forward Kolmogorov* equation.

10.2.1 Time-homogeneous case

Let us now specialize to the time-homogeneous case. In this case the coefficients of the SDE (94) do not depend on time. The process we are concerned with is then solution of the SDE

$$dX_u = b(X_u)du + \sigma(X_u)dW_u, \quad X_t = x\,, \tag{99}$$

where $b(x)$ and $\sigma(x)$ satisfy all the assumptions stated so far. In other words, the function $h(t,x) = \mathbb{E}[f(X_u)|X_t = x]$ becomes

$$h(t,x) = \mathbb{E}[f(X_u)|X_t = x] = \mathbb{E}[f(X_{u-t})|X_0 = x].$$

Setting $\tau = u - t$ and noticing that $\partial_t = -\partial_\tau$, equation (95) becomes now

$$\begin{aligned} &\frac{\partial h(\tau,x)}{\partial \tau} = \mathcal{L}h(\tau,x) \\ &\lim_{\tau\to 0} h(\tau,x) = f(x)\,, \end{aligned} \tag{100}$$

where the differential operator $\mathcal{L}$ is now the generator of the diffusion in the sense of Definition (84),

$$\mathcal{L} = b(x)\frac{\partial}{\partial x} + \frac{1}{2}\sigma^2(x)\frac{\partial^2}{\partial x^2}.$$ [33]

[32] See Remark 10.16 on this point.

[33] You could find this expression also by acting analogously to what we have done in Example 9.8.

Because of (100), it is customary to write

$$h(\tau, x) = (\mathcal{P}_\tau f)(x) = e^{\tau \mathcal{L}}.$$

In this time-homogeneous setting, the Fokker-Planck equation becomes an equation for the transition probability density $p(t, x, y)$ (rather than $p(s, x, u, y)$) and it is precisely

$$\frac{\partial}{\partial t} p(t, x, y) = \mathcal{L}^* p(t, x, y), \quad \lim_{t \to 0} p(t, x, y) = \delta(x - y),$$

for every fixed $x \in \mathbb{R}$, where $\mathcal{L}^*$ is precisely the flat L^2-adjoint of the operator $\mathcal{L}$:

$$\mathcal{L}^* \cdot = -\frac{\partial}{\partial y}(b(y)\cdot) + \frac{1}{2}\frac{\partial^2}{\partial y^2}(\sigma^2(y)\cdot)\,.$$

Example 10.11. Consider the O-U process

$$dX_t = -\alpha X_t dt + \sigma dW_t, \quad \sigma, \alpha > 0.$$

This is a time-homogeneous process with generator

$$\mathcal{L} = -\alpha x \frac{\partial}{\partial x} + \frac{\sigma^2}{2}\frac{\partial^2}{\partial x^2}$$

and Fokker-Planck operator

$$\mathcal{L}^* \cdot = \alpha \frac{\partial}{\partial x}(x\cdot) + \frac{\sigma^2}{2}\frac{\partial^2}{\partial x^2}$$

The equation $\mathcal{L}^* \rho = 0$ has a unique normalized solution

$$\rho = \sqrt{\frac{\alpha}{2\pi D}} e^{-\alpha x^2 / 2D},$$

hence the O-U process is ergodic.

Example 10.12. Both the generator and the Fokker-Planck operator of Brownian motion are simply $\partial^2_{xx}/2$.

Comment. The Fokker-Planck equation is a continuity equation in the sense that it expresses the conservation of probability mass. Indeed, using the Fokker-Planck (FP) equation, it is straightforward to show

$$\frac{d}{dt}\int_{\mathbb{R}} p(t, x, y) dy = 0 \quad \text{for every fixed } x \in \mathbb{R}.$$

Therefore

$$\int_{\mathbb{R}} p(t, x, y) dy = \int_{\mathbb{R}} p(0, x, y) dy = \int \delta(x - y) = 1, \quad \text{for all } t > 0.$$

Remark 10.13. What happens in higher dimensions? Recalling Remark 10.2, this time we have a process $X_t \in \mathbb{R}^n$ solving the multidimensional SDE

$$dX_t = b(x)dt + \sigma(x)dW_t,$$

where $b(x)$ is a vector in $\mathbb{R}^n$ and $\sigma(x)$ is a matrix. X_t is a time-homogeneous diffusion with drift vector $b(x)$ and diffusion coefficient $D(x) = \sigma(x)\sigma^T(x)$. The generator of such a process is the operator

$$\mathcal{L} = \sum_{j=1}^{n} b^j(x)\frac{\partial}{\partial x^j} + \frac{1}{2}\sum_{i,j=1}^{n} D^{i,j}(x)\frac{\partial^2}{\partial x^j x^j}\,.$$

Notice that the diffusion matrix is always symmetric and positive definite.

Example 10.14. Consider the system

$$\begin{aligned} dY_t &= 3Y_t dt - mZ_t dt + \sqrt{2a}\, dW_t^1, \quad a, m > 0 \\ dZ_t &= -6Z_t dt + \sqrt{2} dW_t^2\,. \end{aligned}$$

The generator of the process is

$$\mathcal{L} = 3y\frac{\partial}{\partial y} - mz\frac{\partial}{\partial z} + 2a\frac{\partial^2}{\partial y^2} - 6z\frac{\partial}{\partial z} + \frac{\partial^2}{\partial z^2}.$$

Definition 10.15. *A second order differential operator L on $\mathbb{R}^n$ of the form*

$$L = \sum_{i=1}^{n} a^i(x)\frac{\partial}{\partial x^i} + \sum_{i,j=1}^{n} M^{ij}(x)\frac{\partial^2}{\partial x^i x^j},$$

where $a(x) = (a^i(x))_{1\le i\le n}$ is a $\mathbb{R}^n$ valued function on $\mathbb{R}^n$ and $M(x) = (M^{ij}(x))_{1\le i,j\le n}$ is a matrix valued function on $\mathbb{R}^n$ is said uniformly elliptic *if there exists a positive constant $\alpha > 0$ such that*

$$\sum_{i,j=1}^{n} M^{ij}(x)v^i v^j \ge \alpha\|v\|^2,$$

for all vectors $v \in \mathbb{R}^n$.

Remark 10.16. If the operator $\mathcal{L}^*$ is uniformly elliptic then the process X_t with generator $\mathcal{L}$ has a density. This is a consequence of the good smoothing properties of elliptic operators. Indeed if $\mathcal{L}^*$ is an elliptic operator then the Fokker-Plank equation $\partial_t p = \mathcal{L}^* p$ is a parabolic equation (just the heat

equation in the case of Brownian Motion, for example) and it is a standard result from the theory of PDEs that the solution to this kind of equation is smooth for every positive time, even if the initial datum is not. Also, the maximum principle for parabolic equations guarantees that if we start with a positive initial datum then the solution of the Fokker-Planck equation remains positive for every subsequent time, see [21].

10.3 Reversible diffusions and spectral gap inequality

Also in this section we refer to time-homogeneous processes. Let us start with an example.

Example 10.17. Consider the one dimensional process

$$dX_t = -V'(X_t)dt + \sqrt{2}dW_t\,, \tag{101}$$

where $V(x)$ is a smooth confining potential. The generator of the process is

$$\mathcal{L} = -V'(x)\frac{\partial}{\partial x} + \frac{\partial^2}{\partial x^2}$$

while the FP operator is

$$\mathcal{L}^*\cdot = \frac{\partial}{\partial x}\left(V'(x)\cdot\right) + \frac{\partial^2}{\partial x^2}\cdot\,,$$

and the equation $\mathcal{L}^*\rho = 0$ has a unique normalized solution $\rho = \rho(x)$:

$$\begin{aligned}\mathcal{L}^*\rho = 0 &\Leftrightarrow \frac{\partial}{\partial x}\left[V'(x)\rho(x) + \frac{\partial}{\partial x}\rho(x)\right] = 0\\ &\Leftrightarrow V'(x)\rho(x) + \frac{\partial}{\partial x}\rho(x) = const.\end{aligned}$$

Solving the above SDE gives $\rho(x) = e^{-V(x)}/\mathcal{Z}$, where $\mathcal{Z}$ is a normalization constant. Observe that when the potential is quadratic, this process is precisely the O-U process.

The semigroup generated by $\mathcal{L}$ can be extended to a strongly continuous semigroup on the weighted L^2_ρ (see Lemma 9.13). Therefore, by the Hille-Yosida Theorem, $\mathcal{L}$ is a closed operator. Moreover, $\mathcal{L}$ is a self-adjoint operator in L^2_ρ. Let us start with proving that $\mathcal{L}$ is symmetric. To this end, let us first recall that the scalar product in $\mathcal{L}^2_\rho$ is defined as

$$\langle f, g\rangle_\rho := \int_{\mathbb{R}} f(x)g(x)\,\rho(x)dx\,.$$

Let us now show that $\langle \mathcal{L}f, g\rangle_\rho = \langle f, \mathcal{L}g\rangle_\rho$, for every f, g ***smooth and in L^2_ρ:

$$\begin{aligned}
\langle \mathcal{L}f, g\rangle_\rho &= \int_{\mathbb{R}} (\mathcal{L}f) g\rho \, dx \\
&= \int_{\mathbb{R}} -V'(x) \left(\frac{d}{dx} f\right) g\rho \, dx + \int_{\mathbb{R}} \left(\frac{d^2}{dx^2} f\right) g\rho \, dx \\
&= -\int_{\mathbb{R}} \frac{df}{dx} \frac{dg}{dx} \rho \, dx \\
&= \int_{\mathbb{R}} f \frac{d^2 g}{dx^2} \rho \, dx - \int_{\mathbb{R}} f \frac{dg}{dx} V' \rho \, dx \\
&= \int_{\mathbb{R}} f(\mathcal{L}g)\rho \, dx = \langle f, \mathcal{L}g\rangle_\rho.
\end{aligned}$$

As a byproduct of the above calculation, we also have

$$\langle \mathcal{L}f, g\rangle_\rho = -\int_{\mathbb{R}} \frac{df}{dx} \frac{dg}{dx} \rho \, dx \,, \tag{102}$$

which will be useful in the following - notice also that (102) makes the symmetry property obvious. Because $\mathcal{L}$ is symmetric and closed [34], it is also self-adjoint in $\mathcal{L}^2_\rho$.

Let us now come to explain why diffusions generated by a self-adjoint operator are so important.

Definition 10.18. *A probability measure μ is* reversible *for a Markov semigroup $\mathcal{P}_t$ if for any $f, g \in \mathcal{B}_m$*

$$\int (\mathcal{P}_t f) g \, d\mu(x) = \int f(\mathcal{P}_t g) \, d\mu(x). \tag{103}$$

In this case it is also customary to say that μ satisfies the detailed balance condition *with respect to $\mathcal{P}_t$.*

Analogously to the time discrete case if $\mathcal{P}_t$ is the Markov semigroup associated to some Markov process X_t and (103) is satisfied, then the process X_t is time-reversible (for a proof of this fact in the time-continuous setting see [70], which is an excellent reference). A given Markov semigroup might have more than one reversible measure. We denote $\mathcal{R}(\mathcal{P}_t)$ the set of reversible measures for a given Markov semigroup $\mathcal{P}_t$. Taking $g \equiv 1$, it is obvious that

[34] and defined on a dense subset of $\mathcal{L}^2_\rho$, namely the set of Schwartz functions, see [64].

if μ is reversible for $\mathcal{P}_t$ then it is also an invariant measure for the semigroup. So if $\mathcal{P}_t$ admits a reversible measure μ, then the semigroup can be extended to a strongly continuous semigroup on L^2_μ. With the same reasoning as in Example 10.17, the generator is also closed. Moreover it can be shown that (103) is equivalent to

$$\langle \mathcal{L}f, g\rangle_\mu = \langle f, \mathcal{L}g\rangle_\mu, \quad \forall f, g^{***}\text{smooth enough and in } L^2_\mu. \tag{104}$$

Therefore $\mathcal{L}$ is self-adjoint. This whole reasoning proves the following.

Proposition 10.19. *Let $\mathcal{P}_t$ be a Markov semigroup associated with a Markov process X_t and μ a reversible measure for $\mathcal{P}_t$. Then X_t is time-reversible and the generator of $\mathcal{P}_t$ is self-adjoint. Conversely if $\mathcal{L}$ is the generator of a strongly continuous Markov semigroup on L^2_μ (for some measure μ) and $\mathcal{L}$ satisfies* (104)*, then μ is reversible for the semigroup and the associated process is reversible.*

This is the reason why diffusion processes with self-adjoint generator are also called *reversible diffusions*. The study of exponentially fast convergence to equilibrium for reversible diffusions has been extensively tackled in the literature, see [4, 5] and references therein. In what follows we will often use the notation

$$f_t(x) := (\mathcal{P}_t f)(x)$$

and work in one dimension but everything we say can be rephrased in higher dimensions.

Definition 10.20. *Given a Markov semigroup $\mathcal{P}_t$ *** with generator $\mathcal{L}$, we say that a measure $\mu \in \mathcal{R}(\mathcal{P}_t)$ satisfies a* spectral gap inequality *if there exists a constant $\alpha > 0$ such that*

$$\alpha \int_{\mathbb{R}} \left[f - \int_{\mathbb{R}} f d\mu \right]^2 d\mu \le \langle -\mathcal{L}f, f\rangle_\mu, \quad \textit{for every } f \in L^2_\mu \cap \mathcal{D}(\mathcal{L}). \tag{105}$$

The largest positive number α such that (105) *is satisfied is called the* spectral gap *of the self-adjoint operator $\mathcal{L}$.*

The therm on the RHS of (105) is called the *Dirichlet form* of the operator $\mathcal{L}$.

Remark 10.21. If $\mathcal{L}$ is a self adjoint operator then the form $\langle \mathcal{L}f, f\rangle_\mu$ is real valued (for every $f \in L^2_\mu \cap \mathcal{D}(\mathcal{L})$). In particular the spectrum of $\mathcal{L}$ is real. If $\mathcal{L}$ is the generator of a strongly continuous Markov semigroup and

the semigroup is ergodic then we already know that 0 is a simple eigenvalue of $\mathcal{L}$ (see comment after Lemma 9.13). In the case of Example 10.17, we also know from (102) that $\langle \mathcal{L}f, f\rangle_\mu \leq 0$ for every f, therefore the self-adjoint operator $\mathcal{L}$ is *negative* and all the eigenvalues of $\mathcal{L}$ will be negative. Clearly $-\mathcal{L}$ is positive and the smallest positive α such that (105) holds is the smallest nonzero eigenvalue of $-\mathcal{L}$ (i.e. $-\alpha$ is the biggest nonzero eigenvalue of $\mathcal{L}$). This is the reason why α is called the spectral gap. The next proposition clarifies why spectral gal inequalities are so important.

Proposition 10.22. *A measure $\mu \in \mathcal{R}(\mathcal{P}_t)$ satisfies a spectral gap inequality (with constant α) if and only if*

$$\int_{\mathbb{R}} \left(\mathcal{P}_t f - \int_{\mathbb{R}} f d\mu \right)^2 d\mu \leq e^{-2\alpha t} \int_{\mathbb{R}} \left(f - \int_{\mathbb{R}} f d\mu \right)^2 d\mu \,, \tag{106}$$

for all $t \geq 0$ and $f \in L^2_\mu$.

Proof. Observe first that if $\int f d\mu = 0$ then $\int f_t d\mu = 0$ as well, as μ is invariant. Also, $\mathcal{L}$ is a differential operator, so constant functions are in the kernel of $\mathcal{L}$. Therefore we can work with mean zero functions. We want to prove that (105) holds if and only if (106) does. We prove here the implication (105) $\Rightarrow$ (106), the other implication is left as an exercise, see Exercise 38. ***If we are working with mean zero functions showing that (105) $\Rightarrow$ (106) amounts to proving that

$$\alpha \int_{\mathbb{R}} f^2 d\mu \leq \langle -\mathcal{L}f, f\rangle_\mu \tag{107}$$

implies

$$\int_{\mathbb{R}} (\mathcal{P}_t f)^2 \, d\mu \leq e^{-2\alpha t} \int_{\mathbb{R}} f^2 d\mu$$

In order to do so, apply (107) to f_t and get

$$\begin{aligned} \alpha \int_{\mathbb{R}} f_t^2 d\mu \leq - \int_{\mathbb{R}} f_t (\mathcal{L} f_t) d\mu &= - \int_{\mathbb{R}} f_t \left(\frac{d}{dt} f_t \right) d\mu \\ &= -\frac{1}{2} \int_{\mathbb{R}} \left(\frac{d}{dt} f_t^2 \right) d\mu \,. \end{aligned}$$

Therefore

$$\frac{d}{dt} \int_{\mathbb{R}} \left(f_t^2 \right) d\mu \leq -2\alpha \int_{\mathbb{R}} f_t^2 d\mu.$$

Integrating the above inequality we get

$$\int_{\mathbb{R}} f_t^2 d\mu \leq e^{-2\alpha t} \int_{\mathbb{R}} f^2 \, d\mu.$$

□

Example 10.23 (I.e. Example 10.17 continued)**.** Going back to the process X_t solution of (101), we now have the tools to check whether X_t converges exponentially fast to equilibrium. Working again with mean zero functions (notice that in this case the kernel of the generator is made only of constants), the spectral gap inequality reduces to

$$\alpha \int_{\mathbb{R}} f^2 \rho \, dx \leq \int_{\mathbb{R}} \left| \frac{df}{dx} \right|^2 \rho \, dx. \tag{108}$$

Those of you who have taken a basic course in PDEs will have noticed that this is a *Poincaré Inequality* for the measure ρ. In Appendix B.4, I have recalled the basic Poincaré inequality that most of you will have already encountered. If the potential $V(x)$ is quadratic then the measure ρ does satisfy (108) and in this case we therefore have exponentially fast convergence to equilibrium. For general potentials a classic result is the following.

Lemma 10.24. *Let $V(x) : \mathbb{R}^n \to \mathbb{R}$ be a twice differentiable confining potential such that $\rho = e^{-V(x)}$ is a probability density. Denote by $H^1(\rho)$ the weighted H^1, with norm*

$$\|f\|_{H^1_\rho}^2 := \|f\|_{L^2_\rho}^2 + \|\nabla f\|_{L^2_\rho}^2 .$$

If $V(x)$ is such that

$$\frac{|\nabla V|^2}{2} - \Delta V(x) \overset{|x|\to\infty}{\longrightarrow} \infty,$$

then the measure ρ satisfies a Poincaré inequality: for all functions $h \in H^1(e^{-V(x)})$ and for some $K > 0$:

$$\int_{\mathbb{R}} |\nabla h|^2 \, e^{-V(x)} dx \geq K \left[\int_{\mathbb{R}} h^2 e^{-V(x)} dx - \left(\int_{\mathbb{R}} h e^{-V(x)} dx \right)^2 \right] .$$

Before concluding this section we would like to make a remark, which is useful in computational practice. Given a probability measure μ, there are many processes admitting μ as invariant measure. We will illustrate this fact using the process (101). By the point of view of MCMC, this observation is particularly important as one can then try and find the process that converges fastest to the measure that we want to sample from.

Example 10.25. Let $v(x)$ be a smooth function and consider the process

$$dX_t^v = -(v(X_t^v) + V'(X_t^v))dt + \sqrt{2}dW_t, \tag{109}$$

where the potential $V(x)$ satisfies the same assumptions as in (101). It is clear that we have perturbed the dynamics (101) through the use of the function $v(x)$ [35]. Recall that the unperturbed process has a unique invariant measure with density $\rho = e^{-V(x)}/\mathcal{Z}$. Now the question is: is it possible to find a function $v(x)$ such that X_t^v solution of (109) admits ρ as invariant measure? In other words, is it possible to perturb the drift of the process (101) without altering the invariant measure (and therefore the long time behaviour of the dynamics)? The answer turns out to be simple. The generator of (109) is

$$\mathcal{L}_v = -(v(x) + V'(x))\frac{\partial}{\partial x} + \frac{\partial^2}{\partial x^2}.$$

We know that ρ is the invariant measure of (101), therefore

$$\frac{\partial}{\partial x}(V'\rho) + \frac{\partial^2}{\partial x^2}\rho = 0. \tag{110}$$

If we want ρ to be also the invariant measure of (109) then we need to impose $\mathcal{L}_v^*\rho = 0$. This results in

$$\begin{aligned}\mathcal{L}_v^*\rho = 0 \Leftrightarrow \frac{\partial}{\partial x}\left[(v + V')\rho\right] + \frac{\partial^2}{\partial x^2}\rho = 0 \\ \overset{(110)}{\Leftrightarrow} \frac{\partial}{\partial x}(v\rho) = 0\,.\end{aligned}$$

In higher dimension, i.e. if we consider the process $X_t \in \mathbb{R}^d$ satisfying

$$dX_t = -\nabla V(X_t)dt + \sqrt{2}dW_t$$

where W_t is now d-dimensional standard Brownian motion, you can show again that the measure ρ with density $\rho(x) = e^{-V(x)}/\mathcal{Z}$ is the only invariant measure for X_t. In this case the invariant measure of the process

$$dX_t^v = -(v(X_t^v) + \nabla V(X_t^v))dt + \sqrt{2}dW_t,$$

is still ρ if and only if

$$\nabla \cdot (v\rho) = 0,$$

where $\nabla\cdot$ denotes divergence (see Exercise 39).

[35] Notice that the generator of the perturbed dynamics is no longer self-adjoint in L^2_ρ, where $\rho = e^{-V(x)}/\mathcal{Z}$

10.4 The Langevin equation as an example of Hypocoercive diffusion

The take-home message from Section 10.3 is: diffusions generated by operators which are elliptic and self-adjoint are usually the easiest to analize as, roughly speaking, ellipticity gives the existence of a C^∞ density for the invariant measure while the self-adjointedness of the operator gives nice spectral property which are key to proving exponential convergence to equilibrum. In this section we want to see what happens when the generator is non elliptic and non self-adjoint. We shall mainly focus on how to prove exponential convergence to equilibrium for non self-adjoint diffusions. *Nihil recte sine exemplo docetur*, so let us start with the main motivating example, the *Langevin equation*:

$$\begin{aligned} dq &= pdt \\ dp &= -\partial_q V(q)dt - pdt + \sqrt{2}dW_t, \end{aligned} \tag{111}$$

where $V(q)$ is a confinig potential and W_t is one dimensional standard Brownian motion. If we assume that the potential is locally Lipshitz then the following general result, which can be found for example in [72, Chapter 10], guarantees the existence of a strong solution for the system (111).

Theorem 10.26. *Let $b(t,x) : \mathbb{R}_+ \times \mathbb{R}^n \longrightarrow \mathbb{R}^n$ be Locally Lipshitz, uniformly for $t \in [0,T]$ for each $T > 0$ and suppose*

1. $\sup_{0\le t\le T} |b(t,0)| < \infty$*, for all $T > 0$;*
2. *there exists $k > 0$ such that $\langle x, b(t,x)\rangle \le 0$ if x is outside of a ball of radius k.*

Then the SDE

$$dX_t = b(t,X_t)dt + \sigma dW_t$$

admits a unique strong and nonexploding solution for any random initial datum X_0 and any constant $\sigma > 0$.

The generator of (111) is

$$\mathcal{L} = p\partial_q - \partial_q V(q)\partial_p - p\partial_p + \partial_p^2 \tag{112}$$

and the corresponding Fokker-Planck operator is

$$\mathcal{L}^* = -p\partial_q + \partial_q V(q)\partial_p + \partial_p(p\cdot) + \partial_p^2\,. \tag{113}$$

Showing that this process has a unique invariant measure and that such a measure has a density is not straightforward. However the process is indeed ergodic with invariant density

$$\rho(q,p) = \frac{e^{-(V(q)+p^2/2)}}{\mathcal{Z}}\,. \tag{114}$$

What is easy to see is that the kernel of the generator is only made of constants. In showing this fact we will also gain a better understanding of the system (111). The dynamics described by (111) can be thought of as split into a Hamiltonian component,

$$\begin{aligned} \dot{q} &= p \\ \dot{p} &= -\partial_q V(q) \end{aligned} \tag{115}$$

plus a O-U process (in the p variable):

$$\begin{aligned} dq &= p dt \\ dp &= -\partial_q V(q) dt \underbrace{-p dt + \sqrt{2} dW_t}_{\text{O-U process}}\,. \end{aligned}$$

Indeed the equations (115) are the equations of motion of an Hamiltonian system with Hamiltonian

$$H(q,p) = V(q) + \frac{p^2}{2}\,.$$

At the level of the generator this is all very clear:

$$\mathcal{L} = \mathcal{L}_H + \mathcal{L}_{OU},$$

where

$$\mathcal{L}_H := p\partial_q - \partial_q V(q)\partial_p \tag{116}$$

is the Liouville operator of classical Hamiltonian mechanics and

$$\mathcal{L}_{OU} := -p\partial_p + \partial_p^2$$

is the generator of a O-U process in the p variable. By the point of view of our formalism, the Hamiltonian dynamics (115) admits infinitely many invariant measures, indeed

$$\mathcal{L}_H f(H(q,p)) = 0 \quad \text{for every } f,$$

i.e. any function of the Hamiltonian is in the kernel of $\mathcal{L}^*$. So any integrable and normalized function of the Hamiltonian is an invariant probability measure for (115). Adding the O-U process amounts to selecting one equilibrum, indeed

$$Ker(L_{OU}) = \{\text{functions that are constant in } p\},$$

so the kernel of $\mathcal{L}$ is only made of constants.

To distinguish between the flat L^2 adjoint of an operator T and the adjoint in the weighted L^2_ρ, we shall denote the first by T^* and the latter by $T^\star$. Notice now that the generator $\mathcal{L}_H$ of the Hamiltonian part of the Langevin equation is antisymmetric both in L^2 and in L^2_ρ. It is indeed straightforward to see that

$$\mathcal{L}_H = -\mathcal{L}_H^*.$$

Also, $\langle \mathcal{L}_H f, g\rangle_\rho = -\langle f, \mathcal{L}_H g\rangle_\rho$ for every $f, g \in L^2_\rho \cap \mathcal{D}(\mathcal{L}_H)$:

$$\begin{aligned}
\langle \mathcal{L}_H f, g\rangle_\rho &= \int_{\mathbb{R}}\int_{\mathbb{R}} (p\partial_q f - q\partial_p f)\, g\rho \, dpdq \\
&= -\int_{\mathbb{R}}\int_{\mathbb{R}} fp\partial_q(g\rho)\, dpdq + \int_{\mathbb{R}}\int_{\mathbb{R}} fq\partial_p(g\rho)\, dpdq \\
&= -\int_{\mathbb{R}}\int_{\mathbb{R}} fp(\partial_q g)\rho + \int_{\mathbb{R}}\int_{\mathbb{R}} qf(\partial_p g)\rho = -\langle f, \mathcal{L}_H g\rangle_\rho \,.
\end{aligned}$$

The generator of the O-U process is instead symmetric in $\mathcal{L}^2_\rho$ and in particular

$$\mathcal{L}_{OU} = -T^\star T,$$

where

$$T = \partial_p, \quad \text{so that} \quad T^\star = -\partial_p + p.$$

In conclusion, the generator of the Langevin equation decomposes into a symmetric and antisymmetric part. Moreover, the antisymmetric part comes from the Hamiltonian deterministic component of the dynamics, the symmetric part comes from the stochastic component.

Using Stone's Theorem (see Appendix B.3) we also know that the semigroup generated by $\mathcal{L}_H$ is norm-preserving, while it is easy to see that the semigroup generated by $\mathcal{L}_{OU}$ is dissipative, indeed

$$\begin{aligned}
\frac{d}{dt}\|e^{t\mathcal{L}_{OU}} h\|^2_\rho &= 2\langle \mathcal{L}_{OU} e^{t\mathcal{L}_{OU}} h, e^{t\mathcal{L}_{OU}} h\rangle_\rho \\
&= -2\langle T^\star T h_t, h_t\rangle_\rho = -2\|Th_t\|^2_\rho < 0\,,
\end{aligned}$$

where we used the notation $h_t(x) = e^{t\mathcal{L}_{OU}} h(x)$. In conclusion, so far we have the following picture:

$$\begin{array}{rccc} \mathcal{L} = & \underbrace{\mathcal{L}_H} & - & \underbrace{T^\star T} \\ & \text{skew symmetric} & & \text{symmetric} \\ & \downarrow & & \downarrow \\ & \text{deterministic} & & \text{stochastic} \\ & \text{conservative} & & \text{dissipative} \\ & \text{part of the dynamics} & & \text{part of the dynamics.} \end{array}$$

However it would be misleading to think that, for the Langevin equation, decay to equilibrium happens only because of the effect of the dissipative part of the dynamics. This would be true if the operators $\mathcal{L}_{OU}$ and $\mathcal{L}_H$ did commute, but they don't, so more interesting phenomena take place in the Langevin dynamics and exponential convergence to equilibrium happens because of the interaction between the symmetric and the antisymmetric part of the dynamics. But we are going a bit too fast, we are talking about exponential convergence to equilibrium and we haven't yet proved that such a thing happens for the Langevin dynamics. Clearly we cannot use the same technique that we have used for diffusions with symmetric generator. This is precisely the issue that we would like to address in the remainder of this section and that has been tackled in the book [77].

The hypocoercivity theory, subject of [77], is concerned with the problem of exponential convergence to equilibrium for evolution equations of the form

$$\partial_t h + (A^\star A - B)\, h = 0, \text{[36]} \tag{117}$$

where B is an antisymmetric operator [37]. We shall briefly present some of the basic elements of the hypocoercivity theory and then see what are the outcomes of such a technique when we apply it to the Langevin equation (111).

We first introduce the necessary notation. Let $\mathcal{H}$ be a Hilbert space, real and separable, $\|\cdot\|$ and $(\cdot,\cdot)$ the norm and scalar product of $\mathcal{H}$, respectively. Let A and B be unbounded operators with domains $\mathcal{D}(A)$ and $\mathcal{D}(B)$ respectively, and assume that B is antisymmetric, i.e. $B^\star = -B$, where * denotes

[36] Generalizations to the form $\partial_t h + \left(\sum_{i=1}^m A_i^\star A_i - B\right) h = 0$ as well as further generalizations are presented in [77]. We refer the reader to such a monograph for these cases.

[37] Notice that, for less than regularity issues, any second order differential operator can be written in this form.

adjoint in $\mathcal{H}$. We shall also assume that there exists a vector space $\mathcal{S} \subset \mathcal{H}$, dense in $\mathcal{H}$, where all the operations that we will perform involving A and B are well defined.

Writing the involved operator in the form $\mathcal{T} = A^\star A - B$ has several advantages. Some of them are purely computational. For example, for operators of this form checking the contractivity of the semigroup associated with the dynamics (117) becomes trivial. Indeed, the antisymmetry of B implies that

$$(Bx, x) = -(x, Bx) \quad \text{for all } x \in \mathcal{D}(B).$$

Therefore $(Bx, x) = 0$. This fact, together with $(A^\star Ax, x) = \|Ax\|^2 > 0$, immediately gives

$$\frac{1}{2}\frac{\partial}{\partial_t}\bigg|_{t=0^+} \|e^{-t\mathcal{T}}h\|^2 = -\|Ah\|^2 < 0\,.$$

On the other hand, conceptually, the decomposition $A^\star A - B$ is physically meaningful as the symmetric part of the operator, $A^\star A$, corresponds to the stochastic (dissipative) part of the dynamics, whereas the antisymmetric part corresponds to the deterministic (conservative) component.

Definition 10.27. *We say that an unbounded linear operator $\mathcal{T}$ on $\mathcal{H}$ is relatively bounded with respect to the linear operators $T_1, ..., T_n$ if the domain of $\mathcal{T}$, $\mathcal{D}(\mathcal{T})$, is contained in the intersection $\cap\mathcal{D}(T_j)$ and there exists a constant $\alpha > 0$ s.t.*

$$\forall h \in \mathcal{D}(\mathcal{T}), \qquad \|\mathcal{T}h\| \leq \alpha(\|T_1 h\| + ... + \|T_n h\|).$$

Definition 10.28 (Coercivity)**.** *Let $\mathcal{T}$ be an unbounded operator on a Hilbert space $\mathcal{H}$, denote its kernel by $\mathcal{K}$ and assume there exists another Hilbert space $\tilde{\mathcal{H}}$ continuously and densely embedded in $\mathcal{K}^\perp$. If $\|\cdot\|_{\tilde{\mathcal{H}}}$ and $(\cdot,\cdot)_{\tilde{\mathcal{H}}}$ are the norm and scalar product on $\tilde{\mathcal{H}}$, respectively, then the operator $\mathcal{T}$ is said to be λ-coercive on $\tilde{\mathcal{H}}$ if*

$$(\mathcal{T}h, h)_{\tilde{\mathcal{H}}} \geq \lambda \|h\|^2_{\tilde{\mathcal{H}}}\,, \quad \forall h \in \mathcal{K}^\perp \cap D(\mathcal{T}),$$

where $D(\mathcal{T})$ is the domain of $\mathcal{T}$ in $\tilde{\mathcal{H}}$.

Notice the parallel with (105). Not surprisingly, the following Proposition gives an equivalent definition of coercivity.

Proposition 10.29. *With the same notation as in Definition 10.28, $\mathcal{T}$ is λ-coercive on $\tilde{\mathcal{H}}$ iff*

$$\| e^{-\mathcal{T}t}h \|_{\tilde{\mathcal{H}}} \leq e^{-\lambda t} \| h \|_{\tilde{\mathcal{H}}} \quad \forall h \in \tilde{\mathcal{H}} \text{ and } t \geq 0.$$

Definition 10.30 (Hypocoercivity)**.** *With the same notation of Definition 10.28, assume $\mathcal{T}$ generates a continuous semigroup. Then $\mathcal{T}$ is said to be λ-hypocoercive on $\tilde{\mathcal{H}}$ if there exists a constant $\kappa > 0$ such that*

$$\| e^{-\mathcal{T}t} h \|_{\tilde{\mathcal{H}}} \leq \kappa e^{-\lambda t} \| h \|_{\tilde{\mathcal{H}}}, \quad \forall h \in \tilde{\mathcal{H}} \text{ and } t \geq 0. \tag{118}$$

Remark 10.31. We remark that the only difference between Definition 10.28 and Definition 10.30 is in the constant κ on the right hand side of (118), when $\kappa > 1$. Thanks to this constant, the notion of hypocoercivity is invariant under a change of equivalent norm, as opposed to the definition of coercivity which relies on the choice of the Hilbert norm. Hence the basic idea employed in the proof of exponentially fast convergence to equilibrium for degenerate diffusions generated by operators in the form (117), is to appropriately construct a norm on $\tilde{\mathcal{H}}$, equivalent to the existing one, and such that in this norm the operator is coercive.

We will state in the following the basic theorem in the theory of hypocoercivity. Generalizations can be found in [77].

Theorem 10.32. *With the notation introduced so far, let $\mathcal{T}$ be an operator of the form $\mathcal{T} = A^\star A - B$, with $B^\star = -B$. Let $\mathcal{K} = Ker\mathcal{T}$, define $C := [A, B]$ and consider on $\mathcal{K}^\perp$ [38] the norm*

$$\|h\|^2_{\mathcal{H}^1} := \|h\|^2 + \|Ah\|^2 + \|Ch\|^2 .$$

Suppose the following holds:

1. *A and $A^\star$ commute with C;*
2. *$[A, A^\star]$ is relatively bounded with respect to I and A;*
3. *$[B, C]$ is relatively bounded with respect to A, A^2, C and AC,*

then there exists a scalar product $((\cdot, \cdot))$ on $\mathcal{H}^1/\mathcal{K}$ defining a norm equivalent to the $\mathcal{H}^1$ norm such that

$$((h, \mathcal{T}h)) \geq k(\|Ah\|^2 + \|Ch\|^2), \quad \forall h \in \mathcal{H}^1/\mathcal{K}, \tag{119}$$

for some constant $k > 0$. If, in addition to the above assumptions, we have

$$A^\star A + C^\star C \text{ is } \kappa - \text{coercive for some } \kappa > 0,$$

then $\mathcal{T}$ is hypocoercive in $\mathcal{H}^1/\mathcal{K}$: there exist constants c, $\lambda > 0$ such that

$$\|e^{-t\mathcal{L}}\|_{\mathcal{H}^1/\mathcal{K} \to \mathcal{H}^1/\mathcal{K}} \leq c e^{-\lambda t}.$$

[38] One can prove that space $\mathcal{K}^\perp$ is the same irrespective of whether we consider the scalar product $\langle \cdot, \cdot \rangle$ of $\mathcal{H}$ or the scalar product $\langle \cdot, \cdot \rangle_{\mathcal{H}^1}$ associated with the norm $\| \cdot \|_{\mathcal{H}^1}$.

Comment. I will not write a proof of this theorem but I will explain how it works. The idea is the same that we have explained in Remark 10.31. Consider the norm

$$((h,h)) := \|h\|^2 + a\|Ah\|^2 + c\|Ch\|^2 + 2b(Ah, Ch),$$

where a, b and c are three strictly positive constants to be chosen. The assumptions 1. , 2. and 3. are needed to ensure that this norm is equivalent to the $\mathcal{H}^1$ norm, i.e. that there exist constant $c_1, c_2 > 0$ such that

$$c_1\|h\|_{\mathcal{H}^1} \leq ((h,h)) \leq c_2\|h\|_{\mathcal{H}^1}.$$

If we can prove that $\mathcal{T}$ is coercive in this norm, then by Proposition 10.29 and Remark 10.31 we have also shown exponential convergence to equilibrium in the $\mathcal{H}^1$ norm i.e. hypocoercivity. So the whole point is proving that

$$((\mathcal{T}h, h)) \geq K((h,h)),$$

for some $K > 0$. If 1. ,2. and 3. hold then (with a few lengthy but surprisingly not at all complicated calculations) (119) follows. From now on $K > 0$ will denote a generic constant which might not be the same from line to line. The coercivity of $A^*A + C^*C$ means that we can write

$$\begin{aligned}\|Ah\|^2 + \|Ch\|^2 &= \frac{1}{2}(\|Ah\|^2 + \|Ch\|^2) + \frac{1}{2}(\|Ah\|^2 + \|Ch\|^2) \\ &\geq \frac{1}{2}(\|Ah\|^2 + \|Ch\|^2) + \frac{\kappa}{2}\|h\|^2 \\ &\geq K\|h\|_{\mathcal{H}^1}\,.\end{aligned}$$

Combining this with (119), we obtain

$$((h, \mathcal{T}h)) \geq k(\|Ah\|^2 + \|Ch\|^2) \geq K\|h\|_{\mathcal{H}^1} \geq ((h,h)).$$

This concludes the proof. Another important observation is that, in practice, the coercivity of $A^\star A + C^\star C$ boils down to a Poincaré inequality. This will be clear when we apply this machinery to the Langevin equation, see proof of Theorem 10.35.

Remark 10.33. Let $\mathcal{K}$ be the kernel of $\mathcal{T}$ and notice that $Ker(A^\star A) = Ker(A)$ and $\mathcal{K} = Ker(A) \cap Ker(B)$. Suppose $KerA \subset KerB$; then $Ker\mathcal{L} = KerA$. In this case the coercivity of $\mathcal{T}$ is equivalent to the coercivity of $A^\star A$. So the case we are interested in is the case in which $A^\star A$ is coercive and $\mathcal{T}$ is not. In order for this to happen $A^\star A$ and B cannot commute; if they did, then $e^{-t\mathcal{L}} = e^{-tA^\star A}e^{-tB}$. Therefore, since e^{-tB} is norm preserving, we would have $\|e^{-t\mathcal{L}}\| = \|e^{-tA^\star A}\|$. This is the intuitive reason to look at commutators of the form $[A, B]$.

We can use Theorem 10.32 to prove exponentially fast convergence to equilibrium for the Langevin dynamics. We shall apply such a theorem to the operator $\mathcal{L}$ defined in (112) on the space $\mathcal{H} = L^2_\rho$, where ρ is the equilibrium distribution (112). (The space $\mathcal{S}$ can be taken to be the space of Schwartz functions.) The operators A and B are then

$$A = \partial_p \quad \text{and} \quad B = p\partial_q - \partial_q V \partial_p,$$

so that

$$C := [A, B] = AB - BA = \partial_q \,.$$

The kernel $\mathcal{K}$ of the operator $\mathcal{L}$ is made of constants and in this case the norm $\mathcal{H}^1$ will be the Sobolev norm of the weighted $H^1(\rho)$:

$$\|f\|^2_{H^1_\rho} := \|f\|^2_{L^2_\rho} + \|\partial_q f\|^2_{L^2_\rho} + \|\partial_p f\|^2_{L^2_\rho} \,.$$

Let us first calculate the commutators needed to check the assumptions of Theorem 10.32.

$$[A, C] = [A^\star, C] = 0, \quad [A, A^\star] = Id \tag{120}$$

and

$$[B, C] = -\sqrt{\beta^{-1}} \partial_q^2 V(q) \partial_p \,. \tag{121}$$

Lemma 10.34. *Let $V \in C^\infty(\mathbb{R})$. Suppose $V(q)$ satisfies*

$$|\partial_q^2 V| \le C\big(1 + |\partial_q V|\big), \tag{122}$$

for some constant $C \ge 0$. Then, for all $f \in H^1(\rho)$, there exists a constant $C > 0$ such that

$$\|(\partial_q^2 V)\, \partial_p f\|^2_{L^2(\rho)} \le C\Big(\|f\|^2_{L^2(\rho)} + \|(\partial_q V) f\|^2_{L^2(\rho)}\Big). \tag{123}$$

Theorem 10.35. *Let $V \in C^\infty(\mathbb{R})$, satisfying (122) and the assumptions of Lemma 10.24. Then, there exist constants C, $\lambda > 0$ such that for all $h_0 \in H^1(\rho)$,*

$$\left\| e^{-t\,\mathcal{L}} h_0 - \int h_0 \, d\rho \right\|_{H^1(\rho)} \le C e^{-\lambda t} \|h_0\|_{H^1(\rho)}. \tag{124}$$

Proof. We will use Theorem 10.32. We need to check that conditions (i) to (iv) of the theorem are satisfied. Conditions (i) and (ii) are satisfied, due to (120). Having calculated (121), condition (iii) requires $\partial_q^2 V \partial_p$ to

be relatively bounded with respect to ∂_p, ∂_p^2, ∂_q and ∂_{qp}^2. From Lemma 10.34 this is trivially true as soon as condition (122) holds. Now we turn to condition (iv). Let us first write the operator $\widehat{\mathcal{L}} = A^\star A + C^\star C$:

$$\widehat{\mathcal{L}} = p\partial_p - \partial_p^2 + \partial_q V \partial_q - \partial_q^2.$$

In order for this operator to be coercive, it is sufficient for the Gibbs measure $\rho_\beta(dpdq) = \frac{1}{Z} e^{-H(q,p)} dpdq$ to satisfy a Poincare inequality. This probability measure is the product of a Gaussian measure (in p) which satisfies a Poincaré inequality, and of the probability measure $e^{-V(q)}dq$. It is sufficient, therefore, to prove that $e^{-V(q)}dq$ satisfies a Poincaré inequality. This follows from Lemma 10.24 and Lemma 10.34. □

11 Anomalous diffusion - - - STILL DRAFT

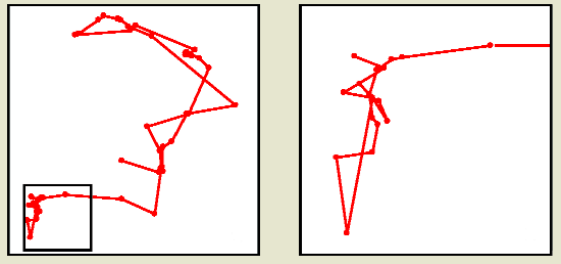


Figure 4: example of superdiffusive path.

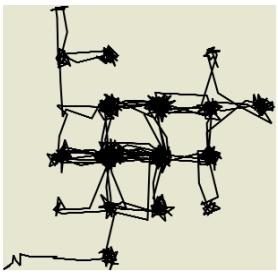


Figure 5: example of subdiffusive path.

Anomalous diffusion processes are characterized by a mean square displacement which, instead of growing linearly in time, grows like $t^{2\gamma}$, $\gamma > 0$, $\gamma \neq \frac{1}{2}$. When $0 < \gamma < \frac{1}{2}$ the process is subdiffusive, when $\gamma > \frac{1}{2}$ it is superdiffusive.
Diffusion phenomena can be described at the microscopic level by BM and macroscopically by the heat equation, i.e. the parabolic problem associated with the Laplacian operator; the link between the two descriptions is, roughly speaking, the fact that the fundamental solution to the diffusion equation is the probability density associated with BM.
A similar picture can be obtained for anomalous diffusion. The main difference is that in nature a variety of anomalous diffusion phenomena can be observed and the question is how to characterize them from both the analytical and the statistical point of view. It has been shown that the microscopical (probabilistic) approach can be understood in the context of continuous time random walks (CTRW) and, in this framework, a process is uniquely determined once the probability density to move at distance r

in time t is known ([**?**]- [11], [**?**], [**?**] and references therein). The analytical approach is based on the theory of fractional differentiation operators, where the derivative can be fractional either in time or in space (see [**?**]-[**?**], [**?**] and references therein).

For $f(s)$ regular enough (e.g. $f \in \mathcal{C}(0,t]$ with an integrable singularity at $s=0$), let us introduce the Riemann-Liouville fractional derivative,

$$D_t^\gamma(f) := \frac{1}{\Gamma(2\gamma)} \frac{d}{dt} \int_0^t ds \frac{f(s)}{(t-s)^{1-2\gamma}}, \qquad 0 < \gamma < \frac{1}{2}, \tag{125}$$

and the Riemann-Liouville fractional integral,

$$I_t^\gamma(f) := \frac{1}{\Gamma(2\gamma-1)} \int_0^t ds \frac{f(s)}{(t-s)^{2-2\gamma}}, \qquad \frac{1}{2} < \gamma < 1, \tag{126}$$

where Γ is the Euler Gamma function ([**?**]). Appendix **??** contains a motivation for introducing such operators. For $\frac{1}{2} < \gamma < 1$ let us also introduce the fractional Laplacian $\Delta^{(\gamma)}$, defined through its Fourier transform: if the Laplacian corresponds, in Fourier space, to a multiplication by $-k^2$, the fractional Laplacian corresponds to a multiplication by $-\mid k \mid^{\frac{1}{\gamma}}$. (125) and (126) can be defined in a more general way (see [**?**]), but to our purposes the above definition is sufficient. Furthermore, notice that the operators in (125) and (126) are fractional in time, whereas the fractional Laplacian is fractional in space.

Let us now consider the function $\rho^\gamma(t,x)$, solution to

$$\partial_t \rho^\gamma(t,x) = \frac{1}{\Gamma(2\gamma)} \frac{d}{dt} \int_0^t ds \, \frac{\partial_x^2 \rho^\gamma(s,x)}{(t-s)^{1-2\gamma}} \quad \text{when } 0 < \gamma < \frac{1}{2}, \tag{127}$$

$$\partial_t \rho^\gamma(t,x) = \frac{1}{\Gamma(2\gamma-1)} \int_0^t ds \, \frac{\partial_x^2 \rho^\gamma(s,x)}{(t-s)^{2-2\gamma}} \quad \text{when } \frac{1}{2} < \gamma < 1. \tag{128}$$

It has been shown (see [**?**, **?**] and references therein) that such a kernel is the asymptotic of the probability density of a CTRW run by a particle either moving at constant velocity between stopping points or instantaneously jumping between halt points, where it waits a random time before jumping again. On the other hand, a classic result states that the Fourier transform of the solution $\rho^\gamma(t,x)$ to

$$\partial_t \rho^\gamma(t,x) = \frac{1}{2} \Delta^{(\gamma)} \rho^\gamma(t,x), \qquad \frac{1}{2} < \gamma < 1, \tag{129}$$

is, for $\gamma \geq \frac{1}{2}$, the characteristic function of a (stable) process whose first moment is divergent when $\gamma \geq 1$ (see [**?**]); this justifies the choice $\frac{1}{2} < \gamma < 1$

in equation (129). Processes of this kind are particular CTRWs, the well known Lévy flights; in this case large jumps are allowed with non negligible probability and this results in the process having divergent second moment. We will use the notation $\rho^\gamma(t,x) = \rho^\gamma_t(x)$ to indicate the solution to either (127), (128) or (129), as in the proofs we use only the properties that these kernels have in common.
The above described framework is analogous to the one of Einstein diffusion: for subdiffusion and Riemann-type superdiffusion the statistical description is given by CTRWs, whose (asymptotical) density is the fundamental solution of the evolution equation associated with the operators of fractional differentiation and integration, i.e. (127) and (128), respectively (see Appendix B). For the Lévy-type superdiffusion, the statistical point of view is given by Lévy flights, whose probability density evolves in time according to the evolution equation associated with the fractional Laplacian, i.e. (129) (see [**?**]).

we want to show how the operators D^γ_t and I^γ_t naturally arise in the context of anomalous diffusion and explain in some more detail the link with CTRWs.
We want to determine an operator A s.t.

$$\begin{cases} \partial_t \rho^\gamma_t(x) = A\, \rho^\gamma_t(x) \\ \rho^\gamma_t(0) = \delta_0, \end{cases}$$

with $\rho^\gamma(t,x)$ enjoying the following three properties:

$$\int_{\mathbb{R}} dx \rho^\gamma_t(x) = 1\,, \quad \int_{\mathbb{R}} dx \rho^\gamma_t(x)\, x = 0 \quad \text{e} \quad \int_{\mathbb{R}} dx \rho^\gamma_t(x)\, x^2 \sim t^{2\gamma} \tag{130}$$

(notice that for $\gamma = \frac{1}{2}$ we recover the diffusion equation with $A = \Delta$). We recall that $\hat{f}$, $f^{\#}$ and $\tilde{f}$ denote the Fourier, the Laplace and the Fourier-Laplace transform of the function f, respectively.
By (130), the following must hold

$$\hat{\rho}^\gamma_t(k) = 1 - \frac{1}{2} c t^{2\gamma} k^2 + o(k^2) \qquad \text{and}$$

$$\tilde{\rho}^\gamma(\mu,k) = \frac{1}{\mu} - \frac{c k^2}{2\mu^{2\gamma+1}} \Gamma(2\gamma+1) = \frac{1}{\mu}(1 - c_1 \mu^{-2\gamma} k^2),$$

where $c_1 = \frac{1}{2} c \Gamma(2\gamma+1)$. In definitions (127) and (128) the constant c_1 should appear; we just set it equal to 1 both for simplicity and not being interested, in this context, in estimating the "anomalous diffusion" constant.

We can assume that the expression for $\tilde{\rho}^\gamma(\mu,k)$ is valid in the regime $\mu^{-2\gamma}k^2 << 1$. Actually, condition $(130)_3$ is meant for an infinitely wide system and for long times. In other words, if Λ is the region where the particle moves, we claim that

$$\lim_{t\to\infty}\lim_{\Lambda\to\mathbb{R}}\frac{\int_\Lambda dx\rho_t^\gamma(x)\,x^2}{t^{2\gamma}}=\text{const.}$$

This means that we are interested in the case $k << \mu$. Of course one can in principle find an infinite number of functions s.t. $\tilde{\rho}^\gamma(\mu,k)=\frac{1}{\mu}(1-c_1\epsilon)$ for $\epsilon=\mu^{-2\gamma}k^2$. One possible choice is

$$\tilde{\rho}^\gamma(\mu,k)=\frac{1}{\mu(1+c_1\epsilon)}=\mu^{\gamma-1}\frac{\mu^\gamma}{\mu^{2\gamma}+(c_1k)^2}=\frac{1}{\mu+c_1k^2\mu^{1-2\gamma}},\qquad(131)$$

which leads to an integro-differential equation and, when $\gamma=\frac{1}{2}$, it coincides with the Fourier-Laplace transform of a Gaussian density.
We now find the operator whose fundamental solution is $\tilde{\rho}^\gamma(\mu,k)$. We have

$$\mathcal{L}(\partial_t\hat{\rho}^\gamma(\cdot,k))(\mu)=-1+\mu\tilde{\rho}^\gamma(\mu,k)=-c_1k^2\mu^{1-2\gamma}\tilde{\rho}^\gamma(\mu,k).$$

Let $p=2\gamma-1$ and $\phi_p(t)=\frac{t^{p-1}}{\Gamma(p)}$; then we need to distinguish two cases in order to study the right hand side of the above equation:
when $0<\gamma<\frac{1}{2}$ one can easily check that

$$\mathcal{L}(\phi_p*\hat{\rho}^\gamma(k,\cdot))=\tilde{\rho}^\gamma(\mu,k)\mu^{-p}$$

which implies that

$$\tilde{\rho}^\gamma(\mu,k)\mu^{1-2\gamma}\quad\text{is the Laplace transform of}\quad\frac{1}{\Gamma(2\gamma-1)}\int_0^t ds\frac{\hat{\rho}^\gamma(s,k)}{(t-s)^{2-2\gamma}};$$

when $\frac{1}{2}<\gamma<1$, instead, a straightforward calculation shows that

$$\mathcal{L}[\partial_t(\phi_{p+1}*\hat{\rho}^\gamma(k,\cdot))]=\tilde{\rho}^\gamma(\mu,k)\mu^{-p}$$

so that

$$\tilde{\rho}^\gamma(\mu,k)\mu^{1-2\gamma}\quad\text{is the Laplace transform of}\quad\frac{1}{\Gamma(2\gamma)}\frac{d}{dt}\int_0^t ds\frac{\hat{\rho}^\gamma(s,k)}{(t-s)^{1-2\gamma}}.$$

Finally, taking the inverse Fourier transform, we get that $\rho^\gamma(t,x)$ satisfies (127) when $0<\gamma<\frac{1}{2}$ and (128) when $\frac{1}{2}<\gamma<1$. Moreover, the explicit expression for $\rho_t^\gamma(x)$ holds true: by (131) we get that

$$\tilde{\rho}^\gamma(\mu,k)=\int_R dx\,e^{ikx}\,\frac{\mu^{\gamma-1}}{2\sqrt{c_1}}\,e^{-\frac{\mu^\gamma}{\sqrt{c_1}}|x|}$$

hence

$$\rho^{\#}(x,\mu) = \frac{\mu^{\gamma-1}}{2\sqrt{c_1}}\, e^{-\frac{\mu^{\gamma}}{\sqrt{c_1}}|x|}$$

and now, by the inverse Laplace formula, we obtain (**??**). Obviously, the expression (**??**) has been deduced after having chosen (131) among all possible candidates for $\tilde{\rho}^{\gamma}$ and this choice can now be justified in view of the link with CTRWs.

Appendices

A Miscellaneous facts

This appendix contains some miscellaneous material.

A.1 Why can we think of white noise as the derivative of Brownian Motion

We said that "white noise is the derivative of BM", without justifying this statement. We will give here a formal explanation. If "white noise is the derivative of BM" then in some sense

$$"\xi_t = \lim_{h\to 0} \frac{W_{t+h} - W_t}{h}".$$

By definition of BM, the process on the RHS of the above must be mean zero and Gaussian, so let us check its covariance function:

$$\begin{aligned}
&\mathbb{E}\left[\frac{W_{t+h} - W_t}{h}\,\frac{W_{s+h} - W_s}{h}\right] \\
&= \frac{1}{h^2}[(t+h)\wedge(s+h) - (t+h)\wedge s - (s+h)\wedge t - s\wedge t] \\
&= \begin{cases} [[(s\wedge t)+h] - ([t\vee s]\wedge([s\wedge t]+h))]/h^2 & \text{if } s\neq t \\ 1/h & \text{if } t = s \end{cases} \\
&\longrightarrow \begin{cases} 0 & \text{if } t\neq s \\ +\infty & \text{if } t = s \end{cases} \quad = \delta_0(t-s)\,,
\end{aligned}$$

because if $s < t$ then $t \wedge (s+h) = s+h$ as h gets smaller (analogous in the case $t < s$).

A.2 Gronwall's Lemma

- **Differential Form.** Let $u(t)$ and $a(t)$ be real valued differentiable functions. If
$$\frac{d}{dt}u(t) \leq a(t)u(t)$$
then
$$u(t) \leq u(c)e^{\int_c^t a(s)ds}\,.$$

- **Integral form:** Let $u(t), a(t)$ and $b(t)$ be continuous real valued functions on $\mathbb{R}_+$, with $b(t)$ nonnegative and $a(t)$ non decreasing. If

$$u(t) \le a(t) + \int_c^t b(s)u(s)\,ds\,, \quad \text{for some } c \le t,$$

then

$$u(t) \le a(t)e^{\int_c^t b(s)\,ds}\,.$$

A.3 Kolmogorov's Extension Theorem

A sequence μ_N of probability measures on $(\mathbb{R}^N, \mathcal{B}(\mathbb{R}^N))$ is *consistent* if

$$\mu_{N+1}((a_1,b_1)\times\cdots\times(a_N,b_N)\times\mathbb{R}) = \mu_N((a_1,b_1)\times\cdots\times(a_N,b_N))\,.$$

We denote by $\mathbb{R}^{\mathbb{N}}$ the space of real sequences endowed with the σ-algebra $\mathcal{R}^{\mathbb{N}}$ generated by the cylinder sets, i.e. by sets of the form

$$C_{A_0,\dots,A_m} := \{\omega = (\omega_0,\omega_1,\omega_2,\dots) \in \mathbb{R}^{\mathbb{N}} : \omega_i \in A_i, i = 0,\dots,m\},$$

$m \in \mathbb{N}$ and $A_i = (a_i, b_i) \subset \mathbb{R}$.

Theorem A.1 (Kolmogorov's Extension Theorem – countable version)**.** *Let μ_N be a consistent sequence of probability measures. Then there exists a unique probability measure P on sequence space $(\mathbb{R}^{\mathbb{N}}, \mathcal{R}^{\mathbb{N}})$ such that*

$$P(\omega = (\omega_1,\omega_2,\dots) : \omega_i \in (a_i,b_i), i = 1,\dots,N) = \mu_N((a_1,b_1)\times\cdots\times(a_N,b_N)).$$

The above Theorem still holds if instead of $\mathbb{R}$ we consider any Polish space S endowed with a σ-algebra $\mathcal{S}$ and let S^N be the space of N-vectors with components in S, endowed with the σ-algebra $\mathcal{S}^N$.

B Elements of Functional Analysis

Excellent references for the material of this appendix are the books [80, 64] Let $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ be two Banach spaces.

Definition B.1. *A linear operator $A : X \to Y$ is* bounded *if there exists a constant $M > 0$ such that*

$$\|Ax\|_Y \le M\|x\|_X, \quad \textit{for all } x \in X\,. \tag{132}$$

Notice that the constant M is independent on $x \in X$. The space of bounded linear operators $A : X \to Y$ is a vector space, denoted by $B(X, Y)$, and it can be endowed with the operator norm

$$\|A\|_{B(X,Y)} := \sup_{x \neq 0} \frac{\|Ax\|_Y}{\|x\|_X} = \sup_{\|x\|_X = 1} \frac{\|Ax\|_Y}{\|x\|_X},$$

where the second inequality in the above is a consequence of linearity. If $X = Y$ then we simply write $B(X)$.

Recall the two following facts: 1. A linear operator $A : X \to Y$ is bounded if and only if it is continuous; 2. $\|A\|_{B(X,Y)}$ is the smallest constant M such that (132) holds.

When dealing with bounded opertors there is no need to specify their domain. An unbounded operator L, on the contrary, is specified by its action as well as by its domain. In other words, it is not correct to talk about "the unbounded operator L", it is more precise to talk about the pair $(L, \mathcal{D}(L))$, where $\mathcal{D}(L)$ denotes the domain of L.

Definition B.2. *An operator $A : \mathcal{D}(A) \to Y$ is* closed *if*

$$x_n \in \mathcal{D}(A),\ x_n \to x,\ Ax_n \to y \implies x \in \mathcal{D}(A) \text{ and } Ax = y\,.$$

A bounded linear operator $A : X \to Y$ (X and Y Banach spaces) is closed; in general a closed operator will not be bounded. However if A is closed and defined on the whole of X then it is also bounded.

Remember that given a normed space V, the (real) dual space of V, denoted V^*, is the space of bounded linear functionals on V, i.e. the space of bounded linear maps $T : V \to \mathbb{R}$. Such a space is a vector space and it becomes a normed space when endowed with the operator norm.

We denote the *duality relation* by $\langle \cdot, \cdot \rangle_{V^*,V}$ or, when there is no risk of confusion, by simply $\langle \cdot, \cdot \rangle$; the duality relation is just the action of V^* on V i.e. the action of T on the elements of V. In other words, given the linear functional $T \in V^*$, instead of writing $T(x)$ or Tx, for $x \in V$, we will write $\langle T, x \rangle$.

B.1 Adjoint operator

Definition B.3 (Adjoint of bounded operator)**.** *Let X and Y be two Banach spaces and $A : X \to Y$ a linear bounded operator. We say that $A^* : Y^* \to X^*$ is the* adjoint *or* dual *of A if*

$$\langle Ax, y \rangle_{Y,Y^*} = \langle x, A^* y \rangle_{X,X^*}, \quad \forall x \in X, y \in Y^*.$$

Remark B.4. In a Hilbert space setting the duality relation is just the scalar product. Because the dual of a Hilbert space is the Hilbert space itself, the above definition coincides, in the Hilbert case, with the one that you will have already seen, which is: if H is a Hilbert space and $A : H \to H$ is a bounded linear operator, $A^* : H \to H$ is the adjoint operator of A if $\langle Ax, y\rangle = \langle x, A^*y\rangle$, for all $x, y \in H$ (where this time $\langle \cdot, \cdot \rangle$ is just the scalar product in H).

If the operator is unbounded, the definition of adjoint is slightly more involved.

Definition B.5 (Adjoint of unbounded operator)**.** *Let X and Y be Banach spaces and A be an unbounded operator $A : \mathcal{D}(A) \subset X \to Y$. If*[39] *$\mathcal{D}(A)$ is dense in X then for every $y \in Y^*$ there exists a unique $\tilde{x} \in X^*$ such that*

$$\langle Ax, y\rangle_{Y,Y^*} = \langle x, \tilde{x}\rangle_{X,X^*}, \quad \forall x \in \mathcal{D}(A). \tag{133}$$

Therefore we can define an operator $A^* : \mathcal{D}(A^*) \subset Y^* \to X^*$ with $T^*y = \tilde{x}$. Such an operator will satisfy

$$\langle Ax, y\rangle_{Y,Y^*} = \langle x, A^*y\rangle_{X,X^*}, \quad \forall x \in \mathcal{D}(A).$$

The domain of A^* is $\mathcal{D}(A^*) := \{y \in Y^* : \exists\, \tilde{x} \in X^* \text{satisfying } (133)\}$.

Definition B.6 (Symmetric and self adjoint operators on Hilbert spaces)**.** *Let H be a Hilbert space with scalar product $\langle \cdot, \cdot \rangle$.*
• *Let $A : H \to H$ be a bounded linear operator. A is* self-adjoint *if $A = A^*$ i.e. if*

$$\langle Ax, y\rangle = \langle x, Ay\rangle, \quad \forall x, y \in H.$$

• *Let $A : H \to H$ be an unbounded operator on the Hilbert space H, and suppose that the domain of A is dense in H. Then A is* symmetric *if $\mathcal{D}(A) \subset \mathcal{D}(A^*)$ and $Ax = A^*x$ for all $x \in \mathcal{D}(A)$. Equivalently, it is symmetric if $\langle Ax, y\rangle = \langle x, Ay\rangle$ for all $x, y \in \mathcal{D}(A)$.*
• *Let $A : H \to H$ be an unbounded operator on the Hilbert space H, and suppose that the domain of A is dense in H. A is* self-adjoint *if $A = A^*$, i.e. if it is symmetric and $\mathcal{D}(A) = \mathcal{D}(A^*)$.*

Now one last definition that will be useful in Appendix B.3.

Definition B.7. *A bounded linear operator U on a Hilbert space H is* unitary *if $Range(U) = H$ and U preserves the scalar product:*

$$\langle Ux, Uy\rangle = \langle x, y\rangle, \quad \textit{for all } x, y \in H. \tag{134}$$

[39] This is actually an "if and only if"

Notice that for a bounded linear operator on a Hilbert space condition (134) is equivalent (see [80]) to

$$\|Ux\| = \|x\|, \quad \text{for all } x, y \in H.$$

One can also prove that a bounded linear operator on a Hilbert space is unitary if anf only if $U^* = U^{-1}$.

B.2 Strong, weak and weak-* convergence.

Let $(V, \|\cdot\|)$ be a normed space and V^* its dual (i.e. the space of bounded linear operators $T : V \to \mathbb{R}$), endowed with the operator norm.

i) A sequence x_n of elements of V is said to *converge (strongly)* or *in norm* to $x \in V$ if

$$\lim_{n\to\infty} \|x - x_n\| = 0\,.$$

ii) A sequence x_n of elements of V is said to *converge weakly* to $x \in V$ if

$$\lim_{n\to\infty} Tx_n = Tx \quad \text{for all } T \in V^*.$$

As we have observed, the space V^* is a normed space itself, when endowed with the operator norm. In this section we use the notation $(V^*, \|\cdot\|_*)$. Therefore in V^* it makes sense to consider the two types of convergence described above. However, when working on the dual space, we can also define another type of convergence:

i) A sequence T_n of elements of V^* is said to *converge (strongly)* or *in norm* to $T \in V^*$ if

$$\lim_{n\to\infty} \|T - T_n\| = 0\,.$$

ii) A sequence x_n of elements of V is said to *converge weakly* to $x \in V$ if

$$\lim_{n\to\infty} \mathcal{T}T_n = \mathcal{T}T \quad \text{for all } \mathcal{T} \in V^{**}.$$

iii) A sequence T_n of elements of V^* is said to *converge weak-** to $T \in V^*$ if

$$\lim_{n\to\infty} T_n x = Tx, \quad \text{for all } x \in V.$$

B.3 Groups of bounded operators and Stone's Theorem

Definition B.8. *A one parameter family of linear bounded operators over a Banach space $\mathfrak{B}$, $\{\mathcal{P}_t\}_{t\in\mathbb{R}}$, $\mathcal{P}_t : \mathfrak{B} \to \mathfrak{B}$ for all $t \geq 0$, is a* group *of bounded linear operators if*

1. *$\mathcal{P}_0 = I$, where I denotes the identity on $\mathfrak{B}$;*
2. *$\mathcal{P}_{t+s} = \mathcal{P}_t\mathcal{P}_s = \mathcal{P}_s\mathcal{P}_t$, for all $t, s \in \mathbb{R}$.*

If the map $\mathbb{R} \ni t \longrightarrow \mathcal{P}_t f \in \mathfrak{B}$ is continuous for all $f \in \mathfrak{B}$, then the group is said to be strongly continuous. *A strongly continuous group of bounded linear operators is also called a $\mathcal{C}_0$-*group. *The* infinitesimal generator *of the group $\mathcal{P}_t$ is the operator*

$$\mathcal{L}f := \lim_{t\to 0} \frac{\mathcal{P}_t f - f}{t}, \tag{135}$$

for all $f \in \mathcal{D}(\mathcal{L}) := \{f \in \mathfrak{B} :$ the limit on the RHs of (135) exists in $\mathfrak{B}\}$.

It is clear that if $\mathcal{P}_t$, $t \in \mathbb{R}$, is a $\mathcal{C}_0$-group then $\mathcal{P}_t$, $t \in \mathbb{R}_+$, is a $\mathcal{C}_0$-semigroup generated by $\mathcal{L}$ and $\mathcal{P}_{-t}$, $t \in \mathbb{R}_+$ is a $\mathcal{C}_0$-semigroup generated by $-\mathcal{L}$. If $\mathcal{P}_t$ is invertible (as an operator) then the inverse is precisely $\mathcal{P}_{-t}$.

Theorem B.9 (Stone's Theorem)**.** *$\mathcal{L}$ is the generator of a $\mathcal{C}_0$ group of unitary operators on a Hilbert space if and only if iA is self-adjoint.*

B.4 Functional Inequalities in their basic form

These are the functional inequalities that you will have seen in a first course on PDEs.

- **Poincaré Inequality (in its most classic form)** Let U be a bounded, connected, open subset of $\mathbb{R}^n$ with a C^1 boundary ∂U. Then for every $1 \leq p \leq \infty$ and $f \in W^{1,p}(U)$,

  $$\|f - \langle f\rangle_U\|_{L^p(U)} \leq C\|\nabla f\|_{L^p(U)},$$

  where $\langle f\rangle_U$ denotes the average of f over U and C is a constant depending on n, p and U.

C Laplace method

The Laplace method is a technique to determine the behaviour of integrals of the form

$$I(\epsilon) = \int_{\mathbb{R}} f(x) e^{-h(x)/\epsilon} dx, \quad \text{as } \epsilon \to 0.$$

We assume that the functions $f(x), h(x) : \mathbb{R} \to \mathbb{R}$ are C^∞ and that $h(x)$ admits a unique global minimum, which is attained at $x = x_0$. In other words there exists a unique $x_0 \in \mathbb{R}$ such that

1. $h'(x_0) = 0$ and $h''(x_0) > 0$;
2. $h(x_0) < h(x)$ for all $x \neq x_0$.

Rewriting

$$I(\epsilon) = e^{-h(x_0)/\epsilon} \int_{\mathbb{R}} f(x) e^{[h(x_0)-h(x)]/\epsilon} dx$$

and observing that

$$e^{[h(x_0)-h(x)]/\epsilon} \longrightarrow \begin{cases} 0 & x \neq x_0 \\ 1 & x = x_0 \,, \end{cases}$$

it is clear that the main contribution to the value of the integral $I(\epsilon)$ will come from a neighbourhood of the point x_0. So for some small $\delta = \delta(\epsilon) > 0$ we can write

$$\begin{aligned} I(\epsilon) &= e^{-h(x_0)/\epsilon} \int_{\mathbb{R}} f(x) e^{[h(x_0)-h(x)]/\epsilon} \\ &\approx e^{-h(x_0)/\epsilon} \int_{x_0-\delta}^{x_0+\delta} f(x) e^{[h(x_0)-h(x)]/\epsilon} \\ &\approx e^{-h(x_0)/\epsilon} f(x_0) \int_{x_0-\delta}^{x_0+\delta} e^{[h(x_0)-h(x)]/\epsilon} \\ &\approx e^{-h(x_0)/\epsilon} f(x_0) \int_{\mathbb{R}} e^{[h(x_0)-h(x)]/\epsilon} \\ &\approx f(x_0) \int_{\mathbb{R}} e^{-h(x)/\epsilon}. \end{aligned}$$

In the above, when I write $X \approx Y$ I mean that X is equal to Y plus terms that go to zero as $\epsilon \to 0$. The above formal calculation shows that

$$\frac{I(\epsilon)}{\int_{\mathbb{R}} e^{-h(x)/\epsilon} dx} \xrightarrow{\epsilon \to 0} f(x_0).$$

Clearly all the above clculations are only formal, and in doing a rigorous proof one should quantify all the $\approx$. The Laplace method consists in using an idea similar to the one described above, in order to find the behaviour of $I(\epsilon)$ to leading order (in ϵ):

$$\begin{aligned}
I(\epsilon) &= e^{-h(x_0)/\epsilon} \int_{\mathbb{R}} f(x) e^{[h(x_0)-h(x)]/\epsilon} dx \\
&\approx e^{-h(x_0)/\epsilon} f(x_0) \int_{x_0-\delta}^{x_0+\delta} e^{[-h'(x_0)(x-x_0)/\epsilon} e^{[-h''(x_0)(x-x_0)^2]/2\epsilon} \\
&= e^{-h(x_0)/\epsilon} f(x_0) \int_{x_0-\delta}^{x_0+\delta} e^{[-h''(x_0)(x-x_0)^2]/2\epsilon} dx \\
&\approx e^{-h(x_0)/\epsilon} f(x_0) \int_{\mathbb{R}} e^{[-h''(x_0)(x-x_0)^2]/2\epsilon} dx \\
&= e^{-h(x_0)/\epsilon} f(x_0) \int_{\mathbb{R}} e^{-v^2/2} \sqrt{\frac{\epsilon}{2h''(x_0)}} dv \\
&= e^{-h(x_0)/\epsilon} f(x_0) \sqrt{\frac{2\pi\epsilon}{h''(x_0)}}.
\end{aligned}$$

So as $\epsilon \to 0$,

$$I(\epsilon) \approx e^{-h(x_0)/\epsilon} f(x_0) \sqrt{\frac{2\pi\epsilon}{h''(x_0)}}, \quad \text{to leading order.}$$

References

[1] L. Arnold. Stochastic Differential Equations: Theory and Applications, *Wiley*, 1974

[2] P. Baldi, Calcolo delle Probabilità e Statistica. Second Edition. Mc Graw-Hill, Milano 1998.

[3] A. Beskos, G. Roberts and A. M. Stuart. Optimal Scalings for local Metropolis-Hastings chains on nonproduct targets in high dimensions. *Ann. Appl. Prob.* **19** 863-898

[4] D. Bakry and M. Emery, *Diffusions hypercontractives*, pp. 177-206 in Sém. de Probab. XIX, Lecture Notes in Math., Vol. 1123, Springer-Verlag, Berlin, 1985.

[5] D. Bakry, I. Gentil and M. Ledoux. Analysis and Geometry of Markov Diffusion Operators. Springer, 2013

[6] A. Beskos, G. Roberts, A. M. Stuart and J. Voss. MCMC methods for diffusion bridges. *Stoch. Dyn.* **8** 319-350.

[7] A. Beskos and A. M. Stuart. MCMC methods for sampling function space. In *ICIAM Invited Lecture 2007*. European Mathematical Society, Zürich.

[8] P. Billingsley. *Convergence of probability measures.* Second edition. Wiley Series in Probability and Statistics: Probability and Statistics. New York, 1999.

[9] P. Billingsley. *Ergodic Theory and information.* Robert E. Krieger Publishing Company. Huntington, New York, 1978.

[10] G. Casella and E.I. George. Explaining the Gibbs sampler.(English summary) *Amer. Statist.* 46 (1992), no. 3, 167–174.

[11] A.V. Chechkin, J. Klafter and I.M. Sokolov. *Distributed-Order Fractional Kinetics*, 16th Marian Smoluchowski Symposium on Statistical Physics: Fundamental and Applications, September 2003

[12] G. Da Prato and J. Zabczyk, *Stochastic equations in infinite dimensions*, Cambridge University Press, Cambridge, 1992.

[13] G. Da Prato and J. Zabczyk. *Ergodicity for infinite dimensional systems*, London Mathematical Society Lecture Note Series, 229. Cambridge University Press, Cambridge, 1996.

[14] J. Davidson. *Asymptotic Methods and Functional Central Limit Theorems*. Chapter 5

[15] R. Durrett. *Probability: theory and examples*. Fourth edition. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2010.

[16] J.-P. Eckmann and M. Hairer. Non-equilibrium statistical mechanics of strongly anharmonic chains of oscillators, *Comm. Math. Phys.*, 212(1): 105–164, 2000.

[17] J.-P. Eckmann and M. Hairer. Spectral properties of hypoellliptic operators. *Comm. Math. Phys.*, 235, 233-253 (2003).

[18] J-P. Eckmann, C-A. Pillet, and L. Rey-Bellet. Entropy production in nonlinear, thermally driven Hamiltonian systems, *J. Statist. Phys.*, 95(1-2): 305–331, 1999.

[19] J.-P. Eckmann, C.-A. Pillet and L. Rey-Bellet. Non-equilibrium statistical mechanics of anharmonic chains coupled to two heat baths at different temperatures, *Comm. Math. Phys.*, 201(3): 657–697, 1999.

[20] S.N. Ethier and T.G. Kurtz. *Markov processes*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics (1986).

[21] L. C. Evans. *Partial Differential Equations*. American Mathematical Society, 2010.

[22] I. I. Gihman and A. V. Skorohod. Stochastic differential equations. Translated from the Russian by Kenneth Wickwire. Ergebnisse der Mathematik und ihrer Grenzgebiete, Band 72. *Springer-Verlag, New York-Heidelberg*, 1972.

[23] D. Givon, R. Kupferman and A.M. Stuart. Extracting macroscopic dynamics: model problems and algorithms, *Nonlinearity*, 17(6) 2004.

[24] G. R. Grimmett and D. R. Stirzaker. *Probability and random processes*. Oxford University Press, New York, 2001.

[25] A. Guionnet and B. Zegarlinski. *Lectures on Logarithmic Sobolev inequalities.* Lecture Notes.

[26] M. Hairer. Ergodic Theory for Stochastic PDEs. Lecture notes.

[27] M. Hairer. Ergodic properties of a class of non-Markovian processes. In *Trends in stochastic analysis*, volume 353 of *London Math. Soc. Lecture Note Ser.*, pages 65–98. Cambridge Univ. Press, Cambridge, 2009.

[28] M. Hairer. How hot can a heat bath get? *Comm. Math. Phys.*, 292 (2009), **1**, 131-177.

[29] M. Hairer and G. A. Pavliotis. From ballistic to diffusive behavior in periodic potentials. *J. Stat. Phys.*, 131(1): 175–202, 2008.

[30] W.K. Hastings. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, **57**, No. 1. (Apr., 1970), 97–109.

[31] P. Hanggi, P. Talkner, and M. Borkovec. Reaction-rate theory: fifty years after Kramers. *Rev. Modern Phys.*, 62(2):251–341, 1990.

[32] B. Helffer and F. Nier. *Hypoelliptic estimates and spectral theory for Fokker-Planck operators and Witten Laplacians*, volume 1862 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2005.

[33] F. Hérau. Short and long time behavior of the Fokker-Planck equation in a confining potential and applications, *J. Funct. Anal.*, 244(1): 95–118, 2007.

[34] F. Hérau and F. Nier. Isotropic hypoellipticity and trend to equilibrium for the Fokker-Planck equation with a high-degree potential, *Arch. Ration. Mech. Anal.* 171 (2), 151–218 (2004)

[35] M. Hitrik and K. Pravda-Starov. Spectra and Semigroup smoothing for Non-Elliptic Quadratic Operators, *Math. Ann.* 344, No 4 (2009), pp 801–846.

[36] L. Hörmander. Hypoelliptic second order differential equations. *Acta Math.*, 119:147–171, 1967.

[37] L. Hörmander, *The analysis of linear partial differential operators*, vol. I,II,III,IV, Springer-Verlag, New York (1985)

[38] V. Jakšić and C.-A. Pillet. Ergodic properties of the non-Markovian Langevin equation, *Lett. Math. Phys.*, 41(1): 49–57, 1997.

[39] V. Jakšić and C.-A. Pillet. Spectral theory of thermal relaxation, *J. Math. Phys.*, 38(4): 1757–1780, 1997. Quantum problems in condensed matter physics.

[40] V. Jakšić and C.-A. Pillet. Ergodic properties of classical dissipative systems. I. *Acta Math.*, 181(2): 245–282, 1998.

[41] I. Karatzas and S.E. Shreve. Brownian motion and stochastic calculus. Second edition. Graduate Texts in Mathematics, 113. *Springer-Verlag, New York*, 1991.

[42] H.A. Kramers. Brownian motion in a field of force and the diffusion model of chemical reactions, *Physica*, 7(1940), pp 284–304.

[43] R. Kupferman. Fractional kinetics in Kac-Zwanzig heat bath models. *J. Statist. Phys.*, 114(1-2):291–326, 2004.

[44] C. Landim. Central limit theorem for Markov processes. In *From classical to modern probability*, volume 54 of *Progr. Probab.*, pages 145–205. Birkhäuser, Basel, 2003.

[45] Lanford, III, O. E. Time evolution of large classical systems. In *Dynamical systems, theory and applications* (Recontres, Battelle Res. Inst., Seattle, Wash., 1974). Springer, Berlin, 1975, pp. 1–111. Lecture Notes in Phys., Vol. 38.

[46] J. Lebowitz. Microscopic reversibility and macroscopic behavior: physical explanations and mathematical derivations. In *Twentyfive years of non-equilibrium ststistical mechanics, proceedngs of the XIII Sitges conference (1994)*, J. Brey, J. Marro, J. Rubi and M. S. Miguel , Eds., Lect. Notes in Physics, Springer, pp.1-20.

[47] L. Lorenzi and M. Bertoldi. *Analytical Methods for Markov Semigroups*, CRC Press, New York (2006)

[48] A. Lunardi. On the Ornstein-Uhlenbeck Operator in L^2 Spaces with respect to invariant measures, *Transactions of the AMS* 349, No 1 (1997), pp 155–169.

[49] P. A. Markowich and C. Villani. On the trend to equilibrium for the Fokker-Planck equation: an interplay between physics and functional analysis, *Mat. Contemp.*, 19:1–29, 2000.

[50] J. C. Mattingly, A. M. Stuart and D. J. Higham. Ergodicity for SDEs and approximations: locally Lipschitz vector fields and degenerate noise. *Stochastic Process. Appl.*, 101(2): 185–232, 2002.

Geometric ergodicity of some hypoelliptic diffusions for particle motions, *Markov Processes and Related Fields*, 8(2): 199–214, 2002.

[51] G. Metafune, D. Pallara and E. Priola. *Spectrum of Ornstein-Uhlenbeck operators in L^p spaces with respect to invariant measures*, J. Funct. Anal. 196 (1), 40–60 (2002)

[52] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth and A.H. Teller. Equation of State Calculations by Fast Computing Machines. *Journal of chemical physics*, **21**, 6, June 1953.

[53] S.P. Meyn and R.L. Tweedie. *Markov chains and stochastic stability.* Communications and Control Engineering Series. Springer-Verlag London Ltd., London, 1993.

[54] H. Mori. A continued-fraction representation of the time-correlation functions. *Progress of Theoretical Physics*, 34(3): 399–416, 1965.

[55] E. Nelson. *Dynamical theories of Brownian motion.* Princeton University Press, Princeton, N.J., 1967.

[56] B. Øksendal. *Stochastic differential equations. An introduction with applications.* Sixth edition. Universitext. Springer-Verlag, Berlin, 2003.

[57] M. Ottobre. Random Walk in a Random Environment with non Diffusive Feedback Force: Long Time Behavior, accepted by *Stochastic proc. Appl.*

[58] M. Ottobre and G.A. Pavliotis, Asymptotic analysis for the Generalized Langevin Equation, *Nonlinearity* 24 (2011) 1629-1653

[59] M. Ottobre, G.A. Pavliotis and K. Pravda-Starov. Exponential return to equilibrium for hypoelliptic quadratic systems. *Journal of Functional Analysis*, 262 (2012) 4000-4039.

[60] G.C. Papanicolaou and S. R. S. Varadhan. Ornstein-Uhlenbeck process in a random potential, *Comm. Pure Appl. Math.*, 38(6): 819–834, 1985.

[61] G.A. Pavliotis and A.M. Stuart. *Multiscale methods*, volume 53 of *Texts in Applied Mathematics*. Springer, New York, 2008. Averaging and homogenization.

[62] A. Pazy. *Semigroups of linear operators and applications to partial differential equations*. Applied Mathematical Sciences, 44. Springer-Verlag, New York, 1983.

[63] C. Prevôt and M. Röckner. *A concise course on stochastic partial differential equations.* Lecture Notes in Mathematics, 1905. Springer, Berlin, 2007.

[64] M. Reed and B. Simon. *Methods of Modern Mathematical Physics I: Functional Analysis*. Academic Press, New York, 1980.

[65] L. Rey-Bellet and L. E. Thomas. Exponential convergence to non-equilibrium stationary states in classical statistical mechanics. *Comm. Math. Phys.*, 225(2):305–329, 2002.

[66] L. Rey-Bellet *Ergodic Properties of Markov Processes*, Quantum Open Systems II. The Markovian approach. Lecture Notes in Mathematics 1881 . Berlin: Springer, (2006) pp. 1–39

[67] C.P. Robert and G. Casella. Introducing Monte Carlo methods with R. Use R!. *Springer, New York*, 2010.

[68] D. Ruelle. Smooth Dynamics and new theoretical ideas in Nonequilibrium Statistical Mechanics, *J. Stat. Phys.*, Vol 95, Nos 1/2, 1999.

[69] J. Sjöstrand. Parametrices for pseudodifferential operators with multiple characteristics, Ark. Mat. (12), 85–130 (1974)

[70] D. W. Stroock. *Probability theory. An analytic view.* Second edition. Cambridge University Press, Cambridge, 2011.

[71] D.W. Stroock and S.R.S. Varadhan. *On the support of Diffusion processes with application to the strong maximum principle* Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability (Univ. California, Berkeley, Calif., 1970/1971), Vol. III: Probability theory, pp. 333–359. Univ. California Press, Berkeley, Calif., 1972.

[72] D.W. Stroock and S.R.S. Varadhan.*Multidimensional diffusion processes.* Springer, Berlin, 1979.

[73] H.J. Sussmann. An interpretation of stochastic differential equations as ordinary differential equations which depend on the sample point. *Bull. Amer. Math. Soc.* 83 (1977), no. 2, 296-298.

[74] L. Tierney. *A note on Metropolis-Hastings kernels for general state spaces.* Ann. Appl. Probab. 8 (1998), no. 1, 1–9.

[75] M. Turelli. Random environments and stochastic calculus. *Theoret. Population Biology* 12 (1977), no. 2, 140-178.

[76] K. Twardowska. *Wong-Zakai approximations for stochastic differential equations.* Acta Appl. Math. 43 (1996), no. 3, 317-359.

[77] C. Villani. Hypocoercivity. *Mem. Amer. Math. Soc.*, 202(950), 2009.

[78] D.Williams. To begin at the beginning:.... Proc. Sympos., Univ. Durham, Durham, 1980. pp. 1-55, Lecture Notes in Math., 851, Springer, Berlin-New York, 1981.

[79] D. Williams. *Probability with martingales.* Cambridge Mathematical Textbooks. Cambridge University Press, Cambridge, 1991.

[80] K. Yosida. *Functional analysis.* Classics in Mathematics. Springer-Verlag, Berlin, 1995.

[81] J. Zabczyk. *Mathematical control theory. An introduction.* Modern Birkhauser Classics, Birkhauser Boston MA, 1995.

[82] R. Zwanzig. Nonlinear generalized Langevin equations. *J. Stat. Phys.*, 9(3):215–220, 1973.

[83] R. Zwanzig . *Problems in nonlinear transport theory* in Systems far from equilibrium, L. Garrido, Springer, New York (1980), pp 198–225.

12 A Few Exercises

Warning: all the Markov chains mentioned in this exercise sheet are time-homogeneous Markov chain, unless otherwise stated.

1. Show that continuous functions are measurable.

2. Let $\{X_n\}_n$ be a sequence of i.i.d taking values in the set $\{1, \ldots, m\}$ and $\mathbb{P}(X_j = k) = p_k$, where $p_1, \ldots, p_k$ are positive numbers adding up to one. Let us fix one value $k \in \{1, \ldots, m\}$ and define

$$Y_j = \begin{cases} 1 & \text{if } X_j = k \\ 0 & \text{otherwise.} \end{cases}$$

   What can we say about $\bar{S}_n = \frac{1}{n}\sum_{j=1}^n Y_j$?

3. Let X_n be a sequence of i.i.d random variables with values in $\{0, 1\}$ and with $\mathbb{P}(X_n = 1) = p$ for some given $p \in [0, 1]$. Let $S_n = \sum_{j=1}^n X_j$. Find an upper bound on

$$\mathbb{P}\left(\left|\frac{S_n}{n} - p\right| > \epsilon\right).$$

   (I should have specified, not the trivial upper bound $\mathbb{P}\left(\left|\frac{S_n}{n} - p\right| > \epsilon\right) \leq 1$, but an upper bound that depends on ϵ).

4. Let $b(t)$ be standard Brownian Motion. Set $\xi(t) = e^{-t}b(e^{2t})$. Prove that ξ_t is a wide sense stationary process and find its autocorrelation function.

5. Prove that if $B(t)$ is a standard BM then $W(t) = \frac{1}{c}B(c^2 t)$ is a standard BM as well.

6. Consider the process $X_t = A\cos(\eta\, t + \varphi)$ where A and η are two random variables with arbitrary joint distribution while φ is uniformly distributed on $[0, 2\pi)$ and independent of both A and η. Prove that if A has finite first and second moment then X_t is wide sense stationary. All the above random variables are real valued. (If you want to make it a bit harder, prove that it is strictly stationary.)

7. **Bernoulli-Laplace model of diffusion.** There are two urns, say right and left, each of them containing m balls. Out of the $2m$ total balls, b are black and $2m-b$ are white. We will assume $b \leq m$. At each

time we take one ball from each urn and we exchange them. The state of the system can be described, for example, by keeping track of the number k of black balls in the left urn. Find the transition probability of the Markov chain that describes the evolution of the number k.

8. **Simple Random Walk.** Let ξ_i be i.i.d. random variables taking values in $\{-1, 1\}$. In particular $\mathbb{P}(\xi_i = 1) = p$ and $\mathbb{P}(\xi_i = -1) = 1 - p$, for all i and for some $0 < p < 1$. Let $\Phi_n = \sum_{i=1}^{n} \xi_i$. Find the transition probabilities of the Markov chain Φ_n.

9. Give an example of a set (and a chain) which is closed but not irreducible and, viceversa, of a set that is irreducible but not closed.

10. Show that in the Ehrenfest chain all states are recurrent.

11. Prove Lemma 3.30.

12. Let X_n be a time-homogeneous MC on a finite state space S. Prove that if the state space S is irreducible then there exists a unique stationary distribution π. Write an expression for π.

13. **Queue**. As we have all experienced at least once, a queue is a line where customers wait for services. Suppose that at most n people are allowed in the queue, i.e. at each moment in time there can be at most n people in the queue. Rules of the queue:

    - If the queue has strictly less than n customers then with probability γ a new customer joins the queue.
    - If the queue is not empty, then with probability λ the head of the line is served and leaves the queue.

    From which we deduce that the queue remains unchanged

    - with probability $1 - \gamma$ if it is empty
    - with probability $1 - \lambda$ it is full (i.e. n people waiting)
    - with probability $1 - \gamma - \lambda$ otherwise.

    All other possibilities occur with zero probability. It should be now clear that this situation ca be modelled using a time-homogeneous MC with finite state space $S = \{0, 1, \ldots, n\}$.

    (a) Write down the transition probabilities of the chain

    (b) Prove that the chain has a unique stationary distribution.

(c) Find the stationary distribution π by solving the system of equations

$$\sum_{k \in S} \pi(k) p(k, j) = \pi(j)$$

and then imposing the normalization condition.

14. Prove that for a transient (time-homogeneous) MC the limiting behaviour of the transition probabilities is $p^n(x, y) \to 0$.

15. Prove, as a consequence of Theorem 3.35, that if y is a recurrent state then

$$\frac{1}{n} \sum_{j=1}^{n} p^j(x, y) \to \frac{\rho_{xy}}{\mathbb{E}_y T_y}.$$

Hint: use the bounded convergence theorem.

16. Consider a time-homogeneous Markov chain on a finite state space S and let $P = (p(x, y))_{x,y \in S}$ be the transition matrix of the chain. The transition matrix, and hence the chain, is said to be *regular* if there exists a positive integer $k > 0$ such that $p^k(x, y) > 0$ for all $x, y \in S$. Clearly a regular Markov chain is irreducible. Consider a MC on a finite state space S and prove the following: if for any x and y in S there exists an integer $n > 0$ such that $p^n(x, y) > 0$ and there exists $z \in S$ such that $p(z, z) > 0$ then the chain is regular. (Notice that k is independent of x and y whereas $n = n(x, y)$ i.e. it depends on the choice of x and y.)

17. Regular chains on finite state spaces are very important. Indeed if X_n is a regular chain on a finite state space then the chain has exactly one stationary distribution, π, and

$$\lim_{n \to \infty} p^n(x, y) = \pi(y), \quad \text{for all } x \text{ and } y \in S.$$

Consider the setting and notation of Example 4.1. Use Exercise 16 and the above statement to prove that if π is not the uniform distribution on S then the chain with transition matrix P converges to π.
Hint: show that there exist two states $a, b \in S$ such that $q(a, b) > 0$ and $\pi(b) < \pi(a)$. Then look at $p(a, a)$.

18. Show that the chain produced by Algorithm 4.6 satisfies the detailed balance condition, i.e. it is π-reversible.

19. Consider the accept-reject method, Algorithm 4.4. Calculate the probability of acceptance of the sample $Y \sim \nu$ at each iteration. Let N be the number of iterations needed before successfully producing a sample from the distribution π, i.e. the number of rejections before an acceptance. Calculate $\mathbb{P}(N = n)$ for all $n \geq 1$ and the expected value of N as a function of M.
Hint: N is a random variable with geometric distribution.

20. Write down the Metropolis-Hastings algorithm to simulate the normal distribution $\mathcal{N}(0, 1)$ using a proposal which is uniformly distributed on $[-\delta, \delta]$, for an arbitrary $\delta > 0$. (As you can imagine, in computational practice it is important to choose a good delta).

21. Let X_t be a continuous time Markov process. If the process is not time-homogeneous then the transition probabilities are functions of four arguments:

$$p(s, w, t, A) : \mathbb{R}_+ \times S \times \mathbb{R}_+ \times \mathcal{S} \longrightarrow [0, 1], \qquad 0 \leq s \leq t,$$

with

$$p(s, w, t, A) = \mathbb{P}(X_t \in A | X_s = w).$$

Any Markov process can be turned into a time-homogeneous one, if we consider an extended state space: suppose X_t with state space S is not time-homogeneous and show that the process $Y_t = (t, X_t)$ with state space $\mathbb{R}_+ \times S$ is instead time homogeneous (Hint: just calculate the transition probabilities of Y_t).

22. Let $\xi(t)$ be the process defined in Exercise 4. $\xi(t)$ is called the stationary Ornstein-Uhlenbeck process (O-U process). $\xi(t)$ is a Markov process and the transition probabilities of such a process have a density. Show that

$$\mathbb{P}(\xi_t = y | \xi_s = x) = \frac{1}{\sqrt{2\pi(1 - e^{-2(t-s)})}} \exp\left\{ -\frac{\left| y - xe^{-(t-s)} \right|^2}{2(1 - e^{-2(t-s)})} \right\}.$$

23. Prove that if the coefficients b and σ of an Itô SDE (satisfying the conditions of the existence and uniqueness theorem) do not depend on time, then the solution of the SDE is a time-homogeneous Markov process.

24. Show, using the definition, that

$$\int_0^t s\, dW_s = tW_t - \int_0^t W_s\, ds\,.$$

25. Use the Itô formula to find the SDE (in the form (45)) satisfied by the following processes

$$Y_t = 2 + t + e^{W_t} \quad \text{and} \quad Z_t = W_t^2 + B_t^2,$$

where W_t and B_t are two independent one dimensional Brownians.

26. Calculate

$$\int_0^t W_s^2\, dW_s\,.$$

27. Consider the OU process

$$dX_t = -bX_t + \sigma dW_t, \quad \text{with } b, \sigma > 0.$$

Suppose $X_0 \sim \mathcal{N}(0, \frac{\sigma^2}{2b})$. Calculate the autocovariance function of X_t. Write the equation satisfied by $|X_t|^2$ and hence the equation for $\mathbb{E}\,|X_t|^2$. Using Gronwall's Lemma estimate $\mathbb{E}\,|X_t|^2$. Now suppose X_0 is deterministic and calculate the covariance function of X_t.

28. Consider the geometric BM of Example 8.11. Assuming that the initial datum X_0 is independent of W_t, calculate the expected value of X_t.

29. Verify that $(X_1(t), X_2(t)) = (t, e^t B_t)$ solves

$$\begin{aligned} dX_1 &= dt \\ dX_2 &= X_2\, dt + e^{X_1} dB_t \end{aligned}$$

and that $(X_1(t), X_2(t)) = (\cosh B_t, , \sinh B_t)$ solves

$$\begin{aligned} dX_1 &= \frac{1}{2} X_1 dt + X_2 dB_t \\ dX_2 &= \frac{1}{2} X_2 dt + X_1 dB_t \end{aligned}$$

where in all the above B_t is a one dimensional standard BM.

30. In this exercise I will not be very precise in defining the types of convergence involved, so I do not expect you to be rigorous in the solution, as far as the type of convergence is concerned. Consider a sequence of stochastic processes $\xi^k(t)$ such that

(a) for each $k \in \mathbb{N}$, ξ_t^k is a Gaussian process with $\mathbb{E}\xi^k(t) = 0$ and $\mathbb{E}(\xi_t^k \xi_s^k) = \mathfrak{d}^k(t-s)$, where $\mathfrak{d}^k$ is a sequence of smooth functions that converge to the Dirac delta, $\mathfrak{d}^k \to \delta_0$ as $k \to \infty$;

(b) for each $k \in \mathbb{N}$, the paths $t \to \xi_t^k$ are smooth for all ω. The sequence of processes ξ_t^k constitutes a smooth approximation to white noise ξ_t, i.e. in some sense $\xi_t^k \to \xi_t$ as $k \to \infty$.

Now look at the Itô equation (66), namely

$$dX_t = d(t)X_t\,dt + f(t)X_t\,dW_t\,, \quad X(0) = X_0\,, \tag{136}$$

the solution of which we have found to be given by equation (67).

- Consider the Stratonovich equation

$$d\tilde{X}_t = d(t)\tilde{X}_t\,dt + f(t)\tilde{X}_t \circ dW_t\,, \quad X(0) = X_0\,, \tag{137}$$

  and solve it (in order to do so you can first transform it in Itô form and then use one of the presented methods of solution).

- Now look at the equation

$$\dot{X}_t^k = d(t)X_t^k dt + f(t)X_t^k \xi_t^k, \quad X^k(0) = X_0\,,\, k \in \mathbb{N} \tag{138}$$

  This is equation (136), when we replace white noise with the smooth approximants. Therefore, for each $k \in \mathbb{N}$ and for each ω, this is a simple ODE. Solve it.

- The solution of (138) contains the integral $I^k(t) = \int_0^t f(s)\xi^k(s)$. Observe that $I^k(t)$ is Gaussian with mean zero. Calculate $\mathbb{E}(I_t^k I_s^k)$. With a formal calculation (assume that you can exchange limit and integral) show that $\mathbb{E}(I_t^k I_s^k) \to \mathbb{E}(I_t I_s)$, where I_t is the Itô integral $I_t = \int_0^t f(s)dW_s\,.$

- Deduce that the solution of (138) converges to the solution of (137) (better, to a process which has the same distributions as (137)).

31. Let $\mathcal{P}_t$ be a C_0-semigroup on a Banach space $\mathfrak{B}$. Show that there exist constants $\omega \geq 0$ and $M \geq 1$ such that

$$\|\mathcal{P}_t\|_{B(\mathfrak{B})} \leq Me^{\omega t}\,.$$

Hint: 1. Assume the following statement: if $\mathcal{P}_t$ is a C_0-semigroup then there exists $t_0 > 0$ such that $\|\mathcal{P}_t\|_{B(\mathfrak{B})} \leq M$ for all $0 \leq t \leq t_0$.
2. Observe that $M \geq 1$ (why?)
3. Set $\omega := (\log M)/t_0$.

32. Let $\mathcal{P}_t$ be the semigroup defined in Example 9.3. Find the generator of $\mathcal{P}_t$ in the space $\mathcal{C}_b(\Omega)$.

33. Find the generator of the semigroup

$$T_t f(x) = \begin{cases} f(x+t) & \text{if } x+t \leq 1 \\ 0 & \text{if } x+t > 1 \end{cases}$$

defined on the space $C^1[0,1]$.

34. Show that the semigrop $\mathcal{P}_t$ associated to a time-homogeneous Markov process is a contraction semigroup in L^∞, i.e. $\|P_t f\|_\infty \leq \|f\|_\infty$.

35. Let X_t be a real valued wide sense stationary process and denote by $C(t,s) = C(t-s) = \mathbb{E}(X_t X_s) - \mu^2$, $\mu := \mathbb{E}(X_t)$ its covariance function. Assume that $C(t) \in L^1(\mathbb{R}_+)$, i.e. $\int_{R_+} C(s)ds < \infty$. Show that the process is ergodic in the following sense:

$$\lim_{T\to\infty} \mathbb{E}\left| \frac{1}{T}\int_0^T X_s\, ds - \mu \right|^2 = 0\,.$$

36. Transform the following Stratonovich SDEs into Itô SDEs or viceversa.

- $X_t = X_0 + \int_0^t X_s^2 ds + \int_0^t \cos(X_s) \circ dW_s$
- $dX_t = \sinh(X_t)dt + t^2 \circ dW_t$
- $dX_t = X_t dt + 3X_t dW_t$

37. Prove Lemma 10.7. Hint: Under the assumptions of Lemma 10.7,

$$\lim_{\delta\to 0}\int_t^{t+\delta} \mathbb{E}f(s,X_s)ds = \lim_{\delta\to 0}\int_{t-\delta}^{t} \mathbb{E}f(s,X_s)ds = f(t,x),$$

where X_u is the solution of

$$dX_u = b(u,X_u)du + \sigma(u,X_u)dW_u, \quad X_t = x\,.$$

Use this hint to prove the lemma. If you want to prove also the statement of the hint then you can use the same kind of arguments that we used in the proof of Theorem 10.4.

38. Prove that (106) $\Rightarrow$ (105) (work with mean zero functions).

39. Consider the process $X_t \in \mathbb{R}^d$ satisfying

$$dX_t = -\nabla V(X_t)dt + \sqrt{2}dW_t$$

where W_t is d-dimensional standard Brownian motion. Show that the measure ρ with density $\rho(x) = e^{-V(x)}/\mathcal{Z}$ (here $\mathcal{Z}$ is a normalizing constant) is the only invariant measure for X_t. Let $v : \mathbb{R}^d \to \mathbb{R}^d$ be a smooth vector valued function, $v(x) = (v^i(x))_{i=1,\dots,d}$. Show that the invariant measure of the process

$$dX_t^v = -(v(X_t^v) + \nabla V(X_t^v))dt + \sqrt{2}dW_t,$$

is still ρ if and only if

$$\nabla \cdot (v\rho) = 0,$$

where $\nabla\cdot$ denotes divergence.

40. Let $\mathcal{P}_t$ be a semigroup of bounded operators on a Banach space $\mathfrak{B}$. Show that the condition

$$\lim_{t\downarrow 0} \mathcal{P}_t u = u \quad \text{for all } u \in \mathfrak{B} \tag{139}$$

is equivalent to the map $\mathbb{R}_+ \ni t \longrightarrow \mathcal{P}_t u$ being continuous for every fixed $u \in \mathfrak{B}$.

41. Prove the following:

    (a) Let A be a densely defined operator on a Hilbert space. Then iA is self adjoint iff A is skew adjoint.

    (b) If A is the generator of a $\mathcal{C}_0$ group of unitary operators on a Hilbert space then iA is self adjoint.

42. Use the antisymmetry (in L^2_ρ) of the operator $\mathcal{L}_H$ defined in (116) to prove that the semigroup generated by $\mathcal{L}_H$ is norm preserving.

43. Consider the one dimensional process

$$dX_t = -X_t^2 dt + \sigma dW_t, \quad \sigma > 0.$$

    (a) Recognize that this equation is of the form (101), for an appropriate function $V(x)$. Write $V(x)$ (you may assume $V(0) = 0$).

    (b) Write the generator $\mathcal{L}$ of the process and its invariant density ρ(i.e. the density of the invariant measure).

(c) Check that $\mathcal{L}$ is symmetric in L^2_ρ.

44. Consider the following one dimensional SDEs:

$$dX_t = -X_t dt + \sigma dW_t, \quad dY_t = rY_t dt + 2dB_t$$

where σ and r are a strictly positive real numbers and W_t and B_t are one dimensional independent standard Brownian motions. Write the equation satisfied by

(a) $Z_t := X_t Y_t$;

(b) $g(X_t, Y_t)$, where g is a twice differentiable function $g(x, y) : \mathbb{R}^2 \to \mathbb{R}$;

(c) $V_t := X_t^2 Y_t$.

45. Let $\mathcal{P}_t$ be a Markov semigroup on a Banach space of real valued functions, $\mathcal{L}$ the generator of the semigroup and μ a reversible measure (on $\mathbb{R}$, and admitting density with respect to the Lebesgue measure) satisfying the spectral gap inequality with constant α. Let $U(x)$ be a bounded (above and below) and measurable function on $\mathbb{R}$ and define a new measure ν with density

$$\nu(x) = \frac{1}{\mathcal{Z}} e^{-U(x)} \mu(x),$$

where $\mathcal{Z}$ is the normalizing constant. Notice that, due to the boundedness of U, if $f \in L^2_\mu$ then $f \in L^2_\nu$ as well (same for integrability) and also the other way around.

(a) Prove that for any constant $a \in \mathbb{R}$,

$$\int_{\mathbb{R}} \left(f - \int_{\mathbb{R}} f d\nu \right)^2 d\nu \leq \int_{\mathbb{R}} (f - a)^2 \, d\nu \,,$$

and for any $f \in L^2_\nu$.

(b) Prove that ν satisfies a spectral gap inequality with constant $\alpha e^{-2Osc(U)}$ where $Osc(U) = \sup U - \inf U$. Hint: use the fact that for any Borel set $A \subset \mathbb{R}$,

$$e^{-Osc(U)} \mu(A) \leq \nu(A) \leq e^{Osc(U)} \mu(A) \,.$$

46. Solve the following SDEs

- $dY_t = rdt + \alpha Y_t dB_t$ with initial condition $Y(0) = Y_0$;
- $dX_t = -X_t dt + e^{-t} dB_t$ with initial condition $X(0) = X_0$;
- $dX_t = 2X_t dt + 4X_t dB_t$ with $X_0 = 3$. For this last equation, calculate also $\mathbb{E}(X_t)$ (you can use the method explained in the solution of Exercise 28).

13 Solutions

Some of the following solutions are a bit sketchy.

1. By definition, the preimage of an open set is an open set.

2. The r.v. Y_j are i.i.d (because they are measurable functions of the X_j's) and integrable. Therefore

$$\bar{S}_n \xrightarrow{a.s.} \mathbb{E}(Y_1) = \mathbb{E}(\mathbf{1}_{(X_1=k)}) = \mathbb{P}(X_1 = k) = p_k.$$

3. $\mathbb{E}X_n = p$ and $Var(X_n) = p(1-p) \leq 1/4$. Therefore $\mathbb{E}(S_n/n) = p$ and $Var(S_n/n) = p(1-p)/n \leq 1/(4n)$. Using Chebyshev's inequality, the probability that we wanted to estimate is bounded above by $1/(4n\epsilon^2)$.

4. Clearly, $\mathbb{E}(\xi(t)) = 0$. As for the autocovariance function, suppose $t > s$. Then

$$\mathbb{E}(\xi(t)\xi(s)) = \mathbb{E}(e^{-t}b(e^{2t})e^{-s}b(e^{2s})) = e^{-(t+s)}e^{2s} = e^{-(t-s)},$$

   having used the fact that $\mathbb{E}(b(t)b(s)) = t \wedge s$. Hence for any $t, s > 0$ we have $\mathbb{E}(\xi(t)\xi(s)) = e^{-|t-s|}$.

5. We need to check that the definition of Example 2.7 is satisfied by $W(t)$. i) is trivial. $W(t) - W(s) = (B(c^2t) - B(c^2s))/c$ is Gaussian as it is the sum of Gaussians. It is mean zero and, assuming $s \leq t$, its variance is

$$\begin{aligned}\mathbb{E}(W(t) - W(s))^2 &= \mathbb{E}[(B(c^2t) - B(c^2s))/c]^2 \\ &= \frac{1}{c^2}[\mathbb{E}B(c^2t) + \mathbb{E}B(c^2s) - 2\mathbb{E}B(c^2t)B(c^2s)] \\ &= \frac{1}{c^2}(c^2t + c^2s - 2c^2s) = t - s\,.\end{aligned}$$

   As for iii), it follows after observing that if (a, b) doesn't overlap (d, e) then the same holds for (c^2a, c^2b) and (c^2d, c^2e).

6. First

$$\begin{aligned}\mathbb{E}(A\cos(\eta t + \varphi)) &= \mathbb{E}(A\cos(\eta t)\cos\varphi) - \mathbb{E}(A\sin(\eta t)\sin\varphi) \\ &= \mathbb{E}(A\cos(\eta t))\mathbb{E}(\cos\varphi) - \mathbb{E}(A\sin(\eta t))\mathbb{E}(\sin\varphi).\end{aligned}$$

   Since $\mathbb{E}(\cos\varphi) = \frac{1}{2\pi}\int_0^{2\pi} coszdz = 0$ and analogously $\mathbb{E}(\sin\varphi) = 0$ we have that $\mathbb{E}X_t = 0$ if $\mathbb{E}(A\cos(\eta t)) < \infty$ and $\mathbb{E}(A\sin(\eta t)) < \infty$.

To check that this is the case, recalling that for the joint probability density of A and η we have $f_{A,\eta}(x,y) = f_{\eta|A}(y|x)f_A(x)$, we can write

$$\begin{aligned}\mathbb{E}(A\cos(\eta t)) &= \iint x\cos(yt) f_{A,\eta}(x,y)dxdy \\ &\leq \iint |x|\, f_{\eta|A}(y|x)f_A(x)dxdy \\ &= \int |x|\, f_A(x)dx \int f_{\eta|A}(y|x)dy = \int |x|\, f_A(x)dx,\end{aligned}$$

where in the last step we used $\int f_{\eta|A}(y|x)dy = 1$ for all x. Therefore $\mathbb{E}(A\cos(\eta t)) < \infty$ if A has finite first moment and the same thing can be checked for $\mathbb{E}(A\sin(\eta t))$. In the same way,

$$\begin{aligned}\mathbb{E}(X_t X_s) &= \mathbb{E}(A^2\cos(\eta t+\varphi)\cos(\eta s+\varphi)) \\ &= \frac{1}{2}\mathbb{E}(A^2\cos(\eta(t+s)+2\varphi) + A^2\cos\eta(t-s)).\end{aligned}$$

With steps analogous to what we have done before one can check that $\mathbb{E}(A^2\cos(\eta(t+s)+2\varphi)) = 0$ if A has finite second moment. Therefore $\mathbb{E}(X_tX_s) = \mathbb{E}(A^2\cos\eta(t-s))$ and hence it is a function of $t-s$ only.

7. If w_l is the number of white balls in the left urn, and similarly for $w_r, b_l = k$ and b_r, the transition probabilities are as follows:

$$\begin{aligned}p(k,k+1) &= \frac{w_l}{m}\frac{b_r}{m} = \frac{m-k}{m}\ \frac{b-k}{m} \\ p(k,k-1) &= \frac{b_l}{m}\frac{w_r}{m} = \frac{k}{m}\ \frac{m-b+k}{m} \\ p(k,k) &= \frac{b_l}{m}\frac{b_r}{m} + \frac{w_l}{m}\frac{w_r}{m} = \frac{k}{m}\ \frac{b-k}{m} + \frac{m-k}{m}\ \frac{m-b+k}{m}.\end{aligned}$$

8. $p(j,k) = p$ if $k = j+1$ and $p(j,k) = 1-p$ if $k = j-1$. Otherwise $p(j,k) = 0$.

9. Suppose we have a two-state chain, i.e. $S = \{a,b\}$, with

$$p = \begin{vmatrix} 0 & 1 \\ 0 & 1 \end{vmatrix}.$$

Then S is closed but not irreducible because there is no way of going from b to a. Notice that the state b is recurrent, in particular it is

absorbent.
Now let $S = \{1, 2, 3\}$ and define

$$p = \begin{vmatrix} 0 & 1 & 0 \\ 0.5 & 0 & 0.5 \\ 0 & 0 & 1 \end{vmatrix}.$$

Then the set $A = \{1, 2\}$ is irreducible but not closed as $2 \in A$ and $\rho_{23} > 0$ but $3 \notin A$.

10. Because we have only $N + 1$ possible states, this is easily done by drawing the chain

$$0 \underset{\leftarrow}{\rightarrow} 1 \underset{\leftarrow}{\rightarrow} 2 \underset{\leftarrow}{\rightarrow} 3 \underset{\leftarrow}{\rightarrow} \ldots \underset{\leftarrow}{\rightarrow} N$$

The state space S is finite, closed and irreducible, therefore every state is recurrent. If you didn't want to sketch the graph of the chain, you could have observed that, from the definition of the transition probabilities, every state communicates with each of its two nearest neighbours (except 0 and N, that communicate only with 1 and $N-1$, respectively). Therefore S is (finite) closed and irreducible. If you really want to complicate your life you could do it by induction on N.

11. If π is a stationary distribution then

$$\sum_y \pi(y) p^n(y, x) = \pi(x).$$

Taking sums over n on both sides we get

$$\sum_y \pi(y) \frac{\rho_{yx}}{1 - \rho_{xx}} = \infty.$$

However $\rho_{yx} \leq 1$, so

$$\infty = \sum_y \pi(y) \frac{\rho_{yx}}{1 - \rho_{xx}} \leq \frac{1}{1 - \rho_{xx}} \sum_y \pi(y) = \frac{1}{1 - \rho_{xx}},$$

so it has to be $\rho_{xx} = 1$.

12. S is finite and irreducible. If the whole state space S is irreducible, then it is also closed - this is not true for subsets of S, as we have seen in Exercise 9. Therefore all the states are recurrent. Therefore

stationary distributions are constant multiples of each others. Being the chain recurrent, we also know that $\mu_x(y) = \sum_{n=0}^{T_x-1} p^n(x,y)$ is a stationary measure. However, due to the finiteness of S, $\sum_y \mu_x(y) < \infty$. Therefore $\pi(y) = \mu_x(y)/\sum_y \mu_x(y)$ (which does not depend on the choice of x, why?).

13. (a) Transition probabilities

$$\begin{aligned} p(k,k+1) &= \gamma \quad \text{if } k < n \\ p(k,k-1) &= \lambda \quad \text{if } k > 0, \end{aligned}$$

and also

$$p(k,k) = \begin{cases} 1-\gamma & \text{if } k = 0 \\ 1-\lambda & \text{if } k = n \\ 1-\gamma-\lambda & \text{if } 1 \le k \le n-1\,. \end{cases}$$

Otherwise $p(j,k) = 0$.
(b) Sketching a graph of the chain shows immediately that all the states communicate with each other, so the chain is closed and irreducible, which on a finite state space implies that the chain has only one stationary measure.
(c) Expanding the system of equations gives

$$\begin{aligned} \pi(0) &= (1-\gamma)\,\pi(0) + \lambda\,\pi(1) \\ \pi(k) &= \gamma\,\pi(k-1) + (1-\lambda-\gamma)\,\pi(k) + \lambda\,\pi(k+1) \quad 1 \le k \le n-1 \\ \pi(n) &= \gamma\,\pi_{n-1} + (1-\lambda)\,\pi(n)\,. \end{aligned}$$

This system has clearly infinitely many solutions (all multiples of each others) so we use $\pi(0)$ as a parameter. A solution is $\pi(k) = \pi(0)\,(\gamma/\lambda)^k$. Now impose the normalization condition and get

$$\pi(k) = \frac{(\gamma/\lambda)^k}{\sum_{i=0}^{n}(\gamma/\lambda)^i}.$$

14. If y is transient then $\sum_n p^n(x,y) < \infty \Rightarrow p^n(x,y) \to 0$.
$\sum_n p^n(x,y) < \infty$ because

$$\sum_{n=1}^{\infty} p^n(x,y) = \mathbb{E}_x N(y) = \frac{\rho_{xy}}{1-\rho_{yy}} < \infty\,.$$

15. $0 \leq N(y)/n \leq 1$ so you can use the bounded convergence theorem which gives

$$\mathbb{E}_x \left| \frac{N_n(y)}{n} - \frac{1}{\mathbb{E}_y T_y} \mathbf{1}_{(T_y<\infty)} \right| \longrightarrow 0$$

and hence also

$$\mathbb{E}_x \frac{N_n(y)}{n} \longrightarrow \mathbb{E}_x \left[\frac{1}{\mathbb{E}_y T_y} \mathbf{1}_{(T_y<\infty)} \right] .$$

16. Let $M = \max_{x,w \in S} n(x,w)$, then for any $x, y \in S$ we have

$$p^{2M}(x,y) \geq p^{n(x,z)}(x,z) \cdot \underbrace{p(z,z) \cdot \ldots \cdot p(z,z)}_{2M-n(x,z)-n(z,y) \text{ times}} \cdot p^{n(z,y)}(z,y) > 0.$$

So $k = 2M$ is the integer we were after. In the above we need to consider $k = 2M$ instead of just M to make sure that $k - n(x,z) - n(z,y) > 0$.

17. We want to prove that the chain with transition matrix P is regular. To this end we will show that the sufficient conditions of Exercise 16 are satisfied (unless $x \to \pi(x)$ is constant i.e. unless π is the uniform distribution). Recall that Q is irreducible hence P is irreducible as well, therefore it is true that for all x, y there exists $n = n(x,y) > 0$ such that $p^{n(x,y)}(x,y) > 0$. Therefore we only need to find a state $a \in S$ such that $p(a,a) > 0$. Let M be the set $M = \{x \in S : \pi(x) = \max_{y \in S} \pi(y)\}$. Because Q is irreducible there exist $a \in M$ and $b \in M^c$ such that $q(a,b) > 0$ and clearly by construction $\pi(a) > \pi(b)$. Notice also that from the definition of P, $p(x,y) \leq q(x,y)$ for all $x \neq y$. Then

$$\begin{aligned}
p(a,a) &= 1 - \sum_{x \neq a} p(a,x) = 1 - \sum_{x \neq a,b} p(a,x) - p(a,b) \\
&\geq 1 - \sum_{x \neq a,b} q(a,x) - q(a,b)\pi(b)/\pi(a) \\
&= 1 - \sum_{x \neq a} q(a,x) + q(a,b)\left[1 - \pi(b)/\pi(a)\right] \\
&= q(a,a) + q(a,b)\left[1 - \pi(b)/\pi(a)\right] \geq q(a,b)\left[1 - \pi(b)/\pi(a)\right] > 0.
\end{aligned}$$

18. The transition kernel of the chain produced with the M-H algorithm can be written as

$$p(x,y) = q(x,y)\alpha(x,y) + \delta_x(y) \int_{\mathbb{R}^N} (1 - \alpha(x,w)) q(x,w) dw.$$

The detailed balance condition for p and π reads $\pi(x)p(x,y) = \pi(y)p(y,x)$ and is hence satisfied if and only if $\pi(x)q(x,y)\alpha(x,y) = \pi(y)q(y,x)\alpha(y,x)$. Setting $\mu(x,y) = \pi(x)q(x,y)$, what we want to show is that $\mu(x,y)\alpha(x,y) = \mu(y,x)\alpha(y,x)$. The latter equality is easily verified:

$$\begin{aligned}\mu(x,y)\alpha(x,y) &= \mu(x,y)\min\{1, \mu(y,x)/\mu(x,y)\} = \min\{\mu(x,y), \mu(y,x)\} \\ &= \min\{1, \mu(x,y)/\mu(y,x)\}\mu(y,x) = \alpha(y,x)\mu(y,x).\end{aligned}$$

19. The probability of acceptance is $a = \mathbb{P}(U \leq \pi(Y)/M\nu(Y))$, which can be calculated by using the law of total probability:

$$\begin{aligned}\mathbb{P}(U \leq \pi(Y)/M\nu(Y)) &= \int_{\mathbb{R}} \mathbb{P}(U \leq \pi(Y)/M\nu(Y)|Y = y)\, \nu(y)\, dy \\ &= \int \frac{\pi(y)}{M\nu(y)}\nu(y)\, dy = 1/M,\end{aligned}$$

having used the fact that π is a probability density function. The random variable N is geometrically distributed, with probability of success at each trial equal to a. Then $\mathbb{P}(N = n) = (1-a)^{n-1}a$ and $\mathbb{E}(N) = 1/a = M$.

20. This is a M-H algorithm with acceptance probability $\alpha(x,y) = \min\{\exp[(x^2 - y^2)/2], 1\}$ and proposal kernel $q(x,\cdot) \sim \mathcal{U}(x-\delta, x+\delta)$.

21. Recall the the measure P on sequence space obtained via the extension theorem is

$$P(\omega : \omega_i \in A_i, i = 0, \ldots, N) = \mathbb{P}(X_0 \in A_1, \ldots, X_N \in A_N),$$

where $\mathbb{P}$ is the probability measure on the space Ω where the sequence X_n is defined and the A_i's. are Borel sets of $\mathbb{R}$. Let us show the claim for the simplest cylinder set: let $C = \{\omega \in \mathbb{R}^{\mathbb{N}} : \omega_0 \in A\}$. Then $\varphi^{-1}(C) = \{\omega \in \mathbb{R}^{\mathbb{N}} : \omega_1 \in A\}$. Therefore

$$P(C) = \mathbb{P}(X_0 \in A) \overset{\text{stationarity}}{=} \mathbb{P}(X_1 \in A) = P(\varphi^{-1}(C)).$$

Now you can detail the proof for every set in $\mathcal{R}^{\mathbb{N}}$.

22. Let $x = (s,y) \in \mathbb{R}_+ \times S$ and $B = C \times A \in \mathcal{B}(\mathbb{R}_+) \times \mathcal{S}$. Then the transition function of the process (t, X_t) is simply

$$q_t(x,B) = p(s,y,t+s,A)\mathbf{1}_C(t+s)\,.$$

I suppose you will need to give this a thought.

23. Sketch: using the properties of Brownian motion,

$$\begin{aligned}\mathbb{P}(\xi_t \le y|\xi_s = x) &= \mathbb{P}(e^{-t}W(e^{2t}) \le y|e^{-s}W(e^{2s}) = x)\\ &= \mathbb{P}(W(e^{2t}) \le e^t y|W(e^{2s}) = e^s x)\\ &= \int_{-\infty}^{e^t y} \frac{1}{\sqrt{2\pi(e^{2t}-e^{2s})}} \exp\left\{-\frac{z-xe^s}{2(e^{2t}-e^{2s})}\right\} dz\,.\end{aligned}$$

Now use a change of variables and conclude by observing that

$$\mathbb{P}(\xi_t = y|\xi_s = x) = \frac{\partial}{\partial y}\mathbb{P}(\xi_t \le y|\xi_s = x)\,.$$

24. The Markovianity is a consequence of Theorem 8.13, so we only need to prove time-homogeneity, which means that we need to prove that

$$\mathbb{P}(X_u \in B|X_0 = x) = \mathbb{P}(X_{u+t} \in B|X_t = x), \quad \forall B \in \mathcal{S},\, x \in S, t \ge 0\,.$$

Denote by $X^{x,t}(u+t)$ the solution of the equation

$$\beta(t+u) = x + \int_t^{t+u} b(\beta(s))ds + \int_t^{t+u} \sigma(\beta(s))dW_s \tag{140}$$

and by $X^{x,0}(u)$ the solution of the equation

$$\beta(u) = x + \int_0^u b(\beta(s))ds + \int_0^u \sigma(\beta(s))dW_s\,. \tag{141}$$

What we want to prove is that $X^{x,t}(u+t)$ has the same distribution as $X^{x,0}(u)$. To this end, notice that if W_v is a standard BM then $W_{t+v} - W_t$ is a standard BM as well (check), i.e. $\tilde{W}_v = W_{t+v} - W_t$ has the same distribution as W_v. With this in mind, a simple change of variable concludes the proof. Indeed, the RHS of (140) can be rewritten as

$$\beta(t+u) = x + \int_0^u b(\beta(v+t))dv + \int_0^u \sigma(\beta(v+t))dW_{t+v}\,.$$

At this point, because $\tilde{W}$ and W have the same distribution, and the above is nothing but (141), when we replace W with $\tilde{W}$, it is clear by the uniqueness of the solution that $\beta(t+u)$ has the same distribution as $\beta(u)$, which is, $X^{x,t}(u+t)$ has the same distribution as $X^{x,0}(u)$.

25. Use $s_j(\Delta W_{s_j}) = \Delta(s_j W_{s_j}) - W_{s_{j+1}}\Delta s_j$.

26. Apply Itô formula with $g(t,x) = 2 + t + e^x$ and get

$$dY_t = (1 + \frac{1}{2}e^{W_t})dt + e^{W_t}dW_t\,.$$

For Z_t instead you get

$$dZ_t = 2dt + 2BdB + 2WdW\,.$$

27. For example by integration by parts: set $X_1 = W^2, X_2 = W$. Applying Itô 's rule we have $d(W^2) = 2WdW + dt$ and using the multiplication table $d(W^2)dW = 2Wdt$. Therefore

$$\int_0^t W^2 dW = W_t^3 - \int_0^t 2W^2 dW - \int_0^t W ds - \int_0^t 2W ds\,,$$

so that readjusting

$$\int_0^t W^2 dW = \frac{W_t^3}{3} - \int_0^t W ds\,.$$

28. Before giving the solution I would like to Remark that we will show that $\mathcal{N}(0, \sigma^2/2b)$ is the stationary measure of the O-U process. Now the solution: from Example 8.10 we know that if $\mathbb{E}X_0 = 0$ then also $\mathbb{E}(X_t) = 0$ for all $t > 0$. Therefore $Cov(X_t, X_s) = \mathbb{E}(X_t X_s)$. To calculate $\mathbb{E}(X_t X_s)$ we use Theorem 8.4:

$$\mathbb{E}(X_t X_s) = e^{-b(t+s)}\frac{\sigma^2}{2b} + \sigma^2 e^{-b(t+s)}\int_0^s e^{2bu}du = \frac{\sigma^2}{2b}e^{-b(t-s)}\,.$$

Using the Itô formula, we have

$$\begin{aligned} d(|X_t|^2) &= 2X_t dX_t + \sigma^2 dt \\ &= -2bX_t^2 dt + 2\sigma X_t dW_t \sigma^2 dt, \end{aligned}$$

so that

$$X_t^2 = X_0^2 - \int_0^t 2bX_s^2 ds + 2\int_0^t \sigma X_s dW_s + \sigma^2 t.$$

Taking expectation on both sides we get

$$\mathbb{E}X_t^2 = \mathbb{E}X_0^2 - \int_0^t 2b(\mathbb{E}X_s^2)ds + \sigma^2 t.$$

You could solve the above equation, which in differential form is

$$\dot{u} = -2bu + \sigma^2, \text{ having set } u(t) = \mathbb{E}X_t^2,$$

so that $u(t) = e^{-2bt}u(0) + \int_0^t e^{-2b(t-s)}\sigma^2 ds$. But if you only need an estimate you can apply Gronwall's Lemma and get

$$\mathbb{E}X_t^2 \leq (\mathbb{E}X_0^2 + \sigma^2\, t)e^{-2bt}\,.$$

29. Clearly, $\mathbb{E}X_t = \mathbb{E}(X_0)\exp^{(r-\sigma^2/2)t}\mathbb{E}(e^{\sigma W_t})$. So we need to calculate $\mathbb{E}(e^{\sigma W_t})$. Set $Y_t = e^{\sigma W_t}$. Using ***Itô formula,

$$dY_t = \sigma Y_t\, dW_t + \frac{1}{2}\sigma^2 Y_t\, dt$$

so that

$$Y_t = Y_0 + \int_0^t \sigma Y_s\, dW_s + \frac{1}{2}\int_0^t \sigma^2 Y_s\, ds\,.$$

Taking expectation,

$$\mathbb{E}Y_t = \mathbb{E}Y_0 + \mathbb{E}\left(\int_0^t \sigma Y_s\, dW_s\right) + \frac{1}{2}\int_0^t \sigma^2 \mathbb{E}Y_s\, ds\,.$$

You can check that Theorem 8.4 can be applied to the second addend on the RHS of the above equation, so that $\mathbb{E}\left(\int_0^t \sigma Y_s\, dW_s\right) = 0$. This means that $\mathbb{E}Y_t$ satisfies the ODE

$$\dot{u}(t) = \frac{1}{2}\sigma^2 u(t), \quad u(0) = 1,$$

hence $\mathbb{E}Y_t = \exp(\sigma^2 t/2)$. To conclude, $\mathbb{E}X_t = \mathbb{E}(X_0)e^{rt}$.

30. Just a straightforward application of the Itô chain rule for two-dimensional SDEs, with $g : \mathbb{R}_+ \times \mathbb{R}^2 \to \mathbb{R}^2$.

31. Step by step.

- Using the conversion formula, (137) is equivalent to the Itô equation

$$d\tilde{X}_t = \left[d(t) + \frac{1}{2}f^2(t)\right]\tilde{X}_t dt + f(t)X_t dW_t.$$

The solution to such an equation can be found by using Example 8.12 and it is

$$\tilde{X}_t = X_0 e^{\int_0^t d(s)ds + \int_0^t f(s)dW_s}$$

(which is not the same as (67)).

- (138) is a simple ODE, for each fixed ω, so

$$X^k(t) = X_0 e^{\int_0^t d(s)ds + \int_0^t f(s)\xi^k(s)}\,.$$

- ξ_s^k is mean zero and Gaussian and $f(s)$ is deterministic, so $f(s)\xi_s^k$ is Gaussian and I_t^k is too; therefore I_t^k can only converge to a Gaussian process. Assuming we can exchange limit and integral, because I_t^k is mean zero, also the limiting process has mean zero. As for the covariance function:

$$\begin{aligned}\lim_{k\to\infty} \mathbb{E}(I_t^k I_s^k) &= \lim_{k\to\infty} \mathbb{E}\int_0^t\int_0^s f(u)f(v)\xi_u^k\xi_v^k \\ &= \lim_{k\to\infty}\int_0^t\int_0^s f(u)f(v)\mathbb{E}\xi_u^k\xi_v^k \\ &= \int_0^t\int_0^s f(u)f(v)\lim_{k\to\infty}\mathfrak{d}^k(u-v) = \int_0^{t\wedge s} du f^2(u)du\,.\end{aligned}$$

  Therefore I_t^k converges to a Gaussian process which has the same mean and covariance as I_t. This means that X_t^k converges to a process which has the same distributions as $\tilde{X}_t$.

32. $M \geq 1$ because $\|P_0\|_{B(\mathfrak{B})} = 1$. Now any $t \in \mathbb{R}_+$ can be written as $t = nt_0 + r$ where $0 \leq r \leq t_0$. Therefore, using the semigroup property, we have

$$\|P_t\|_{B(\mathfrak{B})} = \|P_{nt_0}P_r\|_{B(\mathfrak{B})} = \|P_{t_0}^n P_r\|_{B(\mathfrak{B})} \leq M^n M \leq MM^{t/t_0} = Me^{\omega t}\,.$$

33. By the definition of generator,

$$\begin{aligned}\lim_{t\to 0^+}\frac{\mathcal{P}_t f(x) - f(x)}{t} &= \lim_{t\to 0^+}\frac{e^{-mt}-1}{t}f(x) + \frac{1-e^{-mt}}{t}\int f\,d\mu \\ &= -mf(x) + m\int f\,d\mu\,.\end{aligned}$$

34. Using the definition of generator we find $\mathcal{L}f = f'(x)$.

35. Simply, if $f \in \mathcal{B}_m$ the

$$\sup_x\left|\int f(y)p_t(x,dy)\right| \leq \|f\|_\infty \sup_x \int p_t(x,dy) = \|f\|_\infty.$$

36. Notice that the covariance function is a symmetric function, i.e. $C(t,s) = C(s,t)$. Therefore

$$\begin{aligned}\mathbb{E}\left|\frac{1}{T}\int_0^T X(s)ds - \mu\right|^2 &= \frac{1}{T^2}\mathbb{E}\left|\int_0^T X_s ds\right|^2 + \mu^2 - 2\frac{1}{T}\int_0^T \mu^2 ds \\ &= \frac{1}{T^2}\mathbb{E}\left(\int_0^T X_s ds \int_0^T X_t dt\right) - \mu^2 \\ &= \frac{1}{T^2}\int_0^T dt \int_0^T ds C(t,s) \\ &\overset{\text{symmetry}}{=} \frac{2}{T^2}\int_0^T dt \int_0^t ds C(t,s) \leq \frac{const}{T} \longrightarrow 0 \quad \text{as } T \to \infty.\end{aligned}$$

37. The first becomes $X_t = X_0 + \int_0^t (X_s^2 - \frac{1}{2}\sin X_s \cos X_s)ds + \int_0^t \cos X_s dW_s$. The second remains unaltered as the diffusion coefficient does not depend on x. The Itô form of the last one is $dX_t = -\frac{7}{2}X_t dt + 3X_t dW_t$.

38. Let X_u, $u \geq t - \delta$ be the solution of (93). Using Itô formula we have

$$\begin{aligned}df(X_u) = &\frac{\partial f}{\partial x}(X_u)b(u,X_u)du + \frac{\partial f}{\partial x}(X_u)\sigma(u,X_u)dW_u \\ &+ \frac{\partial^2 f}{\partial x^2}(X_u)\sigma^2(u,X_u)du\,.\end{aligned}$$

Integrating both sides and taking expectation:

$$\begin{aligned}\frac{1}{\delta}\left[\mathbb{E}(f(X_t)) - f(x)\right] = &\frac{1}{\delta}\mathbb{E}\int_{t-\delta}^t \frac{\partial f}{\partial x}(X_u)b(u,X_u)du \\ &+ \frac{1}{2\delta}\mathbb{E}\int_{t-\delta}^t \frac{\partial^2 f}{\partial x^2}(X_u)\sigma^2(u,X_u)du.\end{aligned}$$

Now just let $\delta \to 0$ and use the hint.

39. If

$$\int_{\mathbb{R}} f_t^2 d\mu \leq e^{-2\alpha t}\int_{\mathbb{R}} f^2\, d\mu.$$

then the function $t \to e^{2\alpha t}\int_{\mathbb{R}} f_t^2 d\mu$ is decreasing. This means that

$$\frac{d}{dt}\left[e^{2\alpha t}\int_{\mathbb{R}} f_t^2 d\mu\right] \leq 0.$$

The above inequality, calculated in $t = 0$, gives the spectral gap inequality.

40. After you write the generator of the process X_t^v,

$$\mathcal{L}_v = -\sum_{i=1}^{d} \left(v^i(x) + \frac{\partial V}{\partial x^i} \right) \frac{\partial}{\partial x^i} + \sum_{i=1}^{d} \partial_{x^i}^2 \,,$$

all the calculations are completely analogous to all you have seen in one dimension.

41. We use Exercise 31. Assuming (139), if $t, h \geq 0$ then by the semigroup property

$$\|\mathcal{P}_{t+h}u - \mathcal{P}_u\| \leq \|\mathcal{P}_t\| \|\mathcal{P}_h u - u\| \leq Ce^{\alpha t} \|\mathcal{P}_h u - u\| \overset{\text{by assumption}}{\longrightarrow} 0,$$

as $h \to 0$. Now do the same thing for $\|\mathcal{P}_{t-h} - \mathcal{P}_t\|$, $t \geq h \geq 0$. The converse implication is obvious.

42. We will use Hille-Yosida Theorem.

    (a) If iA is self-adjoint then ($\mathcal{D}(A) = \mathcal{D}(A^*)$ and) $(iA)^* = iA$ which implies $-iA^* = iA \Rightarrow A = -A^*$. The reverse implication is analogous.

    (b) If A is the generator of a group of unitary operators then in particular it is densely defined and closed so, using the previous step, showing that A is skew symmetric will do. To this end, for all $x \in \mathcal{D}(A)$,

$$-Ax = \lim_{t \downarrow 0} \frac{U(-t)x - x}{t} = \lim_{t \downarrow 0} \frac{U^*x - x}{t} = A^*x \,.$$

43. $\mathcal{L}_H$ is antisymmetric in L^2_ρ, so

$$\langle h, \mathcal{L}_H h \rangle_\rho = \langle h, \mathcal{L}_H \rangle_\rho = -\langle h, \mathcal{L}_H \rangle_\rho$$

hence $\langle h, \mathcal{L}_H h \rangle_\rho = 0$. Using this fact

$$\frac{d}{dt} \|e^{t\mathcal{L}_H}\|^2 = 2\langle h_t, \mathcal{L}_H h_t \rangle_\rho = 0.$$

44. The potential is $V(x) = x^3/3$, the generator is

$$\mathcal{L} = -x^2 \partial_x + \frac{\sigma^2}{2} \partial_x^2 \,.$$

The density of the invariant measure is $\rho(x) = e^{-\frac{2x^3}{3\sigma^2}} / \mathcal{Z}$, where $\mathcal{Z}$ is a normalization constant. To show the symmetry, see Example 10.17 at the beginning of Section 10.3.

45. Using the product rule in (a) and the multidimensional Itô formula in (b) and (c), we have

    (a) $d(X_tY_t) = X_t dY_t + Y_t dX - t + dX_t dY_t$. However $dX_t dY_t = 0$ in this case so

    $$d(X_tY_t) = rX_tY_t dt - Y_t X_t dt + 2X_t dB_t + \sigma Y_t dW_t\,.$$

    (b) Using $dX_t dY_t = 0$, $dX_t dX_t = \sigma^2 dt$ and $dY_t dY_t = 4dt$, we have

    $$\begin{aligned} dg(X_t, Y_t) = &\frac{\partial g}{\partial x}(X_t, Y_t)dX_t + \frac{\partial g}{\partial y}(X_t, Y_t)dY_t \\ &+ \frac{\sigma^2}{2}\frac{\partial^2 g}{\partial x^2}(X_t, Y_t)dt + 2\frac{\partial^2 g}{\partial y^2}(X_t, Y_t)dt\,. \end{aligned}$$

    (c) Use the result of (b) and apply them to the function $g(x, y) = x^2 y$ to obtain

    $$d(X_t)2X_tY_t dX_t + X_t^2 dY_t + \sigma^2 Y_t dt\,.$$

46. All the inequalities I will write are assumed to hold for $f \in L^2_\nu \cap \mathcal{D}(\mathcal{L})$ (even though for (a) one can just consider functions in L^2_ν.)

    (a) Just consider the function

    $$H(a) = \int_{\mathbb{R}} (f - a)^2 \, d\nu.$$

    Calculate the derivative of $H(a)$ with respect to a:

    $$\frac{d}{da}H(a) = -2\int_{\mathbb{R}} (f - a)\, d\nu = 0 \iff a = \int f d\nu.$$

    Moreover $\frac{d^2}{da^2}H(a) = 1$ so this is a minimum.

    (b) Using the hint and the result of point (a), one gets

    $$\begin{aligned} \int_{\mathbb{R}} \left(f - \int_{\mathbb{R}} f d\nu\right)^2 d\nu &\leq \int_{\mathbb{R}} \left(f - \int_{\mathbb{R}} f d\mu\right)^2 d\nu \\ &\leq e^{Osc(U)} \int_{\mathbb{R}} \left(f - \int_{\mathbb{R}} f d\mu\right)^2 d\mu \\ &\leq \frac{e^{Osc(U)}}{\alpha}\langle -\mathcal{L}f, f\rangle_\mu \\ &\leq \frac{e^{2Osc(U)}}{\alpha}\langle -\mathcal{L}f, f\rangle_\nu, \end{aligned}$$

    which concludes the proof.

47. Results:

- $Y_t = Y_0 \exp(\alpha B_t - \frac{1}{2}\alpha^2 t) + r \int_0^t \exp[\alpha(B_t - B_s) - \frac{1}{2}\alpha^2(t-s)]ds$.
- $X_t = e^{-t}X_0 + e^{-t}B_t$ assuming $B_0 = 0$.
- $X_t = 3\exp[-6t + 4B_t]$ and $\mathbb{E}(X_t) = 3e^{2t}$.