7 Random Systems

So far we have been studying deterministic systems. But the world around us is not very deterministic; there are fluctuations in everything from a cup of coffee to the global economy. In principle, these could be described from first principles, accounting for the actions of every molecule in the cup or every person on the planet. This is hopeless of course, for even if we could determine the initial condition of the whole system, any tiny errors we made would rapidly grow so that our model would no longer be exact. Even if that wasn't a problem, such an enormous model would offer little insight beyond observing the original system.

Fortunately, random (or *stochastic*) systems can be as simple to understand as deterministic systems, if we're careful to ask the right questions. It's a mistake to try to guess the value of the toss of a fair coin, because it is completely uncertain. But we can quite precisely answer questions about the likelihood of events such as seeing a particular string of heads and tails, and easily model a typical sequence of tosses. In fact, simple stochastic systems are as straightforward as simple deterministic ones. Just as deterministic systems get harder to understand as they become more complex, stochastic systems become harder to understand as they become less ideally random, and approximation methods are needed to describe this deviation.

Different problems lie at different ends of this continuum. For a bridge builder, the bridge had better remain standing with near-certain probability, so the uncertainty is $\approx 1 - \epsilon$, where ϵ is very small. For a financial trader, markets are nearly random, and so any (legal) trading strategy must start with the assuption of near-random uncertainty. There, the probabilities are $\approx 0.5 + \epsilon$. The representations appropriate to $1 - \epsilon$ are different from the ones aimed at $0.5 + \epsilon$, but they do share the central role of variables that are random rather than deterministic.

Our first step will be to look at how to describe the properties of a random variable, and the inter-relationships among a collection of them. Then we will turn to stochastic processes, which are random systems that evolve in time. The chapter closes by looking at algorithms for generating random numbers.

7.1 RANDOM VARIABLES

The most important concept in all of stochastic systems is the idea of a *random variable*. This is a fluctuating quantity, such as the hiss from an amplifier, that is described by governing equations in terms of probability distributions. The crucial distinction between

random variables and ordinary deterministic variables is that it is not possible to predict the value of a random variable, but it is possible to predict the probability for seeing a given event in terms of the random variable (such as the likelihood of observing a value in a particular range, or the functional form of the power spectrum). An ensemble of identical stochastic processes will generate different values for the random variables, but each member of the ensemble (or *realization*) will obey these same distribution laws. It is important to distinguish between the random variables that appear in distribution functions and the particular values that are obtained in a single realization of a stochastic process by drawing from the distribution.

The simplest random system consists of values x taken from a distribution $p(x)$. For example, in a coin toss x can be heads or tails, and $p(\text{heads}) = p(\text{tails}) = 1/2$. In this case x takes on discrete values; it is also possible for a random variable to come from a continuous distribution. For a continuous variable, $p(x)\,dx$ is the probability to observe a value between x and $x + dx$, and more generally

$$\int_a^b p(x)\,dx \tag{7.1}$$

is the probability to observe x between a and b. For a continuous variable, remember that $p(x)$ must always appear with a dx if you want to make a statement about the likelihood of an observable event – a common mistake is to try to evaluate the likelihood of an event by using the value of the density $p(x)$ alone without integrating it over an interval. That is meaningless; among many problems it can easily lead to the impossible result of probabilities greater than 1.

Now consider a string $(x_1, x_2, \ldots, x_N)$ of N x's drawn from a distribution $p(x)$, such as a series of coin tosses. The average value (or *expectation value*) of a function $f(x)$ is defined by

$$\begin{aligned} \langle f(x)\rangle &= \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} f(x_i) \\ &= \int_{-\infty}^{\infty} f(x)\,p(x)\,dx \quad . \end{aligned} \tag{7.2}$$

(in the statistical literature, the expectation is usually written $E[f(x)]$ instead of $\langle f(x)\rangle$). The equivalence of these two equations can be taken as an empirical definition of the probability distribution $p(x)$. This is for a continuous variable; here and throughout this chapter, for a discrete variable the integral is replaced by a sum over the allowed values. If $p(x)$ is not defined from $-\infty$ to ∞ then the integral extends over the values for which it is defined (this is called its *support*). A trivial example of an expectation value is

$$\langle 1\rangle = \int_{-\infty}^{\infty} 1\;p(x)\,dx = 1 \tag{7.3}$$

(since probability distributions must be normalized). An important expectation is the *mean value*

$$\bar{x} \equiv \langle x\rangle = \int_{-\infty}^{\infty} x\;p(x)\,dx \quad . \tag{7.4}$$

The mean value might never actually occur (for example $p(x)$ might be a *bimodal* distribution with two peaks that vanish at $p(\bar{x})$; the average number of legs of dogs and their

owners is three). Therefore, another useful quantity is the value of x at which $p(x)$ is a maximum, the *maximum likelihood* value.

The mean value tells us nothing about how big the fluctuations in x are around it. A convenient way to measure this is by the *variance* σ_x^2, defined to be the average value of the square of the deviation around the mean

$$\begin{aligned}\sigma_x^2 &\equiv \langle (x-\bar{x})^2 \rangle \\ &= \langle x^2 - 2x\bar{x} + \bar{x}^2 \rangle \\ &= \langle x^2 \rangle - 2\langle x \rangle \bar{x} + \bar{x}^2 \\ &= \langle x^2 \rangle - 2\bar{x}^2 + \bar{x}^2 \\ &= \langle x^2 \rangle - \langle x \rangle^2 \end{aligned} \tag{7.5}$$

(remember that $\bar{x} = \langle x \rangle$ is just a constant). The square root of the variance is called the *standard deviation* σ_x. To calculate the variance, we need to know the mean value, and the *second moment*

$$\langle x^2 \rangle = \int_{-\infty}^{\infty} x^2 \, p(x) \, dx \quad . \tag{7.6}$$

It is similarly possible to define the higher-order moments $\langle x^n \rangle$.

7.1.1 Joint Distributions

Now let's consider two random variables x and y, such as the result from throwing a pair of dice, that are specified by a joint density $p(x, y)$. The expected value of a function that depends on both x and y is

$$\langle f(x,y) \rangle = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y) \, p(x,y) \, dx \, dy \quad . \tag{7.7}$$

$p(x, y)$ must be normalized, so that

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(x,y) \, dx \, dy = 1 \quad . \tag{7.8}$$

It must also be normalized with respect to each variable, so that

$$p(x) = \int_{-\infty}^{\infty} p(x,y) \, dy \tag{7.9}$$

and

$$p(y) = \int_{-\infty}^{\infty} p(x,y) \, dx \quad . \tag{7.10}$$

Integrating a variable out of a joint distribution is called *marginalizing* over the variable.

For joint random variables a very important quantity is $p(x|y)$ ("the probability of x given y"). This is the probability of seeing a particular value of x if we already know the value of y, and is defined by *Bayes' rule*

$$p(x|y) = \frac{p(x,y)}{p(y)} \quad , \tag{7.11}$$

which takes the joint probability and divides out from it the known scalar probability. This is easily extended to combinations of more variables,

$$\begin{aligned} p(x,y,z) &= p(x|y,z)\ p(y,z) \\ &= p(x|y,z)\ p(y|z)\ p(z) \\ &= p(x,y|z)\ \ p(z) \quad , \end{aligned} \tag{7.12}$$

manipulations that will recur in later chapters.

If x and y are *independent*, $p(x,y) = p(x)p(y)$. The probability of seeing two independent events is the product of the probabilities of seeing each alone. For independent variables, the conditional distribution will then depend on just one variable, $p(x|y) = p(x)$. This provides a convenient way to remember the form of Bayes' rule, because for independent variables $p(x|y) = p(x)p(y)/p(y) = p(x)$. For *uncorrelated* variables, $\langle xy \rangle = \langle x \rangle \langle y \rangle$. Independent variables are always uncorrelated, but the converse need not be true (although it often is).

Bayes' rule has an almost mystical importance to the (frequently vocal) community of *Bayesians*, who object to any distribution that is not conditional. Equation (7.2) represents a *frequentist* viewpoint, in which probabilites are defined by observed fractions of events. If you are told that the probability of flipping a coin and seeing a head is 50%, you can perform a series of trials to check that. But, if you're told that the chance of rain on a particular day is 50%, you can't check that in the same way. The day will happen only once, so it doesn't make sense to discuss an ensemble of that day. What's really being quoted is the probability of the belief that it will rain, given evidence supporting that belief, $p(\text{belief}|\text{evidence})$. That can be written via Bayes' rule as three terms:

$$\begin{aligned} p(\text{belief}|\text{evidence}) &= \frac{p(\text{belief}, \text{evidence})}{p(\text{evidence})} \\ &= \frac{p(\text{evidence}|\text{belief})\ p(\text{belief})}{p(\text{evidence})} \\ &= \frac{p(\text{evidence}|\text{belief})\ p(\text{belief})}{\int_{\text{beliefs}} p(\text{evidence}|\text{belief})\ p(\text{belief})} \\ \textit{posterior} &= \frac{\textit{likelihood} \times \textit{prior}}{\textit{evidence}} \quad . \end{aligned} \tag{7.13}$$

$p(\text{evidence}|\text{belief})$ is called the *likelihood*, and is a predictive model for the evidence given a belief. $p(\text{belief})$ is called a *prior*, and no well-equipped Bayesian would be caught without one. While it might sound more like religion than science to have a probability for a belief without evidence, it turns out that scientific inference is full of these whether or not they're written down explicitly. $p(\text{evidence})$ is called the *evidence* term; in some cases there's an independent basis for obtaining it, otherwise it must be found by integrating the numerator over all possible beliefs in the model, which is computationally intractable for a non-trivial model but can be approximated by statistical sampling [Gelman *et al.*, 2013]. And $p(\text{belief}|\text{evidence})$ is the *posterior*; it's how a belief is updated following observations about it. We'll have much more to say about all of this in Chapter 13.

To be able to work with random variables we must understand how their distributions change when they are transformed in mathematical expressions. First, if we know a distribution $p(x)$, and are given a change of coordinates $y(x)$, then what is the distribution

$p(y)$? This is easily found by remembering that probabilities are defined in terms of the value of the distribution at a point times a differential element, and locally equating these:

$$\begin{aligned} p(y)\ |dy| &= p(x)\ |dx| \\ \Rightarrow p(y) &= p(x) \left| \frac{dy}{dx} \right|^{-1} \quad . \end{aligned} \tag{7.14}$$

In higher dimensions, the transformation of a distribution is done by multiplying it by the *Jacobian*, the absolute value of the determinant of the matrix of partial derivatives of the change of coordinates (Problem 5.2).

If we take $p(y) = 1$ in the unit interval and integrate both sides of equation (7.14), this relationship has a useful implication

$$\int_{-\infty}^{x} p(x')\ dx' = \int_{-\infty}^{x} 1\ \frac{dy(x')}{dx}\ dx'$$

$$P(x) = y(x) \quad . \tag{7.15}$$

This means that if we choose the transformation $y(x)$ to be the integral $P(x)$ of a given distribution $p(x)$ (P is called the *cumulative distribution*), and pick random values of y from a uniform distribution, then the corresponding values of x will be distributed according to p.

The next question is how to combine random variables. The simplest operation is adding two independent random variables, x_1 drawn from p_1, and x_2 drawn from p_2, to create a new random variable $y = x_1 + x_2$. To find the probability distribution $p(y)$ for y, we must add up the probabilities for each of the different ways that x_1 and x_2 can add up to that value of y. The probability of seeing a particular pair of x_1 and x_2 is given by the product of their probabilities, and so integrating we see that

$$\begin{aligned} p(y) &= \int_{-\infty}^{\infty} p_1(x) p_2(y - x)\ dx \\ &= p_1(x) * p_2(x) \quad . \end{aligned} \tag{7.16}$$

The probability distribution for the sum of two random variables is the convolution of the individual distributions. Now consider the average of N variables

$$y_N = \frac{x_1 + x_2 + \cdots + x_N}{N} \tag{7.17}$$

that are independent and identically distributed (often abbreviated as *iid*), and let's look at what happens as $N \to \infty$. The distribution of y is equal to the distribution of x convolved with itself N times, and since taking a Fourier transform turns convolution into multiplication, the Fourier transform of the distribution of y is equal to the product of the N transforms of the distribution of x. This suggests an important role for Fourier transforms in studying random processes.

7.1.2 Characteristic Functions

The Fourier transform of a probability distribution is called the *characteristic function*, and is equal to the expectation value of the complex exponential

$$\langle e^{ikx} \rangle = \int_{-\infty}^{\infty} e^{ikx} p(x)\, dx \quad . \tag{7.18}$$

For a multivariate distribution, the characteristic function is

$$\langle e^{i\vec{k}\cdot\vec{x}} \rangle = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{i\vec{k}\cdot\vec{x}} p(\vec{x})\, d\vec{x} \quad . \tag{7.19}$$

Now let's look at the characteristic function for the deviation of y_N, the average of N iid random numbers $x_1, \ldots, x_N$, from the average value $\bar{x}$:

$$\begin{aligned}
\langle e^{ik(y_N - \bar{x})} \rangle &= \langle e^{ik(x_1 + x_2 + \cdots + x_N - N\bar{x})/N} \rangle \\
&= \langle e^{ik[(x_1 - \bar{x}) + \cdots + (x_N - \bar{x})]/N} \rangle \\
&= \langle e^{ik(x - \bar{x})/N} \rangle^N \\
&= \left\langle 1 + \frac{ik}{N}(x - \bar{x}) - \frac{k^2}{2N^2}(x - \bar{x})^2 + \mathcal{O}\left(\frac{k^3}{N^3}\right) \right\rangle^N \\
&= \left[1 + 0 - \frac{k^2 \sigma_x^2}{2N^2} + \mathcal{O}\left(\frac{k^3}{N^3}\right) \right]^N \\
&\approx e^{-k^2 \sigma_x^2 / 2N} \quad .
\end{aligned} \tag{7.20}$$

In deriving this we have used the fact that the x_i are independent and identically distributed, and done a Taylor expansion around the average value. The last line follows because

$$\lim_{N \to \infty} \left[1 + \frac{x}{N} \right]^N = e^x \tag{7.21}$$

(which can be verified by comparing the Taylor series of both sides), and we can neglect higher-order terms in the limit $N \to \infty$. To find the probability distribution for y we now take the inverse transform

$$\begin{aligned}
p(y_N - \bar{x}) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-k^2 \sigma_x^2 / 2N} e^{-ik(y - \bar{x})}\, dk \\
&= \sqrt{\frac{N}{2\pi\sigma_x^2}} e^{-N(y_N - \bar{x})^2 / 2\sigma_x^2} \quad .
\end{aligned} \tag{7.22}$$

Something remarkable has happened: in the limit $N \to \infty$, the sum of N variables approaches a *Gaussian* distribution, independent of the distribution of the variables! This is called the *Central Limit Theorem*, and explains why Gaussians are so important in studying random processes. The Gaussian distribution is also called the *normal* distribution, because it is so, well, normal. Since the standard deviation is $\sigma_x / \sqrt{N}$, which vanishes as $N \to \infty$, equation (7.22) also contains the *Law of Large Numbers*: the average of a large number of random variables approaches the mean, independent of the distribution.

The characteristic function takes on an interesting form if the complex exponential is expanded in a power series:

$$\langle \mathrm{e}^{i\vec{k}\cdot\vec{x}} \rangle = \int \mathrm{e}^{i\vec{k}\cdot\vec{x}}\, p(\vec{x})\, d\vec{x}$$

$$= \sum_{n_1=0}^{\infty} \sum_{n_2=0}^{\infty} \cdots \sum_{n_d=0}^{\infty} \frac{(ik_1)^{n_1}}{n_1!} \frac{(ik_2)^{n_2}}{n_2!} \cdots \frac{(ik_N)^{n_N}}{n_N!} \langle x_1^{n_1} x_2^{n_2} \cdots x_N^{n_N} \rangle \quad . \tag{7.23}$$

This provides a relationship between the moments of a distribution and the distribution itself (via its Fourier transform). For this reason, the characteristic function is also called the *moment generating function*. If the characteristic function is known, the moments can be found by taking derivatives of the appropriate order and evaluating at $\vec{k} = 0$ (only one term will be nonzero).

Another important object is the logarithm of the characteristic function. If we choose to write this as a power series in k of the form (for the 1D case)

$$\log\langle e^{ikx} \rangle = \sum_{n=1}^{\infty} \frac{(ik)^n}{n!} C_n \quad , \tag{7.24}$$

this defines the *cumulants* C_n (note that the sum starts at $n = 1$ because $\log 1 = 0$ and so there is no constant term). The cumulants can be found by comparing this to the power series expansion of the characteristic function,

$$\exp\left(\sum_{n=1}^{\infty} \frac{(ik)^n}{n!} C_n \right) = \sum_{n=0}^{\infty} \frac{(ik)^n}{n!} \langle x^n \rangle \quad , \tag{7.25}$$

expanding the exponential as

$$e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \cdots \quad , \tag{7.26}$$

and grouping terms by order of k. The cumulants have an interesting connection to Gaussianity (Problem 5.1).

7.1.3 Entropy

In the next Section we'll relate the Fourier transform of a random process to its correlation function, but in Chapter 21 we'll see that signals from even simple nonlinear systems can have broadband power spectra and hence featureless correlation structure. Information-theoretic quantities provide an elegant alternative that captures the essential features of a correlation function, and more. Let's start by assuming that we have an observable that can take on one of M values

$$x \in \{1, \ldots, M\} \quad . \tag{7.27}$$

From a series of samples of x we can estimate the probability distribution; naively this can be done by binning, taking the ratio of the number of times a value x was seen N_x to the total number of points N

$$p(x) \approx N_x / N \quad . \tag{7.28}$$

Chapter 17 discusses the limitations of, and alternatives to, this simple estimator.

The *entropy* of this distribution is given by

$$H(x) = -\sum_{x=1}^{M} p(x) \log_2 p(x) \quad . \tag{7.29}$$

It is the average number of bits required to describe a sample taken from the distribution, i.e., the expected value of the *information* in a sample. The entropy is a maximum if the distribution is flat (we don't know anything about the next point), and a minimum if the distribution is sharp (we know everything about the next point).

The information is the minimum number of bits a code can use to communicate a value [Shannon & Weaver, 1949], and the entropy is the expected number of bits for a sample drawn from the distribution. If a code designed for a distribution q is used, but the samples are actually drawn from p, then the expected number of bits is

$$\begin{aligned} -\sum_{x=1}^{M} p(x) \log_2 q(x) &= H(p) + \sum_{x=1}^{M} p(x) \log_2 \frac{p(x)}{q(x)} \\ &\equiv H(p) + D(p||q) \quad . \end{aligned} \tag{7.30}$$

$D(p||q)$ is the *Kullback–Leibler distance* or *relative entropy*. It's nonnegative, because $H(p)$ is the most efficient coding possible. It is not a true distance, because it's not symmetrical in p and q and there isn't a triangle inequality relating distances among three points, however it appears widely in areas including inference and statistical physics as a way to measure how similar two distributions are.

Consider now a sample drawn from a continuous distribution. The probability to see a value between x and $x+dx$ is $p(x)dx$, and the information this provides is $\log_2[p(x)dx] = \log_2 p(x) + \log_2 dx$. As $dx \to 0$ this diverges! That's in fact what it should do, because there can be an infinite amount of information in an infinite-precision real number. The *differential entropy* of a continuous distribution is the part that does not diverge:

$$h(x) = -\int p(x) \log_2 p(x)\, dx \quad . \tag{7.31}$$

(where the integral is over the support of p). Because we've ignored the diverging part the value of the differential entropy is not meaningful; it can be be positive or negative. However differences between differential entropies are meaningful, because the diverging parts will cancel.

For a sequence of samples, we can ask for the number of times $N_{\vec{x}}$ a particular set of values $\vec{x} = (x_1, x_2, \ldots)$ was seen in N observations of $\vec{x}$:

$$p(\vec{x}) = N_{\vec{x}}/N \quad , \tag{7.32}$$

and from this measure the *block entropy*

$$H(\vec{x}) = -\sum_{x_1=1}^{M} \sum_{x_2=1}^{M} \cdots\ p(\vec{x}) \log_2 p(\vec{x}) \tag{7.33}$$

which gives the average number of bits needed to describe the sequence.

The *mutual information* is defined to be the difference in the expected information

between two samples taken independently and taken together

$$I(x,y) = H(x) + H(y) - H(x,y) \tag{7.34}$$

$$= -\sum_{x=1}^{M} p(x)\log_2 p(x) - \sum_{y=1}^{M} p(y)\log_2 p(y) + \sum_{x=1}^{M}\sum_{y=1}^{M} p(x,y)\log_2 p(x,y) \quad . \tag{7.35}$$

If the points don't depend on each other then the mutual information is zero:

$$p(x,y) = p(x)p(y) \;\Rightarrow\; I(x,y) = 0 \quad , \tag{7.36}$$

and if they are completely dependent then the mutual information is equal to the bits in one sample:

$$p(x,y) = p(y) \;\Rightarrow\; I(x,y) = H(x) \quad . \tag{7.37}$$

Here then is an alternative to correlation functions, measuring the connection between two variables without assuming any functional form other than what is needed to estimate a probability distribution. Chapter 21 will develop the deep connection between entropy and dynamics.

7.2 STOCHASTIC PROCESSES

It is now time for time to appear in our discussion of random systems. When it does, this becomes the study of *stochastic processes*. We will look at two ways to bring in time: the evolution of probability distributions for variables correlated in time, and stochastic differential equations.

If $x(t)$ is a time-dependent random variable, its Fourier transform

$$X(\nu) = \lim_{T\to\infty} \int_{-T/2}^{T/2} \mathrm{e}^{i2\pi\nu t} x(t)\, dt \tag{7.38}$$

is also a random variable but its *power spectral density* $S(\nu)$ is not:

$$S(\nu) = \langle |X(\nu)|^2 \rangle = \langle X(\nu)X^*(\nu) \rangle \tag{7.39}$$

$$= \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} e^{i2\pi\nu t} x(t)\, dt \int_{-T/2}^{T/2} e^{-i2\pi\nu t'} x(t')\, dt'$$

(where X^* is the complex conjugate of X, replacing i with $-i$). The inverse Fourier transform of the power spectral density has an interesting form,

$$\int_{-\infty}^{\infty} S(\nu) e^{-i2\pi\nu\tau}\, d\nu$$

$$= \int_{-\infty}^{\infty} \langle X(\nu)X^*(\nu) \rangle e^{-i2\pi\nu\tau}\, d\nu$$

$$
\begin{aligned}
&= \lim_{T\to\infty} \frac{1}{T} \int_{-\infty}^{\infty} \int_{-T/2}^{T/2} e^{i2\pi\nu t} x(t)\, dt \int_{-T/2}^{T/2} e^{-i2\pi\nu t'} x(t')\, dt'\; e^{-i2\pi\nu\tau}\, d\nu \\
&= \lim_{T\to\infty} \frac{1}{T} \int_{-\infty}^{\infty} \int_{-T/2}^{T/2} \int_{-T/2}^{T/2} e^{i2\pi\nu(t-t'-\tau)}\, d\nu\; x(t)x(t')\, dt\, dt' \\
&= \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} \int_{-T/2}^{T/2} \delta(t-t'-\tau)x(t)x(t')\, dt\, dt' \\
&= \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} x(t)x(t-\tau)\, dt \\
&= \langle x(t)x(t-\tau)\rangle \quad , \qquad (7.40)
\end{aligned}
$$

found by using the Fourier transform of a *delta function*

$$\int_{-\infty}^{\infty} e^{-i2\pi\nu t}\delta(t)\, dt = 1 \quad \Rightarrow \quad \delta(t) = \int_{-\infty}^{\infty} e^{i2\pi\nu t}\, dt \quad , \qquad (7.41)$$

where the delta function is defined by

$$\int_{-\infty}^{\infty} f(x)\delta(x-x_0)\, dx = f(x_0) \quad . \qquad (7.42)$$

This is the *Wiener–Khinchin* theorem. It relates the spectrum of a random process to its *autocovariance function*, or, if it is normalized by the variance, the *autocorrelation function* (which features prominently in time series analysis, Chapter 21).

7.2.1 Distribution Evolution Equations

A natural way to describe a stochastic process is in terms of the probability to see a sample value x at a time t (written x_t) given a history of earlier values

$$p(x_t|x_{t_1}, x_{t_2}, \ldots) \quad . \qquad (7.43)$$

Given starting values for x this determines the probability distribution of the future values. If the distribution depends only on the time differences and not on the absolute time,

$$p(x_t|x_{t-\tau_1}, x_{t-\tau_2}, \ldots) \quad , \qquad (7.44)$$

then the process is said to be *stationary* (sometimes qualified by calling it *narrow-sense* or *strict* stationarity). A more modest definition asks only that the means and covariances of a process be independent of time, in which case the process is said to be *weak* or *wide-sense* stationary.

If the conditional distribution is limited to a finite history

$$p(x_t|x_{t-\tau_1}, x_{t-\tau_2}, \ldots, x_{t-\tau_N}) \qquad (7.45)$$

this is said to be an Nth-order *Markov process*. If it depends on just the previous value

$$p(x_t|x_{t-\tau}) \qquad (7.46)$$

it is simply called a Markov process, and if x and t are discrete variables

$$p(x_t|x_{t-1}) \qquad (7.47)$$

it becomes a *Markov Chain*. As with ODEs, an Nth-order Markov process for a scalar variable can always be converted to a first-order Markov process in an N-dimensional variable.

Given the conditional distribution for one time step $p(x_t|x_{t-\tau})$ we can find the distribution two time steps ahead $p(x_{t+\tau}|x_{t-\tau})$ by adding up the probabilities for all of the ways to get from $x_{t-\tau}$ to $x_{t+\tau}$ through the intermediate value x_t. The probability for each possible path is the product of the probability to get from $x_{t-\tau}$ to x_t times the probability to go from x_t to $x_{t+\tau}$. For a discrete system this is a sum, and for a continuous system it is the integral

$$p(x_{t+\tau}|x_{t-\tau}) = \int_{-\infty}^{\infty} p(x_{t+\tau}|x_t)\, p(x_t|x_{t-\tau})\, dx_t \quad . \tag{7.48}$$

This is called the *Chapman–Kolmogorov* equation. It can be rewritten by multiplying both sides by $p(x_{t-\tau})$ and then integrating over $x_{t-\tau}$:

$$\begin{aligned} p(x_{t+\tau}|x_{t-\tau})\, p(x_{t-\tau}) &= \int_{-\infty}^{\infty} p(x_{t+\tau}|x_t)\, p(x_t|x_{t-\tau})\, p(x_{t-\tau})\, dx_t \\ p(x_{t+\tau}, x_{t-\tau}) &= \int_{-\infty}^{\infty} p(x_{t+\tau}|x_t)\, p(x_t, x_{t-\tau})\, dx_t \\ p(x_{t+\tau}) &= \int_{-\infty}^{\infty} p(x_{t+\tau}|x_t)\, p(x_t)\, dx_t \\ &= \int_{-\infty}^{\infty} p(x_{t+\tau}, x_t)\, dx_t \quad . \end{aligned} \tag{7.49}$$

For a Markov chain with N states $x = (1, \ldots, N)$ this becomes

$$p(x_{t+1}) = \sum_{x_t=1}^{N} p(x_{t+1}|x_t)\, p(x_t) \quad . \tag{7.50}$$

If we define an N-component vector of the state probabilities $\vec{p}_t = \{p(x_t = 1), \ldots, p(x_t = N)\}$, and a matrix of transition probabilities $\mathbf{P}_{ij} = p(x_{t+1} = i|x_t = j)$, then the update for all of the states can be written as

$$\begin{aligned} \vec{p}_{t+1} &= \mathbf{P} \cdot \vec{p}_t \\ \Rightarrow \vec{p}_{t+n} &= \mathbf{P}^n \cdot \vec{p}_t \quad . \end{aligned} \tag{7.51}$$

The powers of $\mathbf{P}$ hence determine the evolution; in particular, if it's possible to get from every state to every other state then the system is said to be *ergodic* [Reichl, 1984].

It's easy to understand what a Markov model can do. After all, equation (7.51) is a simple linear first-order finite difference equation (Section 4.5). If $\mathbf{P}$ has eigenvectors $\vec{v}_i$ with eigenvalues λ_i, and the starting distribution is written in terms of the eigenvectors as

$$\vec{p}_t = \sum_i \alpha_i \vec{v}_i \quad , \tag{7.52}$$

then

$$\vec{p}_{t+n} = \mathbf{P}^n \cdot \sum_i \alpha_i \vec{v}_i$$

$$= \sum_i \alpha_i \lambda_i^n \vec{v}_i \quad . \tag{7.53}$$

One of the earliest uses of the Chapman–Kolmogorov equation was in the context of *Brownian motion*. When the Scottish botanist Robert Brown in 1827 used a microscope to look at pollen grains suspended in a solution, he saw them move in a path like the one shown in Figure 7.1. While he originally thought this was a sign of life, it was of course due to the fluctuating impacts of the solvent molecules on the pollen grain. This was a significant piece of evidence for the atomic theory of matter, and in 1905 Einstein developed a quantitative description of Brownian motion that could predict the distribution for how far a particle would travel in a given time.

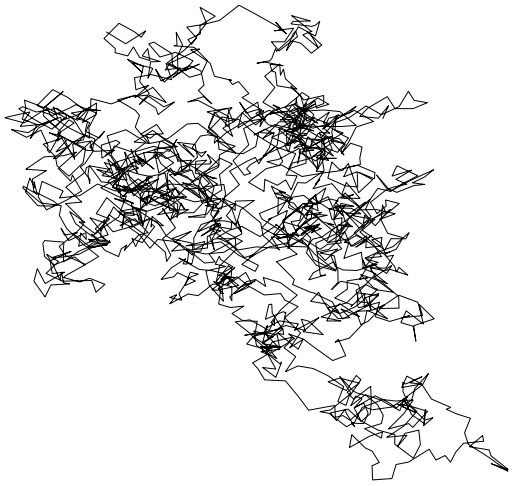

Figure 7.1. 2D Brownian motion.

For 1D Brownian motion (the generalization from 1D to higher dimensions is straightforward), let $p(x,t)\,dx$ be the probability to find the particle between x and $x + dx$ at time t. Because it is a probability, p must be normalized, $\int_{-\infty}^{\infty} p(x,t)\,dx = 1$. Brownian motion arises from the many impacts of the fluid molecules on the particle; let τ be a time that is long compared to these impacts, but short compared to the time for macroscopic motion of the particle. We can then define $p_\tau(x|x')\,dx$, which is the probability for the particle to move from x' to x in a time τ due to the fluctuating impacts on it. This can be used in the Chapman–Kolmogorov equation to write an update equation for p:

$$p(x, t+\tau)\,dx = dx \int_{-\infty}^{\infty} p_\tau(x|x')p(x',t)\,dx' \quad . \tag{7.54}$$

If the fluid is isotropic and homogeneous, then $p_\tau(x|x')$ will depend only on the position difference $x = x' + \delta$, so we can then write $p_\tau(x|x') = p_\tau(\delta)$, where by symmetry $p(\delta) = p(-\delta)$. Then the Chapman-Kolmogorov equation becomes

$$p(x, t+\tau)\,dx = dx \int_{-\infty}^{\infty} p_\tau(\delta)p(x+\delta,t)\,d\delta \quad . \tag{7.55}$$

Since τ has been chosen to be very small compared to the time scale for macroscopic

changes in the particle position, we can expand p in a Taylor series in x and t and keep the lowest-order terms:

$$\left[p(x,t) + \frac{\partial p}{\partial t}\,\tau + \mathcal{O}(\tau^2)\right] dx$$

$$= dx \int_{-\infty}^{\infty} p_\tau(\delta)\left[p(x,t) + \frac{\partial p}{\partial x}\,\delta + \frac{1}{2}\frac{\partial^2 p}{\partial x^2}\,\delta^2 + \cdots\right] d\delta$$

$$= dx\; p(x,t) \underbrace{\int_{-\infty}^{\infty} p_\tau(\delta)\; d\delta}_{1} \;+\; dx\; \frac{\partial p}{\partial x} \underbrace{\int_{-\infty}^{\infty} p_\tau(\delta)\;\delta\; d\delta}_{0} \;+$$

$$dx\; \frac{1}{2}\frac{\partial^2 p}{\partial x^2} \underbrace{\int_{-\infty}^{\infty} p_\tau(\delta)\delta^2\; d\delta}_{\langle\delta^2\rangle} + \mathcal{O}(\delta^3) \quad . \tag{7.56}$$

In the last line, the first integral is 1 because $p_\tau(\delta)$ must be normalized, the second one vanishes because it is the first moment of a symmetrical function $p_\tau(\delta) = p_\tau(-\delta)$, and the last integral is the variance. Cancelling out the $p(x,t)$ on both sides, we're left with

$$\frac{\partial p}{\partial t} = \underbrace{\frac{\langle\delta^2\rangle}{2\tau}}_{\equiv D} \frac{\partial^2 p}{\partial x^2} \quad . \tag{7.57}$$

We've found that the evolution of the probability distribution for Brownian motion is governed by the familiar diffusion equation. In the context of stochastic processes, this is called a *Wiener process*. The diffusion equation is a particular case of the *Fokker–Plank* PDE which governs the evolution of a probability distribution for an underlying Markov process.

We obtained the diffusion equation for a system in which time and space are continuous, but for many problems they are discrete. Consider a 1D random walker that at each time step t_i can hop from a site x_n one unit to the left or right, $x_{n\pm1}$. The change in probability $p(x_n, t_i)$ to find it at a point is then equal to the sum of the probabilities for it to hop into the point, minus the sum of probabilities to hop off the point

$$p(x_n,t_i) - p(x_n,t_{i-1}) = \frac{1}{2}[p(x_{n+1},t_{i-1}) + p(x_{n-1},t_{i-1})] - p(x_n,t_{i-1})$$

$$p(x_n,t_i) = \frac{1}{2}[p(x_{n+1},t_{i-1}) + p(x_{n-1},t_{i-1})] \quad . \tag{7.58}$$

This can be rewritten suggestively by subtracting $p(x_n, t_{i-1})$ from both sides and rearranging:

$$\frac{p(x_n,t_i) - p(x_n,t_{i-1})}{\delta_t} = \underbrace{\frac{\delta_x^2}{2\delta_t}}_{\equiv D}\left[\frac{p(x_{n+1},t_{i-1}) - 2p(x_n,t_{i-1}) + p(x_{n-1},t_{i-1})}{\delta_x^2}\right] \tag{7.59}$$

These are discrete approximations of partial derivatives (Chapter 9). If we take the limit $\delta_t, \delta_x \to 0$ (keeping D finite), we find once again that

$$\frac{\partial p}{\partial t} = D\frac{\partial^2 p}{\partial x^2} \quad . \tag{7.60}$$

More generally, the time derivative of the probability to be at a state x_n is equal to the sum over states x_m of the probability to be at x_m times the rate $W_{x_m \to x_n}$ at which transitions are made from there to x_n, minus the probability to be at x_n times the rate at which transitions are made back to x_m:

$$\frac{\partial p(x_n, t)}{\partial t} = \sum_m W_{x_m \to x_n}\, p(x_m, t) - W_{x_n \to x_m}\, p(x_n, t) \quad . \tag{7.61}$$

This is called the *Master equation*. For a stationary solution the transition rate between two sites is equal in both directions, a condition called *detailed balance*.

7.2.2 Stochastic Differential Equations

An alternative analysis of Brownian motion was first done by Langevin in terms of a stochastic differential equation. A particle moving in a fluid feels a drag force, and as long as the velocity is not too great the force is given by the Stokes drag formula

$$\vec{F} = -6\pi\mu a\vec{v} \quad , \tag{7.62}$$

where μ is the viscosity of the fluid, a is the diameter of the particle, and $\vec{v}$ is the velocity of the particle [Batchelor, 1967]. In addition to this force, we can model Brownian motion by including a fluctuating force η that is due to the molecular impacts on the particle. In terms of these forces, $\vec{F} = m\vec{a}$ for the particle becomes (in 1D):

$$m\frac{d^2x}{dt^2} = -6\pi\mu a\frac{dx}{dt} + \eta \quad . \tag{7.63}$$

This is an example of what is now called a *Langevin equation*. Because η is a random variable, x becomes one, much like the promotion of operator types in a computer program. Therefore we cannot solve for x directly; we must instead use this differential equation to solve for observable quantities that depend on it. To do this, first recognize that

$$\frac{d(x^2)}{dt} = 2x\frac{dx}{dt} \tag{7.64}$$

and

$$\frac{d^2(x^2)}{dt^2} = 2\left(\frac{dx}{dt}\right)^2 + 2x\frac{d^2x}{dt^2} = 2v^2 + 2x\frac{d^2x}{dt^2} \quad . \tag{7.65}$$

Using this, if we multiply both sides of equation (7.63) by x we can rewrite it as

$$\frac{m}{2}\frac{d^2(x^2)}{dt^2} - mv^2 = -3\pi\mu a\frac{d(x^2)}{dt} + \eta x \quad . \tag{7.66}$$

Next, let's take the time expectation value

$$\frac{m}{2}\frac{d^2\langle x^2\rangle}{dt^2} + 3\pi\mu a\frac{d\langle x^2\rangle}{dt} = \underbrace{m\langle v^2\rangle}_{kT} + \underbrace{\langle \eta x\rangle}_{0} \quad . \tag{7.67}$$

In the first term on the right hand side, we've used the fact that the particle is in thermodynamic equilibrium with the fluid to apply the *Equipartition Theorem* [Gershenfeld,

2000], which tells us that

$$\frac{1}{2}m\langle v^2\rangle = \frac{D}{2}kT \quad , \tag{7.68}$$

where D is the dimension (1 in this case), k is Boltzmann's constant 1.38×10^{-23} (J/K), and T is the temperature (in Kelvin). The second term vanishes because the rapidly fluctuating noise term is uncorrelated with the slowly moving particle. Therefore,

$$\frac{d^2\langle x^2\rangle}{dt^2} + \frac{3\pi\mu a}{m}\frac{d\langle x^2\rangle}{dt} = \frac{2kT}{m} \quad . \tag{7.69}$$

This is now an ordinary differential equation for the variance, which can easily be solved to find

$$\langle x^2\rangle = Ae^{-6\pi\mu at/m} + \frac{kT}{3\pi\mu a}\,t \quad . \tag{7.70}$$

The first term is a rapidly decaying exponential transient, leaving us with

$$\langle x^2\rangle = \frac{kT}{3\pi\mu a}\,t \quad . \tag{7.71}$$

This result agrees with the form of Einstein's calculation (Problem 5.4), even though we got here by a very different path. Solving more general stochastic differential equations, and justifying assumptions such as throwing out the $\langle \eta x\rangle$ term, requires extending ordinary calculus to integrals of stochastic functions. This is done by the Ito and the Stratonovich calculus [Gardiner, 2004].

7.3 RANDOM NUMBER GENERATORS

There is a frequent and apparently paradoxical need to use a computer to generate random numbers. In modeling a stochastic system it is necessary to include a source of noise, but computers are (hopefully) not noisy. One solution is to attach a computer peripheral that performs a quantum measurement (which as far as we know can be completely random), or perhaps measures the molecular fluctuations in your coffee cup (which are extremely close to random), but for most people this is not a convenient option. Instead, a more reasonable alternative is to use an algorithm that produces pseudo–random numbers that appear to be more random than can be detected by your application. There is a large difference in what is required to fool a player of a video game and a cryptographic analyst. There is a corresponding broad range of choices for random number generators, based on how sensitive your problem is to the hidden order that must be present in any deterministic algorithm. While these are numerical rather than analytical methods, and so rightfully belong in the next part of this book, they are so closely connected with the rest of this chapter that it is more natural to include them here.

Why discuss random number generators when most every programming language has one built in? There are two reasons: portability and reliability. By explicitly including an algorithm for generating needed random numbers, a program will be sure to give the same answer whatever system it is run on. And built-in generators range from being much too complex for simple tasks to much too simple for complex needs.

7.3.1 Linear Congruential

Linear congruential random number generators, and more sophisticated variants, are the most common technique used for producing random numbers. The simplest example is the map

$$x_{n+1} = 2x_n \pmod 1 \quad . \tag{7.72}$$

Starting with an initial value for x_0 (chosen to be between 0 and 1), this generates a series of new values. The procedure is to multiply the old value by 2, take the part that is left after dividing by 1 (i.e., the fractional part), and use this as the new value. This string of numbers is our first candidate as a random number generator. But how random are the successive values of x? Not very; Figure 7.2 plots x_{n+1} versus x_n for a series of 1000 points. Two things are immediately apparent: the points lie on two lines rather than being uniformly distributed (as they should be for two independent random numbers), and it doesn't look like there are 1000 points in the figure. The first problem can easily be explained by looking back at the definition of the map, which shows that successive pairs of points lie on a line of slope 2, which gets wrapped around because of the mod operator. To understand the second problem, consider x written in a fractional binary expansion (where each digit to the right of the binary point stands for $2^{-1}, 2^{-2}, \ldots$). Each iteration of this map shifts all the digits one place to the left, and throws out the digit that crosses the binary point. This means that it brings up all the bits of the starting position, and finally settles down to a fixed point at $x = 0$ when all the bits are used up.

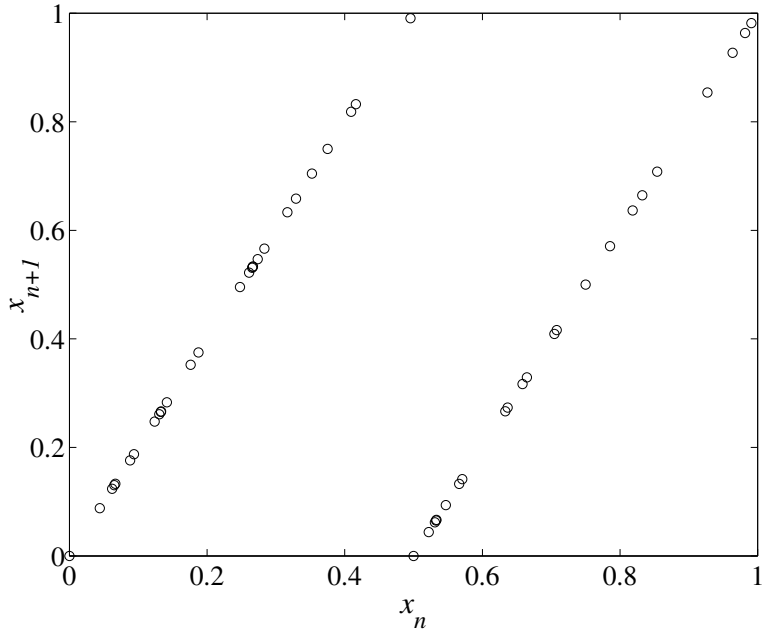

Figure 7.2. 1000 points from the map $x_{n+1} = 2x_n \pmod 1$.

This bad example can be generalized to the class of maps

$$x_{n+1} = ax_n + b \pmod{c} \quad . \tag{7.73}$$

The value of a determines the slope of the lines that the points are on and how many lines there will be ($a = 2$ gave us 2 lines). We want this to be as large as possible, so that the lines fill the space as densely as possible. Then b and c must be chosen relative to a so that the period of the map is as long as possible (it doesn't repeat after a few iterations), there are no fixed points that it can get stuck at, and the digits are otherwise as random as they can be. Choosing optimal values for a, b, and c is a surprisingly subtle problem, but good choices have been worked out as a function of the machine word size used to represent x [Knuth, 1997]. For the common case of a 32-bit integer, with the leading bit used as a sign bit, an optimal choice is

$$x_{n+1} = 8121x_n + 28411 \pmod{134456} \quad . \tag{7.74}$$

Iterating this map produces a string of integers between 1 and 134456 (or a fraction between 0 and 1 if the integer is divided by 134456), shown in Figure 7.3. This now appears to be much more random, and is adequate when there is a simple need for some numbers that "look" random.

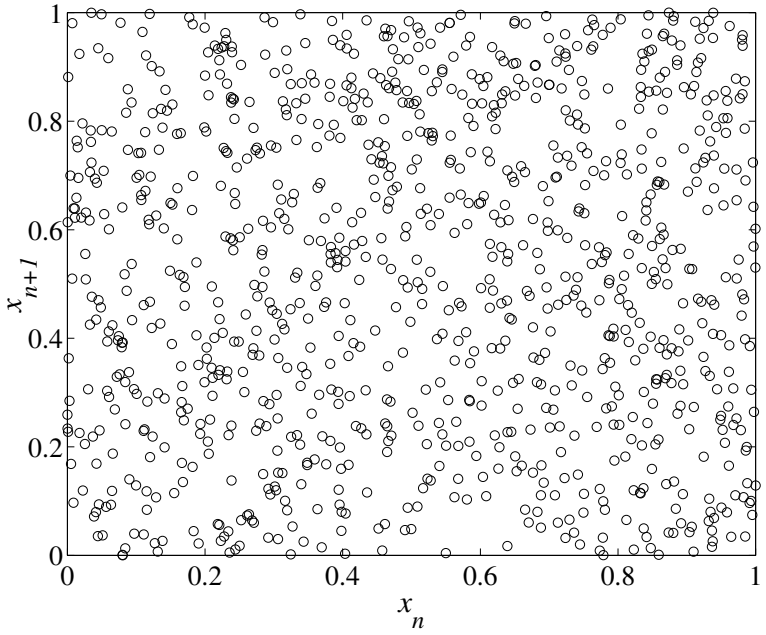

Figure 7.3. 1000 points from the map $x_{n+1} = 8121x_n + 28411 \pmod{134456}$.

This is still not a great generator, because there are only 134456 distinct possible values, and so in a string that long it is possible to detect the predictability. It's also easy to see here why the bits of x are not equally random: if x_n is even then x_{n+1} will be odd, and *vice versa*, so the lowest order bit of x simply oscillates at each step. Not very random. To further improve such a simple linear congruential generator, it is possible

to add degrees of freedom by techniques such as running multiple generators in parallel and using them to shuffle entries in a large array [Press *et al.*, 2007].

7.3.2 Linear Feedback

Linear congruential generators have the problem that all of the bits in each number are usually not equally random; *linear feedback shift registers* (*LFSRs*) provide a powerful alternative that can be used to generate truly pseudo–random bits. A binary linear feedback shift register is specified by a *recursion relation*

$$x_n = \sum_{i=1}^{M} a_i x_{n-i} \pmod 2 \quad . \tag{7.75}$$

This can be viewed as a series of registers through which the bits are shifted, with taps specified by the a_i's that select the values to be added mod 2 (Figure 7.4).

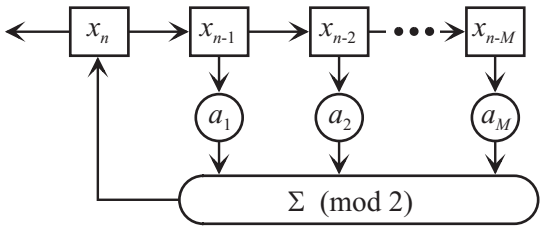

Figure 7.4. A linear feedback shift register.

If the taps are properly chosen, the bits that come out of the end of the shift register are as random as possible. This means that the power spectrum is flat (up to the repeat time, which is $2^M - 1$ for a register with M steps), all possible substrings of bits occur equally often, and so forth. Such a *maximal* LFSR is designed by taking the z-transform of the recursion relation, and finding the taps that make this polynomial have no smaller polynomial factors [Simon *et al.*, 1994]. Table 7.1 gives a (non-unique) choice for maximal taps for a range of register lengths. For example, for order 12 the tap values are 1, 4, 6, 12, so the recursion relation is

$$x_n = x_{n-1} + x_{n-4} + x_{n-6} + x_{n-12} \quad . \tag{7.76}$$

Because the recurrence time is exponential in the length of the register, a surprisingly modest LFSR can have an extremely long period and be hard to distinguish from random (Problem 5.3).

7.3.3 Cryptographic

cryptographic random number requirement, vulnerability [Bellare *et al.*, 1997]
True Random Number Generator (*TRNG*) [Stipčević & Koç, 2014]
thermodynamic: shot noise, avalanche noise
quantum: nuclear decay, photon arrival
theoretically perfect, but vulnerability to interference, bandwidth
construct from *one-way hash function* [Håstad *et al.*, 1999]

Table 7.1. *For an LFSR* $x_n = \sum_{i=1}^{M} a_i x_{n-i} \pmod 2$*, lag* i *values for which* $a_i = 1$ *for the given order* M *(all of the other* a_i*'s are 0).*

M	i	M	i	M	i
2	1, 2	13	1, 3, 4, 13	24	1, 2, 7, 24
3	1, 3	14	1, 6, 10, 14	25	3, 25
4	1, 4	15	1, 15	26	1, 2, 6, 26
5	2, 5	16	1, 3, 12, 16	27	1, 2, 5, 27
6	1, 6	17	3, 17	28	3, 28
7	3, 7	18	7, 18	29	2, 29
8	2, 3, 4, 8	19	1, 2, 5, 19	30	1, 2, 23, 30
9	4, 9	20	3, 20	31	3, 31
10	3, 10	21	2, 21	32	1, 2, 22, 32
11	2, 11	22	1, 22	33	13, 33
12	1, 4, 6, 12	23	5, 23	34	1, 2, 27, 34

SHA256 [Dang, 2013]
computer password file
similar to random
efficient, fixed-length output, infeasible to find input from output, infeasible to find two messages with same output, small input change makes large output change

7.4 RANDOM ALGORITHMS

one of the most significant impacts on computational complexity
[Motwani & Raghavan, 1995]
large difference between typical and worst-case
Las Vegas correct, no time guarantee
Monte-Carlo time guarantee, probaility of error
quicksort random pivot
factoring
min cut

7.5 SELECTED REFERENCES

[Feller, 1968] Feller, William (1968). *An Introduction to Probability Theory and its Applications*. 3rd edn. New York, NY: Wiley.

A classic reference for probability theory.

[Gardiner, 2004] Gardiner, C.W. (2004). *Handbook of Stochastic Methods*. 3rd edn. New York, NY: Springer-Verlag.

This is a beautiful survey of techniques for working with stochastic systems.

[Oksendal, 2010] Oksendal, Bernt K. (2010). *Stochastic Differential Equations: An Introduction with Applications*. 6th edn. New York: Springer-Verlag.

The world of stochastic calculus.

[Cover & Thomas, 2006] Cover, Thomas M., & Thomas, Joy A. (2006). *Elements of Information Theory*. 2nd edn. New York: Wiley-Interscience.

Good information on information.

[Knuth, 1997] Knuth, Donald E. (1997). *Semi-Numerical Algorithms*. 3rd edn. *The Art of Computer Programming*, vol. 2. Reading, MA: Addison-Wesley.

The standard starting point for questions about generating and testing random numbers.

[Press *et al*., 2007] Press, William H., Teukolsky, Saul A., Vetterling, William T., & Flannery, Brian P. (2007). *Numerical Recipes in C: The Art of Scientific Computing*. 3nd edn. Cambridge: Cambridge University Press.

Numerical Recipes has a good collection of practical routines for generating random numbers.

[Hull, 2008] Hull, John. (2008). *Options, Futures and Other Derivatives*. 7th edn. Paramus, NJ: Prentice Hall.

The study of stochasticity is perhaps best-developed where the most money is, in analyzing financial markets.

7.6 PROBLEMS

(5.1) (*a*) Work out the first three cumulants C_1, C_2, and C_3.

(*b*) Evaluate the first three cumulants for a Gaussian distribution

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\bar{x})^2/2\sigma^2} \quad . \tag{7.77}$$

(5.2) (*a*) If $\vec{y}(\vec{x}) = (y_1(x_1, x_2), y_2(x_1, x_2))$ is a coordinate transformation, what is the area of a differential element $dx_1\ dx_2$ after it is mapped into the $\vec{y}$ plane? Recall that the area of a parallelogram is equal to the length of its base times its height.

(*b*) Let

$$y_1 = \sqrt{-2\ln x_1}\ \sin(x_2) \qquad y_2 = \sqrt{-2\ln x_1}\ \cos(x_2) \quad . \tag{7.78}$$

If $p(x_1, x_2)$ is uniform, what is $p(y_1, y_2)$?

(*c*) Write a uniform random number generator, and transform it by equation (7.78). Numerically evaluate the first three cumulants of its output.

(5.3) (*a*) For an order 4 maximal LFSR write down the bit sequence.

(*b*) If an LFSR has a clock rate of 1 GHz, how long must the register be for the time between repeats to be the age of the universe ($\sim 10^{10}$ years)?

(5.4) (*a*) Use a Fourier transform to solve the diffusion equation (7.57) (assume that the initial condition is a normalized delta function at the origin).

(*b*) What is the variance as a function of time?

(*c*) How is the diffusion coefficient for Brownian motion related to the viscosity of a fluid?

(*d*) Write a program (including the random number generator) to plot the position as a function of time of a random walker in 1D that at each time step has an equal probability of making a step of $\pm$ 1. Plot an ensemble of 10 trajectories, each 1000 points long, and overlay error bars of width $3\sigma(t)$ on the plot.

(*e*) What fraction of the trajectories should be contained in the error bars?