

Lecture Notes: Introduction to Stochastic Processes (MATH 5652)

By Kevin D. Stubbs

Last Updated: 2026-10-11

University of Minnesota – Twin Cities

Fall 2026

Outline

Chapter 0: Introduction

§0.1 Why Stochastic Processes?	4
§0.2 Examples	4
0.2.1 Brownian Motion	4
0.2.2 Medical Epidemiology	4
0.2.3 Diffusion Models for Image Generation	5
§0.3 A Review of Probability Theory	5
0.3.1 Axioms of Probability	5
0.3.2 Random Variables	6

Chapter 1: Discrete Time Markov Chains

§1.1 Definitions and Examples	7
§1.2 Multistep Transition Probabilities	10
1.2.1 Predicting the Future	11
1.2.2 General k -step transition formula	13
1.2.3 Chapman–Kolmogorov Equation and Probabilistic Interpretation	14
§1.3 Classification of States	15
1.3.1 The Time of First Return, Recurrence, and Transience	17
1.3.2 Expected Number of Returns	20
1.3.3 Closed and Irreducible Sets	21
§1.4 Stationary Distributions	23
1.4.1 Calculating Stationary Distributions in Practice	24
§1.5 Simple Random Walks on $\mathbb{Z}^d$	26
1.5.1 Simple Random Walk on $\mathbb{Z}$	26
1.5.2 Simple Random Walk on $\mathbb{Z}^2$	28
1.5.3 Comparing Dimensions	28
§1.6 Positive and Null Recurrence	29
1.6.1 The Expected Time of Return	29
1.6.2 Stationary Distributions and the Expected Time of Return	30
§1.7 Periodicity and Limiting Behavior	31
1.7.1 Periodicity	31

1.7.2 Limiting Behavior	32
§1.8 Doubly Stochastic Matrices	32
§1.9 Detailed Balance and Reversibility	33
1.9.1 Reversibility	34
§1.10 The Metropolis–Hastings Algorithm	35
1.10.1 The Ising Model: Setup	36
1.10.2 The Ising Model Example	37
§1.11 Exit Distributions and Hitting Probabilities	39
§1.12 Proof of the Law of Total Probability	42

References

Chapter 0: Introduction

§0.1 Why Stochastic Processes?

The world is a complicated place and no matter how hard we try, some things are purely a matter of chance. While probability theory studies randomness in general, stochastic processes try and understand how random evolves with time. Some of the reasons people study stochastic processes include:

1. To understand the physical world,
2. To predict the future given uncertainty,
3. To design fast algorithms for extremely difficult problems.

§0.2 Examples

0.2.1 Brownian Motion

In 1827, Robert Brown, an English botanist, started looking at grains of pollen suspended in water under a microscope. What he found was somewhat surprising, the grains of pollen seemed to randomly jiggle ([click here](#) for a video of pollen grains vibrating). Brown later tried the same experiment with small particles glass and rock dust so ruled out the motion was related to life.

A mathematical understanding of Brown's observations, now called Brownian motion, was eventually provided by Albert Einstein in 1905 and was instrumental in the wide acceptance of atomic and molecular theory. Per Einstein's explanation, Brownian motion happens because the water molecules are moving randomly and hit the pollen grain at random times which causes the vibrations ([click here](#) for an animation of Brownian motion).

0.2.2 Medical Epidemiology

Since the transmission of infectious diseases is inherently uncertain, stochastic processes play an important role in modeling the incidence and spread of disease. By incorporating stochastic processes into disease models, we can account for random fluctuations arising from demographic variability, environmental differences, and discrete transmission events. This provides greater insight into how diseases spread and affect different segments of the population. More broadly, stochastic processes gives the right mathematical framework for quantifying uncertainty and characterizing the range of possible outcomes in disease dynamics.

0.2.3 Diffusion Models for Image Generation

In recent years, we’ve seen an explosion in the power of artificial intelligence (AI) in a wide variety of tasks. While generally people use “AI” as a catchall term, the underlying approach depends a lot on what problem you’re looking at.

For image generation, one of the current best approaches are a family of methods known as “diffusion methods”. These methods work by first taking a set of images and slowly by add random noise to the image. After enough steps of adding noise, you then determine then undo the noise you added at each step. Once you train the diffusion model on enough data, you learn two things: (1) a “forward map” which takes images to noise, and (2) a “backwards map” which takes noise to images. You can use the backwards map to generate new images from noise. Both the forwards map and the backwards map are modeled with stochastic processes.

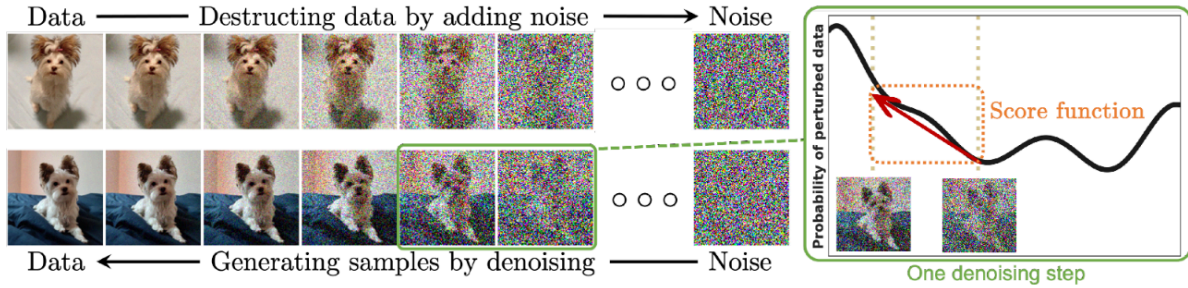


Figure 1: An illustration of how a diffusion model works for image generation. Taken from [1]. In the first row, an image of a dog has noise slowly added to it until it looks completely random. In the second row, denoising is performed to random noise to change the noise into a picture of a different dog.

§0.3 A Review of Probability Theory

0.3.1 Axioms of Probability

We start by recalling the definition of a probability space

Definition 1

A **probability space** is defined by three pieces of information:

1. *Sample space* (denoted Ω) which contains all possible outcomes,
2. *Event space* (denoted $\mathcal{F}$) a set of (reasonable) subsets the sample space, and
3. *Probability function* (denoted $\mathbb{P}$) which assigns each event a probability of occurring.

A probability function is required to satisfy three axioms:

1. If A is an event, then $0 \leq \mathbb{P}(A) \leq 1$.
2. The sum over all outcomes equals 1: $\sum_{s \in S} \mathbb{P}(s) = 1$

3. If A and B are mutually exclusive events, then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.

One of the key concepts in probability theory is *conditional probability*:

Definition 2

Given two events A, B where $\mathbb{P}(B) \neq 0$, the **conditional probability** of A given B is defined:

$$\mathbb{P}(A|B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \quad (1)$$

Theorem 1 (Bayes' Theorem)

Given two events A, B where $\mathbb{P}(B) \neq 0$ we have

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)} \quad (2)$$

Definition 3

We say two events are **independent** if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. Note that this implies $\mathbb{P}(A|B) = \mathbb{P}(A)$ so long as $\mathbb{P}(B) \neq 0$.

Some useful facts which follow from the axioms:

- $\mathbb{P}(A \cap B) \leq \mathbb{P}(A)$ (restriction makes things less likely).
- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$ (mutually exclusive means $\mathbb{P}(A \cap B) = 0$).
- $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$ (the probability A does NOT happen equals 1 minus the probability A does happen).
- If $\mathbb{P}(A) = 1$, then we say the event is certain (or occurs almost surely).
- $\mathbb{P}(\emptyset) = 0$ (the probability nothing happens is zero).

0.3.2 Random Variables

A random variable X is a function which assigns values to random events. Discrete random variables are described by their probability mass function (pmf) and continuous random variables are described by their probability density function (pdf). For example, if X is a random variable representing the roll of a 6-sided die then $X \in \{1, 2, 3, 4, 5, 6\}$ and for a fair die this random variable has pmf $p_X(x) = \mathbb{P}(X = x) = \frac{1}{6}$.

Some important definitions for discrete random variables:

$$\begin{aligned} \mathbb{E}(X) &:= \sum_x x\mathbb{P}(X = x) && \text{expected value of } X \\ \mathbb{E}(f(X)) &:= \sum_x f(x)\mathbb{P}(X = x) && \text{expected value of } f(X) \\ \text{Var}(X) &:= \mathbb{E}(X^2) - \mathbb{E}(X)^2 && \text{variance of } X \end{aligned} \quad (3)$$

Chapter 1: Discrete Time Markov Chains

Markov chains are named after Andrey Markov, a Russian mathematician who did pioneering work in probability and stochastic processes. Discrete time Markov chains are the basis for Markov Chain Monte Carlo (MCMC) methods which were named one of the top 10 algorithms of the 20th century [2] and used today in a variety of applications.

§1.1 Definitions and Examples

Definition 1

A **stochastic process** is defined by a collection of random variables $\{X_t\}_{t \in T}$ all defined on the same probability space. The index set T denotes time; typical choices for T are $T := \mathbb{Z}_{(\geq 0)} = \{0, 1, 2, \dots\}$ for discrete stochastic processes and $T = [0, \infty)$ for continuous stochastic processes.

For simplicity, let's assume that each X_t takes values in a state space S which is countable. To fully specify a stochastic process, we must know its joint distribution function for all finite sets of random variables from the set. For example, we need to know the probabilities

$$\begin{aligned} \mathbb{P}(X_0 = a) \quad \mathbb{P}(X_{10} = b) \quad \mathbb{P}(X_0 = c, X_{21} = d) \\ \mathbb{P}(X_0 = e, X_1 = f, X_3 = g, X_4 = h) \end{aligned} \quad (4)$$

where $a, b, \dots, h \in S$. Generally speaking, we can determine these probabilities from the joint distribution function

$$\mathbb{P}(X_0 = x_0, X_1 = x_1, X_2 = x_2, \dots, X_{n-1} = x_{n-1}, X_n = x_n) \quad \text{where } x_i \in S \quad (5)$$

Applying the definition of conditional probability twice, we have

$$\begin{aligned} \mathbb{P}(X_0 = x_0, X_1 = x_1, X_2 = x_2) \\ = \mathbb{P}(X_0 = x_0) \mathbb{P}(X_1 = x_1, X_2 = x_2 | X_0 = x_0) \\ = \mathbb{P}(X_0 = x_0) \mathbb{P}(X_1 = x_1 | X_0 = x_0) \mathbb{P}(X_2 = x_2 | X_1 = x_1, X_0 = x_0) \end{aligned} \quad (6)$$

Since X_t represents the state at time t , the probability $\mathbb{P}(X_1 = x_1 | X_0 = x_0)$ has the interpretation:

If x_0 happened at $t = 0$, what is the probability that x_1 happens at $t = 1$ and similarly for $\mathbb{P}(X_2 = x_2 | X_1 = x_1, X_0 = x_0)$.

In general, we can write

$$\begin{aligned} \mathbb{P}(X_0 = x_0, X_1 = x_1, X_2 = x_2, \dots, X_{n-1} = x_{n-1}, X_n = x_n) \\ = P(X_0 = x_0) P(X_1 = x_1 | X_0 = x_0) \cdots P(X_n = x_n | X_{n-1} = x_{n-1}, \dots, X_0 = x_0) \end{aligned} \quad (7)$$

Calculating the conditional probability $P(X_n = x_n | X_{n-1} = x_{n-1}, \dots, X_0 = x_0)$ can be very complicated because it depends on ALL the previous times $\{X_0, \dots, X_{n-1}\}$; this is not the case for Markov chains.

Definition 2

Let $\{X_n\}_{n \geq 0}$ be a discrete stochastic processes taking values in a countable sample space S . We say $\{X_n\}_{n \geq 0}$ is a **discrete Markov chain** if for all $n \geq 0$ and all $x_1, \dots, x_{n+1} \in S$:

$$\mathbb{P}(X_{n+1} = x_{n+1} | X_n = x_n, \dots, X_0 = x_0) = \mathbb{P}(X_{n+1} = x_{n+1} | X_n = x_n). \quad (8)$$

Equation 8 is referred to as the “Markov property”.

Another name for the Markov property used in engineering is “memorylessness”. Given all of the previous states $X_0, \dots, X_n$ to determine the conditional distribution of the next state, X_{n+1} , only the current state X_n matters because the system has forgotten about all of the states before X_n .

In this course, we will only consider Markov chains which are “time homogeneous”, meaning that

$$\mathbb{P}(X_{n+1} = x | X_n = y) = \mathbb{P}(X_1 = x | X_0 = y) = p(y, x) \quad \text{for all } x, y \in S \quad (9)$$

The function $p(y, x)$ is a conditional probability so it satisfies

1. $p(y, x) \geq 0$ for all $x, y \in S$, and
2. $\sum_{x \in S} p(y, x) = \sum_{x \in S} \mathbb{P}(X_1 = x | X_0 = y) = 1$ for all $y \in S$

The function $p(y, x)$ defines a matrix P with positive entries and whose rows sum to 1. Importantly, $\sum_{y \in S} p(y, x) = \sum_{y \in S} \mathbb{P}(X_1 = x | X_0 = y)$ which does not necessarily equal 1.

We’ll now go through a couple of examples of Markov chains:

Example 1 (i.i.d. random variables)

Suppose that $\{X_n\}_{n \geq 0}$ are independent, identically distributed (i.i.d.) random variables, then by independence

$$\begin{aligned} \mathbb{P}(X_{n+1} = x_{n+1} | X_n = x_n, \dots, X_1 = x_1) &= \mathbb{P}(X_{n+1} = x_{n+1}) \\ &= \mathbb{P}(X_{n+1} = x_{n+1} | X_n = x_n) \end{aligned} \quad (10)$$

So this defines a Markov chain. The transition function is

$$p(y, x) = \mathbb{P}(X_1 = x | X_0 = y) = \mathbb{P}(X_1 = x) = p_X(x) \quad (11)$$

where $p_X(x)$ is the p.m.f. of X_1 .

Example 2 (Weather Chain)

Let X_n be the weather on day n in sunny California, which we assume is either: 1 = *sunny*, or 2 = *rainy*. Even though the weather is not exactly a Markov chain, we can propose a Markov chain model for the weather by writing down a transition probability matrix

$$P = \begin{pmatrix} .9 & .1 \\ .8 & .2 \end{pmatrix} \quad (12)$$

Q: In the long run, what percentage of days are sunny?

Example 3 (Binary Erasure Channel)

Consider a communications system which attempts to transmit the bits 0 and 1. Each bit transmitted must pass through multiple stages and each stage there is a probability $\theta \in (0, 1)$ that the bit becomes erased (denoted ε) Let X_n denote the digit exiting the n -th stage. Then $\{X_n\}_{n \geq 0}$ is a two-state MC with state space $S = \{0, 1, \varepsilon\}$ and a transition matrix

$$P = \begin{pmatrix} 1 - \theta & 0 & \theta \\ 0 & 1 - \theta & \theta \\ 0 & 0 & 1 \end{pmatrix} \quad (13)$$

Q: After n stages, what's the probability the bit hasn't been erased? What happens as $n \rightarrow \infty$?

Example 4 (Ising Model on Two Spins)

A simple model for magnetism was studied by Ernst Ising in his PhD thesis in 1924; this model is now called the Ising model. The Ising model was completely solved in 1D by Ising in his thesis and about 20 years later, Lars Onsagar solved the Ising model in 2D. Some aspects of the Ising model in 3D remain unsolved to this day.

For two spins, Ising model can be described by a stochastic process $\{X_n\}_{n \geq 0}$ with state space $S = \{ \uparrow \uparrow, \uparrow \downarrow, \downarrow \uparrow, \downarrow \downarrow \}$ and transition matrix

$$P = \begin{pmatrix} p & (1-p)/2 & (1-p)/2 & 0 \\ p/2 & 1-p & 0 & p/2 \\ p/2 & 0 & 1-p & p/2 \\ 0 & (1-p)/2 & (1-p)/2 & p \end{pmatrix} \quad (14)$$

The parameter $p \in [0, 1]$ depends on the ambient temperature. At zero temperature $p = 0$ whereas at high temperatures $p \approx \frac{1}{2}$.

Q: In the long run, does the material favor ferromagnetic states, $\{\uparrow\uparrow, \downarrow\downarrow\}$, or anti-ferromagnetic states, $\{\uparrow\downarrow, \downarrow\uparrow\}$? How does this depend on p ?

§1.2 Multistep Transition Probabilities

Now that we know a bit about how Markov chains are defined, it's time to start looking into how to calculate with them. Our running example, will be the sunny/rainy chain from the previous section which is defined by the transition matrix:

$$P = \begin{pmatrix} .9 & .1 \\ .8 & .2 \end{pmatrix} \quad (15)$$

We start by recalling three equivalent ways of specifying a Markov chain

Conditional Probability	Transition Function	Transition Matrix
$\mathbb{P}(X_{n+1} = x X_n = y)$	$p(y, x)$	$P = \begin{pmatrix} .9 & .1 \\ .8 & .2 \end{pmatrix}$

You should be comfortable with converting between these different ways of specifying Markov chains.

Relating Conditional Probability and the Transition Function

For time homogeneous Markov chains (see Equation 9 above), the conditional probability doesn't depend on the time n , that is:

$$\begin{aligned} \mathbb{P}(X_{n+1} = x | X_n = y) &= \mathbb{P}(X_n = x | X_{n-1} = y) \\ &= \mathbb{P}(X_{n-1} = x | X_{n-2} = y) \\ &= \vdots \\ &= \mathbb{P}(X_1 = x | X_0 = y) \end{aligned} \quad (16)$$

There are many real world systems which satisfy the Markov property but are time inhomogenous but we won't consider them in this class. Because of this homogeneity in time, the conditional probability satisfies:

$$\mathbb{P}(X_{n+1} = x | X_n = y) = p(y, x) \quad \text{for all } n \geq 0 \quad (17)$$

It's important to note that the order of the variables gets reversed when defining the transition function. The mnemonic to remember is that $\mathbb{P}(X_{n+1} = x | X_n = y)$ is the probability of starting at y and going to x so $p(y, x)$ “ $= p(y \rightarrow x)$ ”.

Relating the Transition Function and the Transition Matrix

The entries of the transition matrix correspond to the rows/columns of the transition matrix

$$p(\underbrace{y}_{\text{row}}, \underbrace{x}_{\text{column}}) = P_{y,x} \quad (18)$$

For the sunny/rainy transition matrix

$$p(1, 1) = .9 \quad p(2, 2) = .2 \quad p(1, 2) = .1 \quad p(2, 1) = .8. \quad (19)$$

1.2.1 Predicting the Future

A natural question which arises in studying Markov chains is what the probability of a certain outcome

Question 1

Suppose that it's sunny (1) today, what's the probability it'll be rainy (2) tomorrow?

Answer

By definition, the probability we're looking for is

$$\mathbb{P}(X_1 = 2 | X_0 = 1) = p(1, 2) = .1 \quad (20)$$

Easy!

Now let's ask a more complicated question

Question 2

Suppose that it's sunny (1) today, what's the probability it'll be rainy (2) **two** days from now?

In this case the probability we're looking for is $\mathbb{P}(X_2 = 2 | X_0 = 1)$. Since X_2 is two steps in the future, we cannot directly use the transition matrix (which only has one step). Instead, we use the law of total probability which is proven in Section 2.12.

Answer

By the law of total probability

$$\begin{aligned} & \mathbb{P}(X_2 = 2 | X_0 = 1) \\ &= \sum_{z \in \{1, 2\}} \mathbb{P}(X_2 = 2 | X_1 = z, X_0 = 1) \mathbb{P}(X_1 = z | X_0 = 1) \\ &= \sum_{z \in \{1, 2\}} \mathbb{P}(X_2 = 2 | X_1 = z) \mathbb{P}(X_1 = z | X_0 = 1) \quad (\text{Markov property}) \\ &= \mathbb{P}(X_2 = 2 | X_1 = 1) \mathbb{P}(X_1 = 1 | X_0 = 1) + \mathbb{P}(X_2 = 2 | X_1 = 2) \mathbb{P}(X_1 = 2 | X_0 = 1) \\ &= p(1, 2)p(1, 1) + p(2, 2)p(1, 2) \\ &= (.1)(.9) + (.2)(.1) = \boxed{.11} \end{aligned}$$

Question 3

Suppose that it's sunny (1) today, what's the probability it'll be rainy (2) **three** days from now?

Answer

We're going to approach this same way as the two days, by applying the law of total probability twice.

$$\begin{aligned}\mathbb{P}(X_3 = 2|X_0 = 1) &= \sum_{z_1 \in \{1,2\}} \mathbb{P}(X_3 = 2|X_1 = z_1, X_0 = 1)\mathbb{P}(X_1 = z_1|X_0 = 1) \\ &= \sum_{z_1 \in \{1,2\}} \mathbb{P}(X_3 = 2|X_1 = z_1, X_0 = 1)p(1, z_1)\end{aligned}$$

where in the second line, we've used that since $\{X_n\}_{n \geq 0}$ is a time homogeneous Markov chain $\mathbb{P}(X_1 = z_1|X_0 = 1) = p(1, z_1)$.

To simplify the remaining $\mathbb{P}(X_3 = 2| \dots)$ we can apply the law of total probability again:

$$\begin{aligned}\mathbb{P}(X_3 = 2|X_1 = z_1, X_0 = 1) &= \sum_{z_2 \in \{1,2\}} \mathbb{P}(X_3 = 2|X_2 = z_2, X_1 = z_1, X_0 = 1)\mathbb{P}(X_2 = z_2|X_1 = z_1, X_0 = 1) \\ &= \sum_{z_2 \in \{1,2\}} \mathbb{P}(X_3 = 2|X_2 = z_2)\mathbb{P}(X_2 = z_2|X_1 = z_1) \\ &= \sum_{z_2 \in \{1,2\}} p(z_2, 2)p(z_1, z_2)\end{aligned}$$

It's important to note that *first line* is true for any stochastic process, but the *second line* (which removed the $X_1 = z_1, X_0 = 1$ from the first $\mathbb{P}$ and $X_0 = 1$ from the second $\mathbb{P}$), is only because $\{X_n\}_{n \geq 0}$ forms a Markov chain (and the third line is because of time homogeneity). Putting it all together, we have

$$\begin{aligned}\mathbb{P}(X_3 = 2|X_1 = z_1, X_0 = 1) &= \sum_{z_1 \in \{1,2\}} \left(\sum_{z_2 \in \{1,2\}} p(z_2, 2)p(z_1, z_2) \right) p(1, z_1) \\ &= \sum_{z_1, z_2 \in \{1,2\}} p(z_2, 2)p(z_1, z_2)p(1, z_1) \\ &= \sum_{z_1, z_2 \in \{1,2\}} p(1, z_1)p(z_1, z_2)p(z_2, 2).\end{aligned}$$

For those who know linear algebra, this last line is actually equal to the $(1, 2)$ entry of the matrix $P^3 = P \cdot P \cdot P$; proving this fact will be our next step.

1.2.2 General k -step transition formula

Let A be a $n \times p$ and B be a $p \times m$ matrix, the product of AB is defined to be

$$[AB]_{ij} = \sum_{k=1}^p [A]_{ik} [B]_{kj}. \quad (21)$$

Theorem 1

Let $\{X_n\}_{n \geq 0}$ be a Markov chain with transition matrix P .

$$\mathbb{P}(X_{n+k} = x | X_n = y) = [P^k]_{y,x} \quad (22)$$

PROOF: We'll prove this by mathematical induction. Let's start by thinking about $k = 1$

$$\mathbb{P}(X_{n+1} = x | X_n = y) = p(y, x) = [P]_{y,x} \quad (\text{definition}) \quad (23)$$

As for $k = 2$, by law of total probability

$$\begin{aligned} & \mathbb{P}(X_{n+2} = x | X_n = y) \\ &= \sum_{z \in S} \mathbb{P}(X_{n+2} = x | X_{n+1} = z, X_n = y) \mathbb{P}(X_{n+1} = z | X_n = y) \\ &= \sum_{z \in S} \mathbb{P}(X_{n+2} = x | X_{n+1} = z) \mathbb{P}(X_{n+1} = z | X_n = y) \quad (\text{Markov property}) \\ &= \sum_{z \in S} [P]_{z,x} [P]_{y,z} \quad (\text{definition of the transition matrix}) \\ &= \sum_{z \in S} [P]_{y,z} [P]_{z,x} \\ &= [P^2]_{y,x} \quad (\text{definition of matrix multiplication}) \end{aligned} \quad (24)$$

For $k = 3$, the steps are almost the same:

$$\begin{aligned} & \mathbb{P}(X_{n+3} = x | X_n = y) \\ &= \sum_{z \in S} \mathbb{P}(X_{n+3} = x | X_{n+2} = z, X_n = y) \mathbb{P}(X_{n+2} = z | X_n = y) \\ &= \sum_{z \in S} \mathbb{P}(X_{n+3} = x | X_{n+1} = z) \mathbb{P}(X_{n+2} = z | X_n = y) \\ &= \sum_{z \in S} \mathbb{P}(X_{n+2} = z | X_n = y) [P]_{z,x} \end{aligned} \quad (25)$$

Now notice that we already know that

$$\mathbb{P}(X_{n+2} = z | X_n = y) = [P^2]_{y,z} \quad (26)$$

from the previous step. Therefore,

$$\mathbb{P}(X_{n+3} = x | X_n = y) = \sum_{z \in S} [P^2]_{y,z} [P]_{z,x} = [P^3]_{y,x}. \quad (27)$$

The key trick here is that we used $k = 2$ to prove the formula for $k = 3$. Therefore, for general k

$$\begin{aligned} \mathbb{P}(X_{n+k} = x | X_n = y) &= \sum_{z \in S} \mathbb{P}(X_{n+k} = x | X_{n+k-1} = z, X_n = y) \mathbb{P}(X_{n+k-1} = z | X_n = y) \\ &= \sum_{z \in S} \mathbb{P}(X_{n+k-1} = z | X_n = y) [P]_{z,x} \\ &= \sum_{z \in S} [P^{k-1}]_{y,z} [P]_{z,x} \quad (\text{since we already proved for the previous step}) \\ &= [P^k]_{y,x}. \end{aligned} \quad (28)$$

This completes the proof. ■

1.2.3 Chapman–Kolmogorov Equation and Probabilistic Interpretation

There are a lot of different approaches to proving Equation 22 and above we used the one which ties most directly to linear algebra. Another way of interpreting this result is through the following equation which is known as the Chapman–Kolmogorov equation:

$$\mathbb{P}(X_{m+n} = x | X_0 = y) = \sum_{z \in S} [P^m]_{y,z} [P^n]_{z,x} \quad (29)$$

The way you can see this from Equation 22 is by noting that

$$\mathbb{P}(X_{m+n} = x | X_0 = y) = [P^{m+n}]_{y,x} = \sum_{z \in S} [P^m]_{y,z} [P^n]_{z,x} \quad (30)$$

where the last equality is because of the definition of matrix multiplication. The Chapman–Kolmogorov equation has the following interpretation

Chapman–Kolmogorov Interpretation For a Markov chain, the following are equivalent:

1. We start the chain at y , let it run for $m + n$ steps, and see what the probability we end up at x is.
2. We start the chain at y , let it run m steps and suppose we end up at z . Then we start the chain at z , let it run for n steps, and see what the probability we end up at x is.

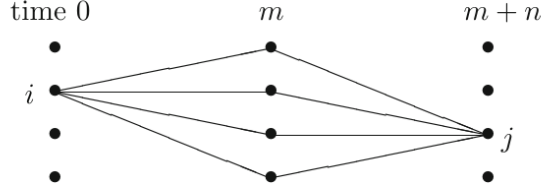


Figure 2: Illustration from the book [3], of the Chapman–Kolmogorov Equation. Here, we start at i and end at j (instead of y and x).

§1.3 Classification of States

Definition 3

From now, given a Markov chain $\{X_n\}_{n \geq 0}$ and for any event A and any $x \in S$, we define

$$\mathbb{P}_x(A) := \mathbb{P}(A|X_0 = x) \quad \text{and} \quad \mathbb{E}_x(A) := \mathbb{E}(A|X_0 = x) \quad (31)$$

So $\mathbb{P}_x(A)$ and $\mathbb{E}_x(A)$ are the probability and expectation of A happening given the chain starts at x .

In Markov chains, there are two classes of states: **recurrent** and **transient**.

Definition 4

Given a Markov chain $\{X_n\}_{n \geq 0}$ on a state space S , we say that:

$$\begin{aligned} x \text{ is } \mathbf{recurrent} &\iff \mathbb{P}_x(\text{we eventually return to } x) = 1, \\ x \text{ is } \mathbf{transient} &\iff \mathbb{P}_x(\text{we eventually return to } x) < 1. \end{aligned} \quad (32)$$

To make sense of recurrence/transience mathematically, we need to first talk about *stopping times* and the *strong Markov property*.

Definition 5

Given a stochastic process $\{X_n\}_{n \geq 0}$ a **stopping time** is any time T so that we can determine whether or not the event $\{T = n\}$ happened by looking at the values of $\{X_0, \dots, X_n\}$.

One important stopping time is the *time of first return*:

Example 5

Given a stochastic process $\{X_n\}_{n \geq 0}$ we define the **time of first return** (to x) as

$$T_x := \min\{n \geq 1 : X_n = x\}. \quad (33)$$

Why is this a stopping time? Consider the following

- $\{T_x = 1\} = \{X_0 = x, X_1 = x\}$
- $\{T_x = 2\} = \{X_0 = x, X_1 \neq x, X_2 = x\}$ because T_x is the FIRST return to x ,
- $\{T_x = 3\} = \{X_0 = x, X_1 \neq x, X_2 \neq x, X_3 = x\}$,
- ...
- $\{T_x = n\} = \{X_0 = x, X_1 \neq x, \dots, X_{n-1} \neq x, X_n = x\}$.

For each n , we can determine whether $\{T_x = n\}$ happened or not by looking at the values of $\{X_0, \dots, X_n\}$.

The reason stopping times are so important is because of the *strong Markov property*:

Theorem 2 (Strong Markov Property, informal)

Let T be a stopping time for a Markov chain $\{X_n\}_{n \geq 0}$. Given $\{T = n\}$ and $\{X_T = x\}$ any other information about the previous times $\{X_0, \dots, X_{T-1}\}$ is irrelevant for predicting the future states X_{T+k} for any $k \geq 0$. In particular, X_{T+k} behaves like a Markov chain with initial state x .

All Markov chains satisfy the strong Markov property so in particular

$$\begin{aligned} \mathbb{P}(X_{T+1} = x | X_T = y, T = n) \\ &= \mathbb{P}(X_{n+1} = x | X_n = y) \quad \text{when } \{X_n\}_{n \geq 0} \text{ is a Markov chain} \\ &= p(y, x) \end{aligned} \quad (34)$$

Caution For general stochastic processes you *cannot* drop the $T = n$. For any stochastic process $\{Y_n\}_{n \geq 0}$ you have

$$\mathbb{P}(Y_{T+1} = y | Y_T = x, T = n) = \mathbb{P}(Y_{n+1} = y | Y_n = x, T = n) \quad (35)$$

however in general $\mathbb{P}(Y_{n+1} = y | Y_n = x, T = n) \neq \mathbb{P}(Y_{n+1} = y | Y_n = x)$.

Proving that Markov chains satisfy the strong Markov property in full generality is beyond the scope of this course, but we'll prove it in a simple case which gives the flavor of the argument. Suppose that T is a stopping time and consider the probability

$$\mathbb{P}(X_{T+1} = x | X_T = y, T = n) \quad (36)$$

Since T is a stopping time, we can write

$$\{T = n\} = \bigcup_{(z_0, \dots, z_n) \in V_n} \{X_0 = z_0, X_1 = z_1, \dots, X_n = z_n\} \quad (37)$$

where V_n is the set of sequences of steps where $\{T = n\}$ is happened. By the law of total probability:

$$\begin{aligned}
 & \mathbb{P}(X_{T+1} = x, X_T = y, T = n) \\
 &= \sum_{(z_0, \dots, z_n) \in V_n} \mathbb{P}(X_{n+1} = x, X_n = y, X_0 = z_0, \dots, X_n = z_n) \\
 &= \sum_{(z_0, \dots, z_n) \in V_n} \mathbb{P}(X_{n+1} = x | X_n = y, X_0 = z_0, \dots, X_n = z_n) \mathbb{P}(X_0 = z_0, \dots, X_n = z_n, X_n = y) \\
 &= \sum_{(z_0, \dots, z_n) \in V_n} \mathbb{P}(X_{n+1} = x | X_n = y) \mathbb{P}(X_0 = z_0, \dots, X_n = z_n, X_n = y) \tag{38} \\
 &= \mathbb{P}(X_{n+1} = x | X_n = y) \left(\sum_{(z_0, \dots, z_n) \in V_n} \mathbb{P}(X_0 = z_0, \dots, X_n = z_n, X_n = y) \right) \\
 &= \mathbb{P}(X_{n+1} = x | X_n = y) \mathbb{P}(X_n = y, T = n)
 \end{aligned}$$

Therefore,

$$\begin{aligned}
 \mathbb{P}(X_{n+1} = x | X_n = y) &= \frac{\mathbb{P}(X_{T+1} = x, X_T = y, T = n)}{\mathbb{P}(X_n = y, T = n)} \\
 &= \mathbb{P}(X_{T+1} = x | X_T = y, T = n).
 \end{aligned} \tag{39}$$

1.3.1 The Time of First Return, Recurrence, and Transience

Definition 6

For all $x, y \in S$, the probability we eventually go to y given we started at x is given by:

$$\rho_{xy} := \mathbb{P}_x(T_y < \infty) \tag{40}$$

Note that

$$\mathbb{P}_x(\text{we eventually return to } x) = \mathbb{P}(T_x < \infty) = \rho_{xx}. \tag{41}$$

We can also consider the time of k^{th} return:

$$T_x^{(k)} := \min\{n \geq T_x^{(k-1)} : X_n = x\}. \tag{42}$$

Question: Why is $T_x^{(k)}$ a stopping time? By the strong Markov property

$$\mathbb{P}(T_x^{(k)} < \infty) = (\mathbb{P}(T_x < \infty))^k = \rho_{xx}^k \tag{43}$$

Before proving this equality, we'll note that fact leads to two cases:

1. If $\rho_{xx} < 1$ then

$$\mathbb{P}(T_x^{(k)} < \infty) = \rho_{xx}^k \rightarrow 0 \text{ as } k \rightarrow \infty \quad (\text{transience}) \tag{44}$$

2. If $\rho_{xx} = 1$ then

$$\mathbb{P}(T_x^{(k)} < \infty) = 1 \text{ for all } k \quad (\text{recurrence}). \quad (45)$$

PROOF: We'll prove this for $k = 2$, the general result is by mathematical induction. The two events

$$\{T_x^{(1)} < \infty\} \cup \{T_x^{(1)} = \infty\} \quad (46)$$

form a partition so by law of total probability

$$\mathbb{P}_x(T_x^{(2)} < \infty) = \mathbb{P}_x(T_x^{(2)} < \infty, T_x < \infty) + \mathbb{P}_x(T_x^{(2)} < \infty, T_x = \infty) \quad (47)$$

But $\mathbb{P}_x(T_x^{(2)} < \infty, T_x = \infty) = 0$ because if the event $\{T_x = \infty\}$ happens, there's no way $\{T_x^{(2)} < \infty\}$ can happen. Therefore,

$$\begin{aligned} \mathbb{P}_x(T_x^{(2)} < \infty) &= \mathbb{P}_x(T_x^{(2)} < \infty, T_x < \infty) \\ &= \sum_{n=1}^{\infty} \mathbb{P}_x(T_x^{(2)} < \infty | T_x = n) \mathbb{P}_x(T_x = n) \end{aligned} \quad (48)$$

But $\mathbb{P}_x(T_x^{(2)} < \infty | T_x = n) = \mathbb{P}_x(T_x < \infty) = \rho_{xx}$ by the strong Markov property. Substituting this back into the previous expression

$$\mathbb{P}_x(T_x^{(2)} < \infty) = \rho_{xx} \sum_{n=1}^{\infty} \mathbb{P}_x(T_x = n) = (\rho_{xx})^2 \quad (49)$$

■

The same argument works when the chain starts somewhere other than x :

Theorem 3

For all $x, y \in S$ and $k \geq 1$,

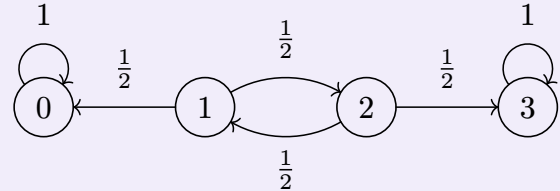
$$\mathbb{P}_y(T_x^{(k)} < \infty) = \rho_{yx} \rho_{xx}^{k-1}. \quad (50)$$

Exercise: Prove Equation 50 for $k = 2$.

Example 6 (Worksheet 1, Problem 5)

Consider the gambler's chain on $S = \{0, 1, 2, 3\}$ with transition matrix and transition diagram

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



For questions about recurrence and transience, only whether $[P]_{x,y} > 0$ or $[P]_{x,y} = 0$ matters, not the exact values of the transition probabilities.

Question 4 (Exercise)

In the chain above, which states are recurrent and which are transient? Why?

Answer

Consider state 1. With probability $\frac{1}{2}$ the chain jumps from 1 to 0, and since 0 is absorbing it never comes back, so $\mathbb{P}_1(T_1 = \infty) \geq \frac{1}{2}$. Since $\mathbb{P}_1(T_1 = \infty) + \mathbb{P}_1(T_1 < \infty) = 1$,

$$\rho_{11} = \mathbb{P}_1(T_1 < \infty) \leq \frac{1}{2} < 1, \quad (51)$$

so 1 is transient. The remaining states are left as an exercise.

Definition 7

We say that x **communicates with** y (denoted $x \rightarrow y$) if

$$\rho_{xy} = \mathbb{P}_x(T_y < \infty) > 0 \quad (52)$$

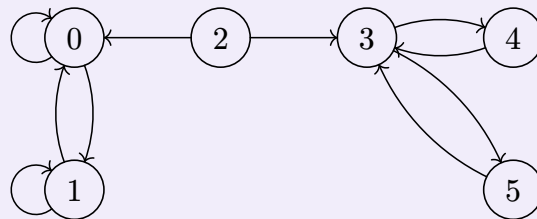
That is, the probability that we eventually reach y starting at x is non-zero.

Remark 1 (Warning)

Communication is not symmetric: in general $\rho_{xy} \neq \rho_{yx}$, just like real life. In the Worksheet 1 chain above, $\rho_{10} > 0$ but $\rho_{01} = 0$, so $1 \rightarrow 0$ but $0 \nrightarrow 1$. On the other hand, $1 \rightarrow 2$ and $1 \rightarrow 3$.

Example 7 (Yet another chain)

Consider the chain with transition diagram



Reading off the arrows we can check that:

$$\begin{aligned} 2 &\rightarrow 0, & 2 &\rightarrow 3, & 2 &\rightarrow 5 \\ 0 &\rightarrow 1, & 1 &\rightarrow 0, & 0 &\nrightarrow 3 \\ 3 &\rightarrow 4, & 4 &\rightarrow 3, & 5 &\rightarrow 3. \end{aligned} \quad (53)$$

Lemma 1

If $x \rightarrow y$ and $y \rightarrow z$, then $x \rightarrow z$.

PROOF: By definition

$$\mathbb{P}_x(T_y < \infty) = \sum_{n=1}^{\infty} \mathbb{P}_x(T_y = n), \quad (54)$$

so if $\rho_{xy} > 0$ then $\mathbb{P}_x(T_y = n) > 0$ for some n . Since $\{T_y = n\} \subseteq \{X_n = y\}$, this means $[P^n]_{x,y} = \mathbb{P}_x(X_n = y) > 0$. Likewise, $\rho_{yz} > 0$ gives $[P^m]_{y,z} > 0$ for some m . Then

$$0 < [P^n]_{x,y} [P^m]_{y,z} \leq \sum_{w \in S} [P^n]_{x,w} [P^m]_{w,z} = [P^{n+m}]_{x,z}, \quad (55)$$

and since $[P^{n+m}]_{x,z} = \mathbb{P}_x(X_{n+m} = z) \leq \mathbb{P}_x(T_z < \infty) = \rho_{xz}$, we get $x \rightarrow z$. ■

Theorem 4

If x is recurrent and $\rho_{xy} > 0$, then $\rho_{yx} = 1$.

PROOF: We prove the equivalent statement: if $\rho_{xy} > 0$ and $\rho_{yx} < 1$, then x is transient. If $\rho_{xy} > 0$ then there must exist a sequence of steps $x, z_1, z_2, \dots, z_k, y$ such that

$$p(x, z_1) > 0, \quad p(z_i, z_{i+1}) > 0, \quad p(z_k, y) > 0. \quad (56)$$

Consider the shortest such path; then $z_i \neq x$ for all i . The chain can follow this path from x to y and then never return to x , so

$$\mathbb{P}_x(T_x = \infty) \geq p(x, z_1)p(z_1, z_2) \cdots p(z_k, y)(1 - \rho_{yx}) > 0, \quad (57)$$

and therefore x is transient. ■

1.3.2 Expected Number of Returns

To decide whether a state is recurrent or transient, it's often easier to count how many times the chain visits it. The number of visits to x (after time 0) is

$$N_x := \sum_{n=1}^{\infty} \mathbb{1}_{\{X_n = x\}} \quad (58)$$

where $\mathbb{1}_A$ is the indicator (characteristic) function of the event A . Taking expectations,

$$\mathbb{E}_y N_x = \sum_{n=1}^{\infty} \mathbb{E}_y(\mathbb{1}_{\{X_n = x\}}) = \sum_{n=1}^{\infty} \mathbb{P}_y(X_n = x) = \sum_{n=1}^{\infty} [P^n]_{y,x}. \quad (59)$$

Theorem 5

$$\mathbb{E}_y N_x = \sum_{k=1}^{\infty} \mathbb{P}_y(N_x \geq k) \quad (60)$$

PROOF: Since N_x takes values in $\{0, 1, 2, \dots\}$,

$$\begin{aligned} \mathbb{E}_y N_x &= \sum_{m=1}^{\infty} m \mathbb{P}_y(N_x = m) \\ &= \underbrace{\mathbb{P}_y(N_x = 1)}_{1 \text{ time}} + \underbrace{\mathbb{P}_y(N_x = 2) + \mathbb{P}_y(N_x = 2)}_{2 \text{ times}} + \underbrace{\mathbb{P}_y(N_x = 3) + \mathbb{P}_y(N_x = 3) + \mathbb{P}_y(N_x = 3)}_{3 \text{ times}} + \dots \end{aligned} \quad (61)$$

Regrouping the terms by columns,

$$\begin{aligned} \mathbb{E}_y N_x &= \mathbb{P}_y(N_x = 1) + \mathbb{P}_y(N_x = 2) + \mathbb{P}_y(N_x = 3) + \dots \\ &\quad + \mathbb{P}_y(N_x = 2) + \mathbb{P}_y(N_x = 3) + \mathbb{P}_y(N_x = 4) + \dots \\ &\quad + \mathbb{P}_y(N_x = 3) + \mathbb{P}_y(N_x = 4) + \dots \\ &= \mathbb{P}_y(N_x \geq 1) + \mathbb{P}_y(N_x \geq 2) + \mathbb{P}_y(N_x \geq 3) + \dots \end{aligned} \quad (62)$$

■

The event $\{N_x \geq k\}$ is the same as $\{T_x^{(k)} < \infty\}$, so by Equation 50

$$\begin{aligned} \mathbb{E}_y N_x &= \sum_{k=1}^{\infty} \mathbb{P}_y(N_x \geq k) \\ &= \sum_{k=1}^{\infty} \mathbb{P}_y(T_x^{(k)} < \infty) \\ &= \sum_{k=1}^{\infty} \rho_{yx} \rho_{xx}^{k-1} \\ &= \begin{cases} 0 & \text{if } \rho_{yx} = 0 \\ \infty & \text{if } \rho_{yx} > 0 \text{ and } \rho_{xx} = 1 \\ \frac{\rho_{yx}}{1 - \rho_{xx}} & \text{if } \rho_{xx} < 1. \end{cases} \end{aligned} \quad (63)$$

In particular, taking $y = x$: x is recurrent if and only if $\mathbb{E}_x N_x = \sum_{n=1}^{\infty} [P^n]_{x,x} = \infty$, and if x is transient then $\mathbb{E}_y N_x < \infty$ for every $y \in S$.

1.3.3 Closed and Irreducible Sets

We want to know what a Markov chain does over long times, and the recurrent states are the only ones which matter for long times. To find recurrent states, we'll start by looking at sets of states which the chain can never leave:

Definition 8

A set of states $C \subseteq S$ is **closed** if once the chain enters C , it can never escape; that is, $p(x, y) = 0$ for all $x \in C$ and $y \notin C$.

Definition 9

A set of states C is **irreducible** if for all $x, y \in C$, $x \rightarrow y$ and $y \rightarrow x$.

For example, in the “yet another chain” example above, $\{0, 1\}$ and $\{3, 4, 5\}$ are closed (and irreducible), while $\{2, 3\}$, $\{4, 5\}$, and $\{0, 1, 2\}$ are not closed. The whole state space S is always closed.

Theorem 6

If C is a finite, closed, irreducible set, then all states in C are recurrent.

PROOF: The proof has two steps.

Step 1. If x is recurrent and $x \rightarrow y$, then y is recurrent.

Since $x \rightarrow y$, $\rho_{xy} > 0$, so by the previous theorem $\rho_{yx} = 1$ and hence $y \rightarrow x$. As in the proof of the lemma, there exist n and m with $[P^n]_{x,y} > 0$ and $[P^m]_{y,x} > 0$. Since x is recurrent,

$$\mathbb{E}_x N_x = \sum_{k=1}^{\infty} [P^k]_{x,x} = \infty. \quad (64)$$

Then, by Equation 59,

$$\begin{aligned} \mathbb{E}_y N_y &= \sum_{k=1}^{\infty} [P^k]_{y,y} \geq \sum_{k=1}^{\infty} [P^{m+k+n}]_{y,y} \quad (\text{index shift}) \\ &= \sum_{k=1}^{\infty} \sum_{z_1, z_2 \in S} [P^m]_{y,z_1} [P^k]_{z_1, z_2} [P^n]_{z_2, y} \\ &\geq \sum_{k=1}^{\infty} [P^m]_{y,x} [P^k]_{x,x} [P^n]_{x,y} \\ &= [P^m]_{y,x} \left(\sum_{k=1}^{\infty} [P^k]_{x,x} \right) [P^n]_{x,y} = \infty, \end{aligned} \quad (65)$$

so y is recurrent.

Step 2. If C is finite and closed, then it has at least one recurrent state.

Suppose all states in C are transient; then $\mathbb{E}_x N_y < \infty$ for all $x, y \in C$. Since C is finite, $\sum_{y \in C} \mathbb{E}_x N_y < \infty$. But since C is closed,

$$\sum_{y \in C} \mathbb{E}_x N_y = \sum_{y \in C} \sum_{k=1}^{\infty} [P^k]_{x,y} = \sum_{k=1}^{\infty} \underbrace{\sum_{y \in C} [P^k]_{x,y}}_{=1} = \infty, \quad (66)$$

a contradiction.

Conclusion. By Step 2, C has a recurrent state x . Since C is irreducible, $x \rightarrow y$ for every $y \in C$, so by Step 1 every state in C is recurrent. ■

Corollary 1

For a Markov chain on a finite state space S , we can decompose

$$S = T \cup C_1 \cup C_2 \cup \dots \cup C_N \quad (67)$$

where T is the set of transient states and each C_i is closed and irreducible.

PROOF: Pick an unmarked state $x \in S$ and mark it.

- If x is transient, add x to T .
- If x is recurrent, find all states y such that $x \rightarrow y$, mark them, and set $C = \{x\} \cup \{y : x \rightarrow y\}$.

Repeat until every state is marked. ■

For the “yet another chain” above, $T = \{2\}$, $C_1 = \{0, 1\}$, and $C_2 = \{3, 4, 5\}$.

§1.4 Stationary Distributions

By Equation 22, if the chain starts at a fixed state x then $\mathbb{P}_x(X_n = y) = [P^n]_{x,y}$. What happens if the initial state is itself random?

Example 8

Suppose $\mathbb{P}(X_0 = 1) = 0.7$ and $\mathbb{P}(X_0 = 2) = 0.3$. What is the distribution of the state at time 3, in terms of P ? By the law of total probability,

$$\begin{aligned} \mathbb{P}(X_3 = 1) &= \mathbb{P}(X_3 = 1 | X_0 = 1) \mathbb{P}(X_0 = 1) + \mathbb{P}(X_3 = 1 | X_0 = 2) \mathbb{P}(X_0 = 2) \\ &= \sum_{x \in \{1,2\}} \mathbb{P}_x(X_3 = 1) \underbrace{\mathbb{P}(X_0 = x)}_{\pi_0(x)} \\ &= \sum_{x \in \{1,2\}} \pi_0(x) [P^3]_{x,1}. \end{aligned} \quad (68)$$

Here $\pi_0(x) := \mathbb{P}(X_0 = x)$ is the **initial distribution**. Writing the distribution at time n as a row vector π_n , the example above shows

$$\pi_3 = \pi_0 P^3, \quad \pi_4 = \pi_0 P^4, \quad \text{and in general} \quad \pi_n = \pi_0 P^n. \quad (69)$$

Definition 10

A distribution π on S is **stationary** if

$$\pi = \pi P, \quad (70)$$

in which case $\pi = \pi P^n$ for all n .

Example 9 (Two-state chain)

Consider

$$P = \begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix}. \quad (71)$$

A stationary distribution must satisfy

$$(\pi(1) \ \pi(2)) \begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix} = (\pi(1) \ \pi(2)), \quad \pi(1) + \pi(2) = 1. \quad (72)$$

This gives the system

$$\begin{aligned} (1-a)\pi(1) + b\pi(2) &= \pi(1) \\ a\pi(1) + (1-b)\pi(2) &= \pi(2) \end{aligned} \quad (73)$$

which simplifies to

$$-a\pi(1) + b\pi(2) = 0. \quad (74)$$

Substituting $\pi(2) = 1 - \pi(1)$:

$$-a\pi(1) + b(1 - \pi(1)) = 0 \implies -(a+b)\pi(1) + b = 0 \implies \pi(1) = \frac{b}{a+b}, \quad \pi(2) = \frac{a}{a+b} \quad (75)$$

For the weather chain from the start of this chapter, $a = .1$ and $b = .8$, giving $\pi(1) = 8/9$ and $\pi(2) = 1/9$. Therefore, in the long run $8/9 \approx 88.9\%$ of days are sunny.

1.4.1 Calculating Stationary Distributions in Practice

There are two natural questions to ask about stationary distributions:

1. Does the Markov chain have a stationary distribution? This is true for all finite, irreducible Markov chains.
2. Starting from anywhere and stepping, will the Markov chain converge to the stationary distribution? This is true for finite, irreducible, and aperiodic Markov chains.

For small chains, we can find the stationary distribution by solving a linear system.

Example 10

Consider the chain with transition matrix

$$P = \begin{pmatrix} .8 & .1 & .1 \\ .2 & .6 & .2 \\ .3 & .3 & .4 \end{pmatrix}. \quad (76)$$

A stationary distribution satisfies

$$(\pi_1 \ \pi_2 \ \pi_3)P = (\pi_1 \ \pi_2 \ \pi_3)I \iff (\pi_1 \ \pi_2 \ \pi_3)(P - I) = 0, \quad (77)$$

that is,

$$(\pi_1 \ \pi_2 \ \pi_3) \begin{pmatrix} -.2 & .1 & .1 \\ .2 & -.4 & .2 \\ .3 & .3 & -.6 \end{pmatrix} = (0 \ 0 \ 0). \quad (78)$$

This gives one equation per column:

$$-.2\pi_1 + .2\pi_2 + .3\pi_3 = 0 \quad (\text{i})$$

$$.1\pi_1 - .4\pi_2 + .3\pi_3 = 0 \quad (\text{ii})$$

$$.1\pi_1 + .2\pi_2 - .6\pi_3 = 0 \quad (\text{iii})$$

Note that (i) + (ii) = -(iii), so these equations are linearly dependent. We also require $\pi_1 + \pi_2 + \pi_3 = 1$, so replacing the third column with ones, the above is equivalent to

$$(\pi_1 \ \pi_2 \ \pi_3) \begin{pmatrix} -.2 & .1 & 1 \\ .2 & -.4 & 1 \\ .3 & .3 & 1 \end{pmatrix} = (0 \ 0 \ 1). \quad (79)$$

Adding 1/2 of column 1 to column 2 and 5 times column 1 to column 3 gives

$$(\pi_1 \ \pi_2 \ \pi_3) \begin{pmatrix} -.2 & 0 & 0 \\ .2 & -.3 & 2 \\ .3 & .45 & 2.5 \end{pmatrix} = (0 \ 0 \ 1), \quad (80)$$

and then adding 20/3 times column 2 to column 3 gives

$$(\pi_1 \ \pi_2 \ \pi_3) \begin{pmatrix} -.2 & 0 & 0 \\ .2 & -.3 & 0 \\ .3 & .45 & 5.5 \end{pmatrix} = (0 \ 0 \ 1). \quad (81)$$

Solving from the last column back to the first:

$$\begin{aligned}
 5.5\pi_3 &= 1 \Rightarrow \boxed{\pi_3 = \frac{1}{5.5}} \\
 -.3\pi_2 + .45 \cdot \frac{1}{5.5} &= 0 \Rightarrow \boxed{\pi_2 = \frac{1.5}{5.5}} \\
 -.2\pi_1 + .2\pi_2 + .3\pi_3 &= 0 \Rightarrow \boxed{\pi_1 = \frac{3}{5.5}}
 \end{aligned}$$

Another approach is to invert the matrix directly:

$$(\pi_1 \ \pi_2 \ \pi_3) = (0 \ 0 \ 1) \begin{pmatrix} -.2 & .1 & 1 \\ .2 & -.4 & 1 \\ .3 & .3 & 1 \end{pmatrix}^{-1}, \quad (82)$$

so the stationary distribution is the last row of the inverse.

§1.5 Simple Random Walks on $\mathbb{Z}^d$

1.5.1 Simple Random Walk on $\mathbb{Z}$

The simple random walk on $\mathbb{Z}$ steps one to the right with probability p and one to the left with probability $1 - p$.

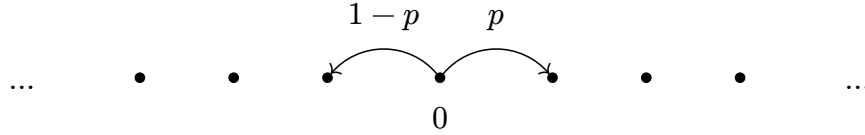


Figure 3: Simple random walk on $\mathbb{Z}$.

Recall that

$$\begin{aligned}
 x \text{ is transient} &\iff \rho_{xx} < 1 \iff \mathbb{E}_x N_x < \infty, \\
 x \text{ is recurrent} &\iff \rho_{xx} = 1 \iff \mathbb{E}_x N_x = \infty.
 \end{aligned} \quad (83)$$

Question: Is zero transient or recurrent?

By Equation 59,

$$\mathbb{E}_0 N_0 = \sum_{n=1}^{\infty} \mathbb{P}_0(X_n = 0) = \sum_{n=1}^{\infty} [P^n]_{0,0}. \quad (84)$$

We make two observations:

1. We are never at zero for n odd.
2. To be at zero, left steps must cancel out right steps, so we need an even number of steps. Of the $2n$ steps, we choose which n go left:

$$\underbrace{\text{L} \ \text{R} \ \text{L} \ \text{L} \ \text{R} \ \text{R}}_{2n \text{ steps}} \Rightarrow \binom{2n}{n} \text{ ways}, \quad (85)$$

so

$$[P^{2n}]_{0,0} = \binom{2n}{n} p^n (1-p)^n = \frac{(2n)!}{n! n!} p^n (1-p)^n. \quad (86)$$

Stirling's Approximation

To understand how $\binom{2n}{n}$ grows, we use Stirling's approximation:

$$\sqrt{2\pi} e^{-n} n^{n+\frac{1}{2}} e^{\frac{1}{12n+1}} < n! < \sqrt{2\pi} e^{-n} n^{n+\frac{1}{2}} e^{\frac{1}{12n}}. \quad (87)$$

For $n = 5$: $e^{\frac{1}{12n+1}} \approx 1.0165$ and $e^{\frac{1}{12n}} \approx 1.0168$, so we usually just take

$$n! \sim \sqrt{2\pi} e^{-n} n^{n+\frac{1}{2}}. \quad (88)$$

Then

$$(2n)! \sim \sqrt{2\pi} e^{-2n} (2n)^{2n+\frac{1}{2}}, \quad (n!)^2 \sim 2\pi e^{-2n} n^{2n+1}, \quad (89)$$

so

$$\begin{aligned} \binom{2n}{n} &\sim \frac{\sqrt{2\pi} e^{-2n} (2n)^{2n+\frac{1}{2}}}{2\pi e^{-2n} n^{2n+1}} \\ &= \frac{1}{\sqrt{2\pi}} 2^{2n+\frac{1}{2}} n^{-\frac{1}{2}} \\ &= \frac{1}{\sqrt{\pi}} 4^n n^{-\frac{1}{2}}. \end{aligned} \quad (90)$$

Therefore

$$[P^{2n}]_{0,0} \sim \frac{1}{\sqrt{\pi}} 4^n n^{-\frac{1}{2}} (p(1-p))^n, \quad (91)$$

and

$$\sum_{n=1}^{\infty} [P^n]_{0,0} = \sum_{n=1}^{\infty} [P^{2n}]_{0,0} \sim \frac{1}{\sqrt{\pi}} \sum_{n=1}^{\infty} n^{-\frac{1}{2}} (4p(1-p))^n. \quad (92)$$

- If $4p(1-p) = 1$, this is

$$\sim \frac{1}{\sqrt{\pi}} \sum_{n=1}^{\infty} \frac{1}{\sqrt{n}} = \infty. \quad (93)$$

- If $4p(1-p) < 1$, this is

$$\sim \frac{1}{\sqrt{\pi}} \sum_{n=1}^{\infty} \frac{1}{\sqrt{n}} (4p(1-p))^n \leq \underbrace{\sum_{n=1}^{\infty} (4p(1-p))^n}_{\text{geometric series}} < \infty. \quad (94)$$

Since $4p(1-p) = 1$ exactly when $p = 1/2$, we conclude: **if $p = 1/2$, then 0 is recurrent;**
if $p \neq 1/2$, then 0 is transient.

Remark 2

If $p \neq 1/2$, all states are transient, but the chain is irreducible.

Remark 3

The result “makes sense” because if $p > 1/2$ then we’re biased to travel to the right (and to the left if $p < 1/2$).

Question: What should the stationary distribution be? Infinity is weird.

1.5.2 Simple Random Walk on $\mathbb{Z}^2$

The simple random walk on $\mathbb{Z}^2$ steps to each of its four neighbors with probability $1/4$.

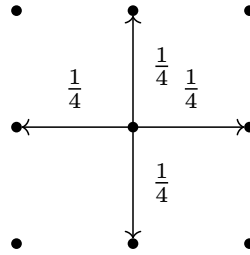


Figure 4: Simple random walk on $\mathbb{Z}^2$.

To return to the origin in $2n$ steps, $2k$ of the steps must be horizontal (with k to the right) and $2(n - k)$ must be vertical (with $n - k$ up), so

$$\begin{aligned}
 [P^{2n}]_{(0,0),(0,0)} &= \left(\sum_{k=0}^n \binom{2n}{2k} \binom{2k}{k} \binom{2(n-k)}{n-k} \right) \left(\frac{1}{4} \right)^{2n} \\
 &= \sum_{k=0}^n \frac{(2n)!}{(2k)! (2(n-k))!} \frac{(2k)!}{(k!)^2} \frac{(2(n-k))!}{((n-k)!)^2} \left(\frac{1}{4} \right)^{2n} \\
 &= \sum_{k=0}^n \frac{(2n)!}{(k!)^2 ((n-k)!)^2} \left(\frac{1}{4} \right)^{2n} \\
 &= \sum_{k=0}^n \frac{(2n)!}{(n!)^2} \frac{n!}{k! (n-k)!} \frac{n!}{k! (n-k)!} \left(\frac{1}{4} \right)^{2n} \\
 &= \binom{2n}{n} \sum_{k=0}^n \binom{n}{k} \binom{n}{n-k} \left(\frac{1}{4} \right)^{2n} \\
 &= \binom{2n}{n} \binom{2n}{n} \left(\frac{1}{4} \right)^{2n}.
 \end{aligned} \tag{95}$$

1.5.3 Comparing Dimensions

For $\mathbb{Z}$ (with $p = 1/2$):

$$[P^{2n}]_{0,0} = \left(\frac{1}{2}\right)^{2n} \binom{2n}{n} \sim \frac{1}{\sqrt{\pi n}}. \quad (96)$$

For $\mathbb{Z}^2$:

$$[P^{2n}]_{0,0} = \left(\frac{1}{2}\right)^{4n} \binom{2n}{n}^2 \sim \frac{1}{\pi n}. \quad (97)$$

Since $\sum_{n=1}^{\infty} \frac{1}{\pi n} = \infty$, the walk on $\mathbb{Z}^2$ is still recurrent. For $\mathbb{Z}^d$:

$$[P^{2n}]_{0,0} \sim C_d n^{-d/2} \implies \sum_{n=1}^{\infty} \frac{C_d}{n^{d/2}} \text{ is summable iff } d > 2. \quad (98)$$

Theorem 7

Simple random walk on $\mathbb{Z}^d$ is recurrent if $d = 1$ or $d = 2$, but transient otherwise.

§1.6 Positive and Null Recurrence

Recall that a state x is recurrent if $\mathbb{P}_x(T_x < \infty) = 1$. We now split recurrent states into two types.

Definition 11

A state x is said to be **positive recurrent** if $\mathbb{E}_x T_x < \infty$. Otherwise, it is **null recurrent**.

Lemma 2

If x is positive recurrent and $x \rightarrow y$, then y is positive recurrent. The same is true for null recurrent.

In particular, if C is a closed, irreducible set of recurrent states, then the states in C are either all positive recurrent or all null recurrent.

Lemma 3

For a finite, irreducible Markov chain, all states are positive recurrent.

Remark 4

In the simple random walks on $\mathbb{Z}$ and $\mathbb{Z}^2$ (with $p = 1/2$), all states are null recurrent.

1.6.1 The Expected Time of Return

So far we've seen three quantities which describe how a chain returns to a state x :

- $\rho_{xx} = \mathbb{P}_x(T_x < \infty)$, the probability we eventually return: x is transient $\iff \rho_{xx} < 1$ and x is recurrent $\iff \rho_{xx} = 1$.

- $\mathbb{E}_x N_x$, the expected number of returns: x is transient $\iff \mathbb{E}_x N_x < \infty$ and x is recurrent $\iff \mathbb{E}_x N_x = \infty$.
- $\mathbb{E}_x T_x$, the expected time until we return:
 - if x is transient, then $\mathbb{E}_x T_x = \infty$;
 - if x is null recurrent, then $\mathbb{E}_x T_x = \infty$;
 - if x is positive recurrent, then $\mathbb{E}_x T_x < \infty$.

What does $\mathbb{E}_x T_x = \infty$ mean? Either we might never return, *or* it's likely we'll take a long time to return.

1.6.2 Stationary Distributions and the Expected Time of Return

Theorem 8

For an irreducible, positive recurrent Markov chain, the stationary distribution is

$$\pi(x) = \frac{1}{\mathbb{E}_x T_x}. \quad (99)$$

Does this make sense? The longer it takes for us to return to x , the less time we spend in x . **PROOF:** Fix a state x and define

$$\mu_x(y) := \mathbb{E}_x \left(\sum_{n=0}^{T_x-1} \mathbb{1}_{\{X_n=y\}} \right) = \sum_{n=0}^{\infty} \mathbb{P}_x(X_n = y, T_x > n). \quad (100)$$

So $\mu_x(y)$ is the expected number of visits to y during the times $\{0, 1, 2, \dots, T_x - 1\}$. Note that $\mu_x(x) = 1$, since the only visit to x during these times is at time 0.

If we compute $\mu_x P$, where $\mu_x = (\mu_x(1) \ \mu_x(2) \ \dots)$ is a row vector, then the y th entry is the expected number of visits to y during the times $\{1, 2, \dots, T_x\}$. Since $X_{T_x} = x$, just like $X_0 = x$, these two counts are the same, so

$$\mu_x P = \mu_x. \quad (101)$$

We want to make μ_x a probability distribution, so we compute

$$\sum_y \mu_x(y) = \sum_y \mathbb{E}_x \sum_{n=0}^{T_x-1} \mathbb{1}_{\{X_n=y\}} = \mathbb{E}_x \left(\sum_{n=0}^{T_x-1} \underbrace{\sum_y \mathbb{1}_{\{X_n=y\}}}_{=1} \right) = \mathbb{E}_x T_x, \quad (102)$$

which is finite because the chain is positive recurrent. Therefore

$$\pi(y) := \mu_x \frac{y}{\sum_z \mu_x(z)} = \mu_x \frac{y}{\mathbb{E}_x T_x} \quad (103)$$

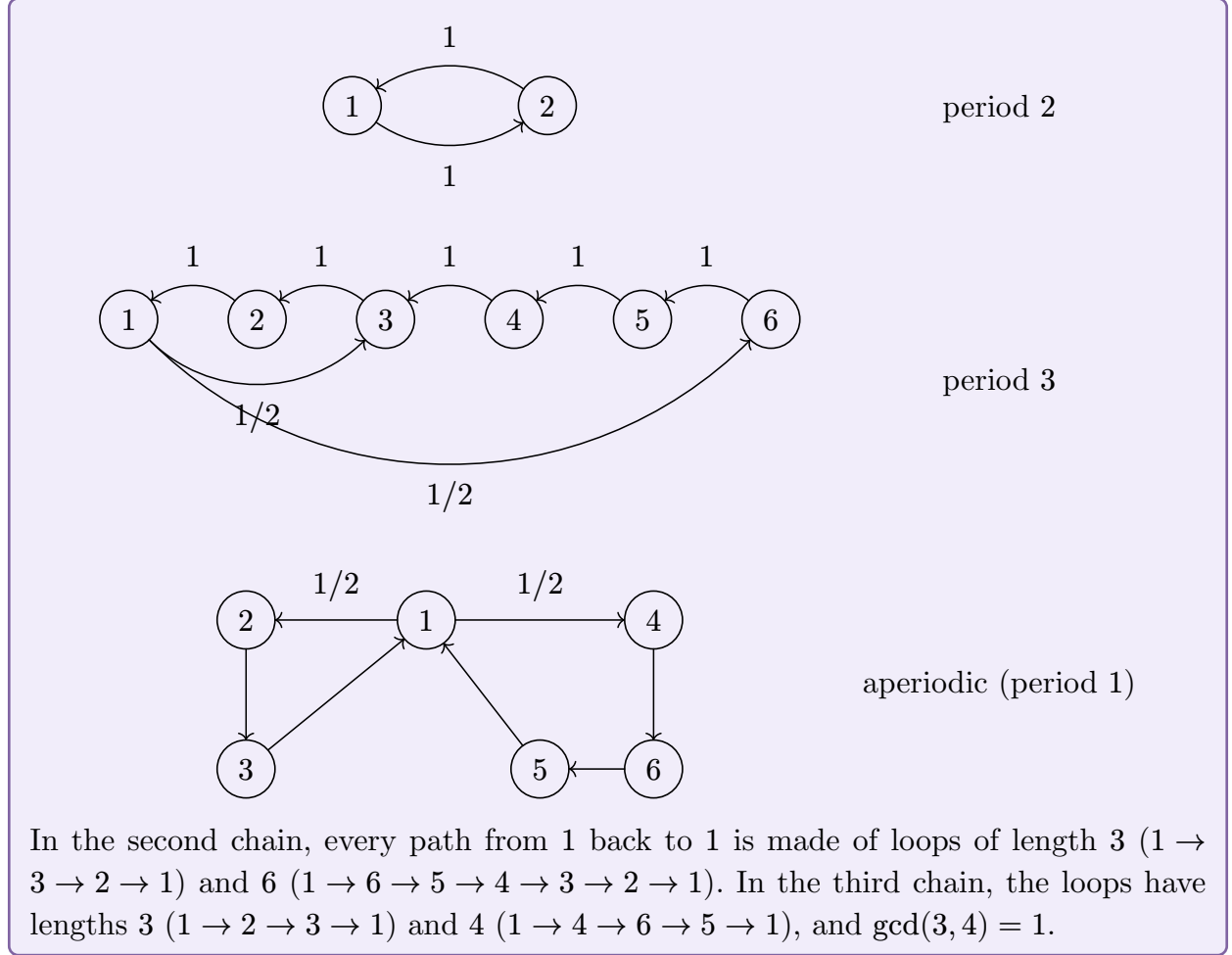
is a stationary distribution. Taking $y = x$, we get $\pi(x) = 1/\mathbb{E}_x T_x$. (Since an irreducible chain has only one stationary distribution, this π doesn't depend on which x we started with, so the formula holds for every x .)



§1.7 Periodicity and Limiting Behavior

1.7.1 Periodicity

Example 11 (Three examples)



Definition 12

Let $I_x := \{n \geq 1 : [P^n]_{x,x} > 0\}$. The **period** of x is

$$d_x := \gcd(I_x). \quad (104)$$

If $d_x = 1$, then we say x is **aperiodic**.

Lemma 4

If $[P]_{x,x} > 0$, then x is aperiodic.

PROOF: $1 \in I_x$, so $\gcd(I_x) = 1$.

■

Lemma 5

If $\rho_{xy} > 0$ and $\rho_{yx} > 0$, then x and y have the same period.

PROOF: Since $x \rightarrow y$, there exists m with $[P^m]_{x,y} > 0$, and since $y \rightarrow x$, there exists n with $[P^n]_{y,x} > 0$. Then

$$0 < [P^m]_{x,y} [P^n]_{y,x} \leq \sum_z [P^m]_{x,z} [P^n]_{z,x} = [P^{m+n}]_{x,x}, \quad (105)$$

so $m+n \in I_x$ and d_x divides $m+n$. Now pick any $\ell \in I_y$. Then

$$0 < [P^m]_{x,y} [P^\ell]_{y,y} [P^n]_{y,x} \leq [P^{m+\ell+n}]_{x,x}, \quad (106)$$

so d_x divides $m+\ell+n$. Since d_x also divides $m+n$, it divides ℓ . This holds for every $\ell \in I_y$, so d_x divides $\gcd(I_y) = d_y$, and hence $d_x \leq d_y$. Repeating the same steps with x and y swapped gives $d_y \leq d_x$, so $d_x = d_y$.

■

Remark 5

For practical purposes, people often replace P with

$$\tilde{P} := \frac{1}{2}I + \frac{1}{2}P. \quad (107)$$

This is also a transition matrix, and $[\tilde{P}]_{x,x} \geq 1/2 > 0$, so it is aperiodic. It also has the same stationary distributions as P , since $\pi\tilde{P} = \frac{1}{2}\pi + \frac{1}{2}\pi P$. $\tilde{P}$ is called the **lazy** version of P .

1.7.2 Limiting Behavior

Theorem 9 (Convergence to the stationary distribution)

For an irreducible, aperiodic, positive recurrent Markov chain,

$$\lim_{n \rightarrow \infty} [P^n]_{x,y} = \pi(y) \quad \text{for all } x, y. \quad (108)$$

The limit does *not* depend on the starting state x !

§1.8 Doubly Stochastic Matrices

Recall that a transition matrix P has two properties:

1. all entries are ≥ 0 , and
2. the rows sum to 1.

For example, $\begin{pmatrix} .8 & .2 \\ .7 & .3 \end{pmatrix}$.

Definition 13

If the columns of P also sum to 1, then we say P is **doubly stochastic**.

Lemma 6

If P is an $N \times N$ doubly stochastic matrix, then $\pi(x) = 1/N$ for all x is a stationary distribution.

PROOF: For every y ,

$$[\pi P]_y = \sum_x \pi(x)[P]_{x,y} = \frac{1}{N} \underbrace{\sum_x [P]_{x,y}}_{\text{column sum}=1} = \frac{1}{N} = \pi(y). \quad (109)$$

■

§1.9 Detailed Balance and Reversibility

Definition 14

We say a Markov chain satisfies **detailed balance** with respect to π if

$$\pi(x)[P]_{x,y} = \pi(y)[P]_{y,x} \quad \text{for all } x, y. \quad (110)$$

One way to picture this is to imagine a pile of dirt of size $\pi(x)$ sitting on each state x . Then $\pi(1)[P]_{1,2}$ is how much dirt goes from $1 \rightarrow 2$ in one step, and $\pi(2)[P]_{2,1}$ is how much goes from $2 \rightarrow 1$. Detailed balance says that for every pair of states, these two amounts are equal.

Lemma 7

If π satisfies detailed balance, then π is a stationary distribution.

PROOF: The x th entry of πP is

$$[\pi P]_x = \sum_y \pi(y)[P]_{y,x} = \sum_y \pi(y)[P]_{x,y} = \pi(x) \underbrace{\sum_y [P]_{x,y}}_{\text{row sum}=1} = \pi(x). \quad (111)$$

■

Remark 6 (Warning)

Not all Markov chains satisfy detailed balance. Consider

$$P = \begin{pmatrix} .5 & .5 & 0 \\ .3 & .1 & .6 \\ .2 & .4 & .4 \end{pmatrix}. \quad (112)$$

If π satisfied detailed balance, then

$$\begin{aligned}\pi(1)[P]_{1,3} &= \pi(3)[P]_{3,1} \implies 0 = .2\pi(3) \implies \pi(3) = 0, \\ \pi(2)[P]_{2,3} &= \pi(3)[P]_{3,2} \implies .6\pi(2) = 0 \implies \pi(2) = 0, \\ \pi(1)[P]_{1,2} &= \pi(2)[P]_{2,1} \implies .5\pi(1) = 0 \implies \pi(1) = 0,\end{aligned}\tag{113}$$

so $\pi = 0$, which is not a distribution. But this matrix is doubly stochastic, so $\pi = (1/3, 1/3, 1/3)$ is a stationary distribution.

1.9.1 Reversibility

Usually we consider $X_0, X_1, \dots$ forward in time, which (assuming the chain is irreducible, aperiodic, and positive recurrent) always converges to the stationary distribution:

$$(\text{anything}) \longrightarrow \pi.\tag{114}$$

Sometimes it's useful to consider the **reversed chain**

$$Y_n := X_{N-n}, \quad \text{with } X_0 \sim \pi,\tag{115}$$

to understand how the chain got to π .

Theorem 10

If $\{X_n\}_{n \geq 0}$ is irreducible, aperiodic, and positive recurrent with stationary distribution π and $X_0 \sim \pi$, then for any $N \geq 0$, $\{Y_n\}_{0 \leq n \leq N}$ is also a Markov chain with

$$\mathbb{P}(Y_{n+1} = x | Y_n = y) = \frac{\pi(x)[P]_{x,y}}{\pi(y)}.\tag{116}$$

Usually, we denote

$$[\hat{P}]_{x,y} := \frac{\pi(y)[P]_{y,x}}{\pi(x)}\tag{117}$$

and call $\hat{P}$ the **reversed** (or **dual**) chain.

Definition 15

A chain is **reversible** if $\hat{P} = P$.

Lemma 8

A Markov chain is reversible if and only if it satisfies detailed balance.

PROOF:

$$[\hat{P}]_{x,y} = [P]_{x,y} \iff \frac{\pi(y)[P]_{y,x}}{\pi(x)} = [P]_{x,y} \iff \pi(y)p(y,x) = \pi(x)p(x,y).\tag{118}$$

■

§1.10 The Metropolis–Hastings Algorithm

Suppose we want to sample from a distribution π , and we have a Markov chain $p(x, y)$ we can easily simulate (the *proposal* chain). The Metropolis–Hastings algorithm (Metropolis et al. 1953, generalized by Hastings in 1970) modifies p so that π becomes the stationary distribution:

1. Start at a state x .
2. Propose a move to y (with probability $p(x, y)$) and let

$$r(x, y) = \min\left\{\frac{\pi(y)p(y, x)}{\pi(x)p(x, y)}, 1\right\}. \quad (119)$$

3. Pick a random number $z \in [0, 1]$.
 - If $z \leq r(x, y)$, accept the proposal and move to y .
 - If $z > r(x, y)$, reject the proposal and stay at x .
4. Repeat.

The resulting Markov chain has transition probabilities

$$p_{\text{MH}}(x, y) := \begin{cases} p(x, y)r(x, y) & x \neq y \\ 1 - \sum_{z \neq x} p_{\text{MH}}(x, z) & x = y. \end{cases} \quad (120)$$

Theorem 11

π satisfies detailed balance for p_{MH} , so it is a stationary distribution.

PROOF: We want to show $\pi(x)p_{\text{MH}}(x, y) = \pi(y)p_{\text{MH}}(y, x)$ for $x \neq y$, that is,

$$\pi(x)p(x, y)r(x, y) = \pi(y)p(y, x)r(y, x), \quad (121)$$

or

$$\pi(x)p(x, y) \min\left\{\frac{\pi(y)p(y, x)}{\pi(x)p(x, y)}, 1\right\} = \pi(y)p(y, x) \min\left\{\frac{\pi(x)p(x, y)}{\pi(y)p(y, x)}, 1\right\}. \quad (122)$$

Suppose $\pi(x)p(x, y) \leq \pi(y)p(y, x)$ (the other case is the same with x and y swapped). Then the left minimum is 1 and the right minimum is the fraction, so this becomes

$$\begin{aligned} \pi(x)p(x, y) &= \pi(y)p(y, x) \frac{\pi(x)p(x, y)}{\pi(y)p(y, x)} \\ \pi(x)p(x, y) &= \pi(x)p(x, y). \quad \checkmark \end{aligned} \quad (123)$$

■

Application. Run Metropolis–Hastings for a long time to generate $\{X_n\}$. By the law of large numbers for Markov chains,

$$\frac{1}{N} \sum_{n=1}^N f(X_n) \rightarrow \sum_x f(x) \pi(x). \quad (124)$$

Observation 1 (Why this is useful)

- **Only ratios of π appear.** In $\pi(y)/\pi(x) = w(y)/w(x)$ the constant Z cancels, so we never need it.
- **Symmetric proposals.** If $p(x, y) = p(y, x)$ the rule simplifies to $r(x, y) = \min\{\pi(y)/\pi(x), 1\}$: always accept a move to a more likely state, and accept a move to a less likely state with probability equal to the ratio. This is the original Metropolis algorithm.
- **Rejections count.** When a proposal is rejected the chain stays at x , and that repeated visit is part of the sample.

Remark 7 (Using the chain)

If the chain is irreducible and the state space is finite, then by the law of large numbers for Markov chains

$$\frac{1}{N} \sum_{n=1}^N f(X_n) \rightarrow \sum_x f(x) \pi(x). \quad (125)$$

To get one sample from π , run the chain for a long time. To get many, record the state at widely separated times. Deciding how long is “long enough” is the hard part in practice.

Remark 8 (Choosing the proposal)

Any proposal that can reach every state, with $p(x, y) > 0$ whenever $p(y, x) > 0$, gives the right stationary distribution. The choice affects only the speed. A huge amount of effort has been spent trying to figure out proposals which mix quickly.

1.10.1 The Ising Model: Setup

We now work an example by hand.

Example 12 (Ising model on a 3×3 grid)

Consider a 3×3 grid Λ with nine sites, which we label

1	2	3
4	5	6
7	8	9

A state is a configuration $\sigma : \Lambda \rightarrow \{-1, +1\}$, so there are $2^9 = 512$ states. Sites are neighbors ($y \sim x$) if they are adjacent horizontally or vertically, so corners have 2 neighbors, edge sites 3, and the center 4. There are 12 neighboring pairs (bonds).

The target distribution is

$$\pi(\sigma) = \frac{1}{Z(\beta)} \exp(-\beta H(\sigma)), \quad H(\sigma) = - \sum_{x \sim y} \sigma_x \sigma_y, \quad (126)$$

where the sum counts each bond once. We never compute $Z(\beta)$.

Proposal. Pick a site x uniformly at random and propose flipping it: $p(\sigma, \sigma^x) = 1/9$, where σ^x is σ with the spin at x reversed. Since p is symmetric, the acceptance probability is

$$r(\sigma, \sigma^x) = \min \left\{ 1, \frac{\pi(\sigma^x)}{\pi(\sigma)} \right\}. \quad (127)$$

Proposition 1 (The ratio is local)

Flipping x changes the sign of exactly the terms $\sigma_x \sigma_y$ with $y \sim x$, so

$$\frac{\pi(\sigma^x)}{\pi(\sigma)} = \exp(-2\beta S_x(\sigma)), \quad S_x(\sigma) = \sigma_x \sum_{y \sim x} \sigma_y. \quad (128)$$

In words, $S_x = \#\{\text{neighbors agreeing with } x\} - \#\{\text{neighbors disagreeing with } x\}$, and the flip changes the energy by $\Delta H = 2S_x$.

Remark 9 (Choice of β)

Take $\beta = \frac{1}{2} \ln 2 \approx 0.347$, so that $e^{-2\beta} = 1/2$ and

$$r(\sigma, \sigma^x) = \min \{ 1, 2^{-S_x(\sigma)} \}. \quad (129)$$

Every acceptance probability is then one of $1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}$:

S_x	≤ 0	1	2	3	4
r	1	1/2	1/4	1/8	1/16

Summary 1 (One step of the chain)

1. Pick a site x uniformly from $\{1, \dots, 9\}$.
2. Compute S_x by looking only at x and its neighbors.
3. Draw $U \sim \text{Uniform}(0, 1)$. If $U < 2^{-S_x}$, flip the spin at x ; otherwise stay put.

1.10.2 The Ising Model Example

Start from

$$X_0 = \begin{array}{|c|c|c|} \hline + & + & - \\ \hline \end{array} \quad H(X_0) = -2$$

+	-	-
+	+	-

(of the 12 bonds, 7 agree and 5 disagree, so $H = -(7 - 5) = -2$). Use the following pre-drawn randomness.

Step n	1	2	3	4	5	6
Site x	5	1	6	3	5	8
U	0.83	0.62	0.31	0.95	0.18	0.44

Step 1. Propose site 5 (center, spin $-$). Neighbors 2, 4, 6, 8 have spins $+, +, -, +$, sum $= +2$, so $S_5 = (-1)(2) = -2$. Ratio $2^2 = 4 \geq 1$: **accept** (no need to look at U). $\Delta H = -4$.

+	+	-
+	-	-
+	+	-

 $\rightarrow$

+	+	-
+	+	-
+	+	-

 $H(X_1) = -6$

Step 2. Propose site 1 (corner, spin $+$). Neighbors 2, 4 are $+, +$, so $S_1 = 2$ and $r = 1/4$. $U = 0.62 > 0.25$: **reject**. The chain stays where it is, and this still counts as a step.

+	+	-
+	+	-
+	+	-

 $\rightarrow$

+	+	-
+	+	-
+	+	-

 $H(X_2) = -6$

Step 3. Propose site 6 (edge, spin $-$). Neighbors 3, 5, 9 are $-, +, -$, sum $= -1$, so $S_6 = (-1)(-1) = 1$ and $r = 1/2$. $U = 0.31 < 0.5$: **accept**, even though the energy goes up. $\Delta H = +2$.

+	+	-
+	+	-
+	+	-

 $\rightarrow$

+	+	-
+	+	+
+	+	-

 $H(X_3) = -4$

Step 4. Propose site 3 (corner, spin $-$). Neighbors 2, 6 are $+, +$, so $S_3 = (-1)(2) = -2$. Ratio $4 \geq 1$: **accept**. $\Delta H = -4$.

+	+	-
+	+	+
+	+	-

 $\rightarrow$

+	+	+
+	+	+
+	+	-

 $H(X_4) = -8$

Step 5. Propose site 5 (center, spin $+$). All four neighbors are $+$, so $S_5 = 4$ and $r = 1/16 = 0.0625$. $U = 0.18 > 0.0625$: **reject**.

+	+	+
---	---	---

 $\rightarrow$

+	+	+
---	---	---

 $H(X_5) = -8$

+	+	+
+	+	-

+	+	+
+	+	-

Step 6. Propose site 8 (edge, spin +). Neighbors 5, 7, 9 are +, +, -, sum = +1, so $S_8 = 1$ and $r = 1/2$. $U = 0.44 < 0.5$: **accept**. $\Delta H = +2$.

+	+	+
+	+	+
+	+	-

 $\rightarrow$

+	+	+
+	+	+
+	-	-

 $H(X_6) = -6$

Summary 2 (The whole run)

n	site x	σ_x	$\sum_{y \sim x} \sigma_y$	S_x	r	U	result
1	5	-	+2	-2	1	0.83	accept
2	1	+	+2	2	1/4	0.62	reject
3	6	-	-1	1	1/2	0.31	accept
4	3	-	+2	-2	1	0.95	accept
5	5	+	+4	4	1/16	0.18	reject
6	8	+	+1	1	1/2	0.44	accept

Energies: $H(X_0), \dots, H(X_6) = -2, -6, -6, -4, -8, -8, -6$.

Question 5

What happens to the chain as $\beta \rightarrow 0$? As $\beta \rightarrow \infty$?

Answer

As $\beta \rightarrow 0$ every ratio tends to 1, every proposal is accepted, and π is uniform on the 512 states. As $\beta \rightarrow \infty$ energy-raising moves are never accepted and the chain only moves downhill or sideways, with π concentrating on the two ground states (all +, all -).

§1.11 Exit Distributions and Hitting Probabilities

Recall the time of first return $T_x = \min\{n \geq 1 : X_n = x\}$. For a set of states we also want to allow “hitting” at time 0.

Definition 16

For a target set $A \subseteq S$, the **hitting time** and **return time** of A are

$$H_A := \min\{n \geq 0 : X_n \in A\}, \quad T_A := \min\{n \geq 1 : X_n \in A\}, \quad (130)$$

with the convention that $\min \emptyset = \infty$. The **hitting probability** of A starting from x is

$$h(x) := h_A(x) = \mathbb{P}_x(H_A < \infty). \quad (131)$$

Note that if $X_0 \in A$ then $H_A = 0$ whereas $T_A \geq 1$ always; if $X_0 \notin A$ then $H_A = T_A$. When $A = \{x\}$ we just write H_x and T_x .

Theorem 12

The hitting probabilities satisfy the linear system

$$\begin{aligned} h(x) &= \sum_{y \in S} p(x, y) h(y) && \text{if } x \notin A, \\ h(x) &= 1 && \text{if } x \in A \quad (\text{boundary condition}). \end{aligned} \quad (132)$$

PROOF: If $x \in A$ then $H_A = 0$, so $h(x) = 1$. For $x \notin A$, condition on the first step. By the law of total probability,

$$\begin{aligned} h(x) &= \mathbb{P}_x(H_A < \infty) = \sum_{y \in S} \mathbb{P}(H_A < \infty | X_1 = y, X_0 = x) \mathbb{P}(X_1 = y | X_0 = x) \\ &= \sum_{y \in S} \mathbb{P}_y(H_A < \infty) p(x, y) \quad (\text{Markov property}) \\ &= \sum_{y \in S} p(x, y) h(y). \end{aligned} \quad (133)$$

In the second line we used that, since $x \notin A$, hitting A from time 0 onwards is the same as hitting A from time 1 onwards, and from time 1 the chain behaves like a fresh chain started at y . ■

Remark 10

The equations Equation 132 only depend on the transition probabilities *out of* the states not in A .

Example 13 (Worksheet 1 chain, revisited)

Consider again the gambler's chain on $S = \{0, 1, 2, 3\}$ where 0 and 3 are absorbing and from 1 and 2 we step left or right with probability $\frac{1}{2}$ each. Starting from 1, what is the probability of getting trapped in 3?

Let $h(i) = \mathbb{P}_i(H_3 < \infty)$, so $h(3) = 1$. By Equation 132,

$$\begin{aligned}
 h(2) &= \frac{1}{2}h(3) + \frac{1}{2}h(1) = \frac{1}{2} + \frac{1}{2}h(1) \\
 h(1) &= \frac{1}{2}h(2) + \frac{1}{2}h(0) \\
 h(0) &= h(0)
 \end{aligned} \tag{134}$$

The third equation is useless, so we have two equations for three unknowns. However, the chain can never leave 0, so $h(0) = 0$ (an extra condition). Solving the remaining two equations gives $h(1) = 1/3$ and $h(2) = 2/3$.

Example 14 (Two year college)

At a local two year college, 60% of freshmen become sophomores, 25% remain freshmen, and 15% drop out. 70% of sophomores graduate and transfer to a four year college, 20% remain sophomores, and 10% drop out. What fraction of new students eventually graduate?

Let 1 = freshman, 2 = sophomore, G = graduate, D = dropout. With the states in this order,

$$P = \begin{pmatrix} .25 & .6 & 0 & .15 \\ 0 & .2 & .7 & .1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{135}$$

Let $h(x) = \mathbb{P}_x(H_G < \infty)$, so $h(G) = 1$ and $h(D) = 0$. Considering what happens in one step:

$$\begin{aligned}
 h(1) &= .25h(1) + .6h(2) \\
 h(2) &= .2h(2) + .7
 \end{aligned} \tag{136}$$

so $h(2) = 7/8$ and $h(1) = 0.7$. That is, 70% of new students eventually graduate.

Example 15 (Gambler's ruin)

A gambler wins \$1 with probability p and loses \$1 with probability $q = 1 - p$, stopping at \$0 or \$ N . Starting with \$ i , what is the probability the gambler's fortune reaches \$ N before \$0?

This is a Markov chain on $\{0, 1, \dots, N\}$ with

$$p(0, 0) = p(N, N) = 1, \quad p(i, i+1) = p, \quad p(i, i-1) = q \quad \text{for } 0 < i < N. \tag{137}$$

The closed irreducible sets are $\{0\}$ and $\{N\}$, and the states $\{1, \dots, N-1\}$ are transient. Let $h(i) = \mathbb{P}_i(H_N < \infty)$, so $h(0) = 0$, $h(N) = 1$, and for $1 \leq i \leq N-1$

$$h(i) = ph(i+1) + qh(i-1). \tag{138}$$

Since $p + q = 1$, we can rewrite this as $p(h(i + 1) - h(i)) = q(h(i) - h(i - 1))$, so

$$\begin{aligned}
 h(i + 1) - h(i) &= \left(\frac{q}{p}\right)(h(i) - h(i - 1)) \\
 &= \left(\frac{q}{p}\right)^2 (h(i - 1) - h(i - 2)) \\
 &= \vdots \\
 &= \left(\frac{q}{p}\right)^i (h(1) - h(0)) \\
 &= \left(\frac{q}{p}\right)^i h(1).
 \end{aligned} \tag{139}$$

Adding these differences together (telescoping series),

$$\begin{aligned}
 h(i) &= \sum_{k=0}^{i-1} (h(k + 1) - h(k)) \\
 &= h(1) \sum_{k=0}^{i-1} \left(\frac{q}{p}\right)^k \\
 &= \begin{cases} \frac{1-(q/p)^i}{1-(q/p)} h(1) & \text{if } p \neq q \\ i h(1) & \text{if } p = q. \end{cases}
 \end{aligned} \tag{140}$$

Using $h(N) = 1$ to solve for $h(1)$, we get

$$h(i) = \begin{cases} \frac{1-(q/p)^i}{1-(q/p)^N} & \text{if } p \neq q \\ i/N & \text{if } p = q. \end{cases} \tag{141}$$

For example, if $p = 0.45$ and $N = 20$ then $h(10) \approx 0.1185$, while if $p = 0.45$ and $N = 200$ then $h(100) \approx 1.9 \times 10^{-9}$.

§1.12 Proof of the Law of Total Probability

We'll start with the abstract statement of the law of total probability

Theorem 13 (Law of Total Probability)

Let A be any event and let $\{B_i\}_i$ be a (countable) partition of probability space Ω , meaning it satisfies

$$\bigcup_i B_i = \Omega \quad \text{and} \quad \mathbb{P}(B_i \cap B_j) = 0 \text{ for all } i, j \tag{142}$$

then

$$\mathbb{P}(A) = \sum_i \mathbb{P}(A|B_i)\mathbb{P}(B_i) = \sum_i \mathbb{P}(A \cap B_i) \quad (143)$$

Additionally, if C is any event with $\mathbb{P}(C) \neq 0$ then

$$\mathbb{P}(A|C) = \sum_i \mathbb{P}(A|C \cap B_i)\mathbb{P}(B_i|C) \quad (144)$$

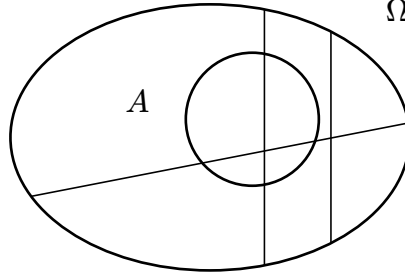


Figure 5: A partition $\{B_i\}_i$ of Ω covers everything and has no intersections, so it splits any event A into the pieces $A \cap B_i$.

Often times in this class, we'll have a random variable X which has possible outcomes, S . We'll apply the law of total probability with the partition $\{\{X = s\} : s \in S\}$ (why is this a partition?). In practice, applying the law of total probability gives you equalities like

$$\begin{aligned} \mathbb{P}(Y = y) &= \sum_{s \in S} \mathbb{P}(Y = y|X = s)\mathbb{P}(X = s), \quad \text{or} \\ \mathbb{P}(Y = y|Z = z) &= \sum_{s \in S} \mathbb{P}(Y = y|Z = z, X = s)\mathbb{P}(X = s|Z = z) \end{aligned} \quad (145)$$

PROOF: The key fact underlying this proof is that if $\{E_i\}$ is a countable family of events then

$$\mathbb{P}(E_i \cap E_j) = 0 \text{ for all } i, j \implies \mathbb{P}\left(\bigcup_i E_i\right) = \sum_i \mathbb{P}(E_i) \quad (146)$$

This fact is one of the axioms of probability.

Remark 11

If E_1, E_2 are two events then this statement simplifies to

$$\mathbb{P}(E_1 \cap E_2) = 0 \implies \mathbb{P}(E_1 \cup E_2) = \mathbb{P}(E_1) + \mathbb{P}(E_2). \quad (147)$$

which can be seen as a consequence of the inclusion-exclusion principle.

Since $\bigcup_i B_i = \Omega$, we have that

$$\mathbb{P}(A) = \mathbb{P}\left(A \cap \bigcup_i B_i\right) = \mathbb{P}\left(\bigcup_i (A \cap B_i)\right). \quad (148)$$

But we can define $E_i := A \cap B_i$ and we have that

$$\begin{aligned}\mathbb{P}(E_i \cap E_j) &= \mathbb{P}((A \cap B_i) \cap (A \cap B_j)) \\ &= \mathbb{P}(A \cap (B_i \cap B_j)) \\ &\leq \mathbb{P}(B_i \cap B_j) = 0.\end{aligned}\tag{149}$$

Therefore,

$$\mathbb{P}(A) = \mathbb{P}\left(\bigcup_i (A \cap B_i)\right) = \sum_i \mathbb{P}(A \cap B_i) = \sum_i \mathbb{P}(A|B_i)\mathbb{P}(B_i).\tag{150}$$

For the second statement, we can use roughly the same steps:

$$\begin{aligned}\mathbb{P}(A|C) &= \mathbb{P}\left(\bigcup_i (A \cap B_i) \mid C\right) \\ &= \sum_i \mathbb{P}(A \cap B_i|C) \\ &= \sum_i \mathbb{P}(A|B_i \cap C)\mathbb{P}(B_i|C).\end{aligned}\tag{151}$$

■

References

- [1] L. Yang *et al.*, “Diffusion Models: A Comprehensive Survey of Methods and Applications,” *ACM Comput. Surv.*, vol. 56, no. 4, pp. 1–39, Apr. 2024, doi: [10.1145/3626235](https://doi.org/10.1145/3626235).
- [2] J. Dongarra and F. Sullivan, “ Guest Editors Introduction to the top 10 algorithms ,” *Computing in Science & Engineering*, vol. 2, no. 01, pp. 22–23, Jan. 2000, doi: [10.1109/MCISE.2000.814652](https://doi.org/10.1109/MCISE.2000.814652).
- [3] R. Durrett, *Essentials of Stochastic Processes*. in Springer Texts in Statistics. Cham: Springer International Publishing, 2016. doi: [10.1007/978-3-319-45614-0](https://doi.org/10.1007/978-3-319-45614-0).