AI2613: Stochastic Processes

Chihao Zhang

June 2, 2022

Contents

1 Probability Space

We start with the notion of probability space. The standard reference for the probability theory is [Dur19].

Definition 1.1 (Probability Space) A *probability space* is a tuple $(\Omega, \mathcal{F}, \mathsf{P}(\cdot))$ satisfying the following requirements.

- The universe Ω is a set of "outcomes" (which can be either countable or uncountable).
- The set $\mathcal{F} \subseteq 2^\Omega$ is a σ-algebra (the set of all possible "events"). Here we say $\mathcal{F}$ is a σ-algebra if $\mathcal{F}$ satisfies:
  - $\varnothing, \Omega \in \mathcal{F}$;
  - $\forall A \in \mathcal{F}$, it holds $A^c \in \mathcal{F}$;[a]
  - for any finite or countable sequence of sets $A_1, \ldots, A_n, \cdots \in \mathcal{F}$, it holds that $\bigcup_{i=1}^\infty A_i \in \mathcal{F}$.
- The probability function $\mathsf{P}(\cdot) : \mathcal{F} \to [0,1]$ satisfies
  - $\mathsf{P}(\varnothing) = 0$, $\mathsf{P}(\Omega) = 1$;
  - $\mathsf{P}(A^c) = 1 - \mathsf{P}(A)$ for all $A \in \mathcal{F}$;
  - for any finite or countable sequence of *disjoint* sets $A_1, \ldots, A_n, \cdots \in \mathcal{F}$, it holds that $\mathsf{P}\left(\bigcup_{i=1}^\infty A_i\right) = \sum_{i=1}^\infty \mathsf{P}(A_i)$.

[a] $A^c \triangleq \Omega \setminus A$.

Let $\mathcal{S} \subseteq 2^\Omega$. We use $\sigma(\mathcal{S})$ to denote the minimal σ-algebra containing sets in $\mathcal{S}$. [1]That is, for any $\mathcal{F} \subseteq 2^\Omega$, $\mathcal{F} = \sigma(\mathcal{S})$ if and only if (1) $\mathcal{F}$ is a σ-algebra; (2) $\mathcal{S} \subseteq \mathcal{F}$; (3) For any $\mathcal{F}' \subseteq \mathcal{F}$ such that $\mathcal{S} \subseteq \mathcal{F}'$, $\mathcal{F}'$ is not a σ-algebra.
[2]

Example 1 (Tossing n fair coins) Let $\Omega = \{0,1\}^n, \mathcal{F} = 2^\Omega$ and for every $S \in \{0,1\}^n$, $\mathsf{P}(\{S\}) = \frac{1}{2^n}$.

[3]

Example 2 (Uniform Reals in $(0,1)$) The uniform distribution on $(0,1)$ is defined as follows:

- $\Omega = (0,1)$;
- $\mathcal{F}$ is the σ-algebra consisiting of all Borel sets on $(0,1)$, namely the collection of subsets of $(0,1)$ obtained from open intervals by repeatedly taking *countable* unions and complements;
- $\forall$ interval $I = (a,b)$, $P(I) = b - a$ (This is the Lebesgue measure).

2 Random Variables

Definition 2.1 (Measurable Space) Consider a set Ω and a σ-algebra $\mathcal{F}$ on Ω. The tuple $(\Omega, \mathcal{F})$ is called a measurable space.

Definition 2.2 (Measurable Function) Let $(\Omega, \mathcal{F})$ and $(\Omega', \mathcal{F}')$ be two measurable spaces and $X : \Omega \to \Omega'$ be a function. We say X is a *$\mathcal{F}$-measurable function* if

$$\forall B' \in \mathcal{F}',\ X^{-1}(B') \in \mathcal{F},$$

[a]

[a] $X^{-1}(B') \triangleq \{\omega \in \Omega | X(\omega) \in B'\}$ is the inverse of X.

For any function, we use $\sigma(X)$ to denote the minimal σ-algebra $\mathcal{F}$ such that X is $\mathcal{F}$-measurable.

Definition 2.3 (Random Variable) . Let Ω' and $\mathcal{F}'$ in Definition 2.2 be $\mathbb{R}$ and the Borel algebra $\mathcal{B}$, then X in Definition 2.2 is a (real-valued) random variable.

We say a random variable X *discrete* if its range $\mathsf{Ran}(X)$ is countable. In other words, X can only take at most countable many distinct values. Otherwise, we say X is a *continuous* random variable.

[1] The term "minimal" here is with respect to the set inclusion relation $\subseteq$.

[2] For every $n \in \mathbb{N}$, we use $[n]$ to denote the set $\{1, 2, \ldots, n\}$.

[3] The definition here, although a bit wired at the first glance, is in fact the simplest way to capture our intuition that the probability that a point is in (a,b) should be $b-a$. We cannot take $\mathcal{F} = 2^\Omega$ in Example 2 as doing so may include some *non-measurable* sets. In fact, $\mathcal{F}$ is called the *Borel algebra*, which is the smallest σ-algebra containing all open intervals. One can construct a non-Borel set in $(0,1)$ assuming the *axiom of choice*. In fact, the existence of a non-Borel set is independent of Zermelo-Fraenkel set theory without the axiom of choice. We use $\mathcal{B}$ to denote the collection of Borel sets on $\mathbb{R}$. For any $A \subseteq \mathbb{R}$, we use $\mathcal{B}(A)$ to denote $\mathcal{B} \cap 2^A$.

[4]

Example 3 (Measurable Functions of Tossing a Dice) . Let $\Omega = [6]$. We have three σ-algebras on Ω: $\mathcal{F}_1 = 2^{[6]}$, $\mathcal{F}_2 = \sigma(\{1,3,5\})$ and $\mathcal{F}_3 = \sigma(\{1,2\})$. Consider three random variables $X_1, X_2, X_3 : \Omega \to \mathbb{R}$ such that $X_1 : \omega \mapsto \omega$, $X_2 : \omega \mapsto \omega \bmod 2$ and $X_3 : \omega \mapsto \mathbf{1}[\omega \le 2]$. Then all these three mappings are $\mathcal{F}_1$-measurable, only X_2 is $\mathcal{F}_2$-measurable and only X_3 is $\mathcal{F}_3$-measurable.

3 Distribution

Let $(\Omega, \mathcal{F}, \mathsf{P})$ be a probability space and $X : \Omega \to \mathbb{R}$ be a $\mathcal{F}$-measurable random variable. Let $\mathscr{B}$ be the Borel algebra on $\mathbb{R}$. The distribution space $(\mathbb{R}, \mathscr{B}, \mathbf{Pr})$ induced by X is defined as

$$\forall A \in \mathscr{B}, \mathbf{Pr}\left[A\right] = \mathbf{Pr}\left[X \in A\right] \triangleq \mathsf{P}[X^{-1}(A)].$$

If a function $f : \mathbb{R} \to \mathbb{R}$ satisfies for any $a \le b$:

$$\mathbf{Pr}\left[a \le X \le b\right] = \mathbf{Pr}\left[X^{-1}([a,b])\right] = \int_a^b f(x)\,\mathrm{d}x,$$

Then we call $f(x)$ the *probability density function (pdf)* of X.

The function $F(x) \triangleq \mathbf{Pr}\left[X \le x\right] = \int_{-\infty}^{x} f(t)\,\mathrm{d}t$ is called the *cumulative distribution function (cdf)* of X.

Example 4 (Exponential Distribution) If $X \sim \mathtt{Exp}(\lambda)$, or equivalently it follows exponential distribution with rate λ for $\lambda > 0$, then its pdf is

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0, & \text{otherwise.} \end{cases}$$

4 Expectation and Variance

Definition 4.1 (Expectation) . Let $(\Omega, \mathcal{F}, \mathsf{P})$ be a probability space and $X : \Omega \to \mathbb{R}$ be a random variable.

- For a discrete random variable X, its expectation is

  $$\mathbf{E}\left[X\right] \triangleq \sum_{a \in \mathsf{Ran}(X)} a \cdot \mathbf{Pr}\left[X = a\right].$$

  If Ω is at most countable, we can also write

  $$\mathbf{E}\left[X\right] = \sum_{\omega \in \Omega} \mathsf{P}(\{\omega\}) \cdot X(\omega).$$

- For a continuous random variable X with pdf f, its expectation is

  $$\mathbf{E}\left[X\right] \triangleq \int_{-\infty}^{\infty} t \cdot f(t)\,dt.$$

  Sometimes it is more convenient to equivalently write the expectation as

  $$\mathbf{E}\left[X\right] = \int_{\Omega} X(\omega)\mu(d\omega) = \int_{\Omega} X\,d\mu.$$

  using Lebesgue integration.

Example 5 (Expectation of Exponential Distribution) Let $X \sim \mathtt{Exp}(\lambda)$ for $\lambda > 0$, then

$$\mathbf{E}\left[X\right] = \int_0^{\infty} t \cdot \lambda e^{-\lambda t}\,\mathrm{d}t = \frac{1}{\lambda}.$$

Definition 4.2 (Variance) The variance of a random variable X is

$$\mathbf{Var}\left[X\right] \triangleq \mathbf{E}\left[(X - \mathbf{E}\left[X\right])^2\right] = \mathbf{E}\left[X^2\right] - \mathbf{E}\left[X\right]^2.$$

[4] The *measurability* of a random variable X captures the intuition that we can safely talk about *the probability of X taking some value.* Intuitively X induces a partition of Ω where two outcomes ω_1 and ω_2 are in the same partition if and only if $X(\omega_1) = X(\omega_2)$. If the partition defined by X is more "coaser" than the partition defined by a σ-algebra $\mathcal{F}$, then X is $\mathcal{F}$ measurable.

Proposition 4.1 Let $X_1, \dots, X_n$ be random variables where n is a finite constant. Then

$$\mathbf{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbf{E}\left[X_i\right].$$

5 Conditional Probability

[5]

Definition 5.1 (Conditional Probability) Let $(\Omega, \mathcal{F}, \mathsf{P})$ be a probability space. Let $A, B \in \mathcal{F}$ be two events with $\mathsf{P}(B) > 0$. The conditional probability of A given B is

$$\mathsf{P}(A \mid B) \triangleq \frac{\mathsf{P}(A \cap B)}{\mathsf{P}(B)}.$$

In the following, we define the notion of *conditional expectation* for those *discrete* random variables.

Definition 5.2 (Conditional Expectation) Let $(\Omega, \mathcal{F}, \mathsf{P})$ be a probability space. Let $A \in \mathcal{F}$ be an event with $\mathsf{P}(A) > 0$. Let $X : \Omega \to \mathbb{R}$ be a *discrete* random variable. The conditional expectation of X conditioned on A is

$$\mathbf{E}\left[X \mid A\right] \triangleq \sum_{a \in \mathsf{Ran}(X)} a \cdot \mathbf{Pr}\left[X = a \mid A\right].$$

Let $Y : \Omega \to \mathbb{R}$ be another discrete random variable. The conditional expectation of X conditioned on Y, written as $\mathbf{E}\left[X \mid Y\right]$, is a random variable $f_Y : \Omega \to \mathbb{R}$ such that

$$\forall \omega \in \Omega: \; f_Y(\omega) = \mathbf{E}\left[X \,\middle|\, Y^{-1}(Y(\omega))\right] = \mathbf{E}\left[X \mid Y = Y(\omega)\right]. \tag{1}$$

Proposition 5.1

- $\mathbf{E}\left[X \mid Y\right]$ is $\sigma(Y)$-measurable.
- $\mathbf{E}\left[\mathbf{E}\left[X \mid Y\right]\right] = \mathbf{E}\left[f_Y\right] = \mathbf{E}\left[X\right]$.

Proof.

- Since the value of $\mathbf{E}\left[X \mid Y\right]$ is determined by $Y(\omega)$, it is clearly $\sigma(Y)$-measurable.
- We compute $\mathbf{E}\left[f_Y\right]$ by definition.

$$\begin{aligned}
\mathbf{E}\left[f_Y\right] &= \sum_{y \in \mathsf{Ran}(Y)} \mathbf{Pr}\left[Y = y\right] \cdot \mathbf{E}\left[X \mid Y = y\right] \\
&= \sum_{y \in \mathsf{Ran}(Y)} \mathbf{Pr}\left[Y = y\right] \cdot \sum_{x \in \mathsf{Ran}(X)} \mathbf{Pr}\left[X = x \mid Y = y\right] \cdot x \\
&= \sum_{x \in \mathsf{Ran}(X)} x \cdot \sum_{y \in \mathsf{Ran}(Y)} \mathbf{Pr}\left[Y = y\right] \cdot \mathbf{Pr}\left[X = x \mid Y = y\right] \\
&= \sum_{x \in \mathsf{Ran}(X)} x \cdot \sum_{y \in \mathsf{Ran}(Y)} \mathbf{Pr}\left[X = x \wedge Y = y\right] \\
&= \sum_{x \in \mathsf{Ran}(X)} x \cdot \mathbf{Pr}\left[X = x\right] \\
&= \mathbf{E}\left[X\right].
\end{aligned}$$

□

6 Conditional Expectation for General Random Variables

The definition of conditional expectation for continuous random variables is more subtle. For example, if $X, Y \sim N(0,1)$ are two independent random variables following standard normal distribution, then intuitively $\mathbf{E}\left[X \mid Y = 0\right]$ should be identical to $\mathbf{E}\left[X\right]$, which is zero. However, we cannot directly adopt the definition before since $\mathbf{Pr}\left[Y = 0\right] = 0$.

[5] This is well-defined since we know from the definition of σ-algebra that $A \cap B \in \mathcal{F}$.

Definition 6.1 Let $(\Omega, \mathcal{F}, \mathsf{P})$ be the probability space. Let X be a random variable with $\mathbf{E}[|X|] < \infty$. The conditional expectation $\mathbf{E}[X \mid Y]$ is a $\sigma(Y)$-measurable random variable f_Y satisfying

$$\forall A \in \sigma(Y), \int_A f_Y \, dP = \int_A X \, dP.$$

The existence and uniqueness of f_Y follow from Radon-Nikodym theorem.

In this lecture, we first introduce the balls-into-bins model, which is a common structure arising in probabilistic analysis. Then we look at the "concentration inequalities", namely a set of inequalities that provide bounds on how a random variable deviates from its expectation. Finally, we start our journey on finite Markov chains.

7 Balls-into-Bins

Balls-into-bins is a simple random process in which a person throws m balls into n bins uniformly at random. Many interesting questions can be asked about the process.

7.1 Birthday Paradox

Birthday paradox refers to the seemly counter-intuitive fact that some students in the class are very likely to share the same birthday. Viewing bins as dates and balls as students, the event that two students have the same birthday can be modeled as the event that some bin contains more than one ball.

Note that each ball is thrown independently. Condition on there is no collision after the $k-1$ balls are thrown, the probability that no collision occurs after throwing the k^{th} ball is $\frac{n-k+1}{n}$. Hence,

$$\begin{aligned}
\mathbf{Pr}[\text{no same birthday}] &= \prod_{k=1}^{m} \frac{n-k+1}{n} \\
&= \prod_{k=1}^{m-1} \left(1 - \frac{k}{n}\right) \\
&\le \exp\left\{-\frac{\sum_{k=1}^{m-1} k}{n}\right\} \quad (\text{by } 1 + x \le e^x) \\
&= \exp\left\{-\frac{m(m-1)}{2n}\right\}. \qquad (2)
\end{aligned}$$

For $m = O(\sqrt{n})$, the probability can be arbitrarily close to 0. [6]

7.2 Coupon Collector

The coupon collector problem asks the following question: If each box of a brand of cereals contains a coupon, randomly chosen from n different types of coupons, what is the number of boxes one needs to buy to collect all n coupons? In the language of balls-into-bins, it asks how many balls one needs to throw until each of the n bins contains at least one ball.

The expectation can be easily calculated using the linearity of expectations. Let X_i be the number of balls to throw to get the i-th distinct type of coupon while exactly $i-1$ distinct types of coupons are already in had. Then the number of draws X to collect all coupons satisfies

$$X = \sum_{i=1}^{n-1} X_i.$$

By the linearity of expectations:

$$\mathbf{E}[X] = \sum_{i=1}^{n} \mathbf{E}[X_i].$$

It is clear that $X_i \sim \mathtt{Gem}(\frac{n-i+1}{n})$ and therefore $\mathbf{E}[X_i] = \frac{n}{n-i+1}$. As a result,

$$\mathbf{E}[X] = \sum_{i=1}^{n} \frac{n}{n-i+1} = n \cdot H(n),$$

where $H(n)$ is the harmonic number satisfying $\lim_{n\to\infty} H_(n) = \log n + \gamma$ for $\gamma = 0.577\ldots$ [7]

8 Concentration Inequalities

In addition to the expectation, we are often interested in how a random variable deviates from certain fixed value. Concentration inequalities are inequalities of this form.

[6] When n is sufficiently large, Equation (2) is quite tight because $\frac{k}{n} \le \frac{m}{n} = O(\frac{1}{\sqrt{n}}) \to 0$ and $1 + x \le e^x$ is quite tight when x is small.

[7] γ is called the Euler constant.

8.1 Markov's Inequality

Theorem 8.1 (Markov's Inequality) . For any non-negative random variable X and $a > 0$,

$$\mathbf{Pr}\left[X \geq a\right] \leq \frac{\mathbf{E}\left[X\right]}{a}.$$

Proof. Since X is non-negative, we have

$$\mathbf{E}\left[X\right] \geq a \cdot \mathbf{Pr}\left[X \geq a\right] + 0 \cdot \mathbf{Pr}\left[X < a\right].$$

This is equivalent to

$$\mathbf{Pr}\left[X \geq a\right] \leq \frac{\mathbf{E}\left[X\right]}{a}.$$

□

Example 6 (Concentration for Coupon Collector) . Recall that X is the number of balls we need. Apply Markov's inequality, for $c > 0$ we have

$$\mathbf{Pr}\left[X \geq c\right] \leq \frac{\mathbf{E}\left[X\right]}{c} = \frac{nH_n}{c}.$$

Thus, the probability that we need to draw the coupon for more than $100 \cdot nH_n$ times is less than 0.01.

8.2 Chebyshev's Inequality

A common trick to improve concentration is to consider $\mathbf{E}\left[f(X)\right]$ instead $\mathbf{E}\left[X\right]$ for some increasing function $f : \mathbb{R} \to \mathbb{R}$ since

$$\mathbf{Pr}\left[X \geq a\right] = \mathbf{Pr}\left[f(X) \geq f(a)\right].$$

Concentration inequalities give a sense that how the random variable deviate from its expectation. Then the probability we care about is actually $\mathbf{Pr}\left[|X - \mathbf{E}\left[X\right] | \geq a\right]$ for some postive constant a. Choosing the increasing function $f(x) = x^2$, we get the following Chebyshev's inequality.

Theorem 8.2 (Chebyshev's Inequality) . For any random variable with bounded $\mathbf{E}\left[X\right]$ and $a \geq 0$, it holds that

$$\mathbf{Pr}\left[|X - \mathbf{E}\left[X\right]| \geq a\right] \leq \frac{\mathbf{Var}\left[X\right]}{a^2}$$

Proof. Let $Y = |X - \mathbf{E}\left[X\right]|$, then clearly $Y \geq 0$. Therefore

$$\begin{aligned}\mathbf{Pr}\left[|X - \mathbf{E}\left[X\right]| \geq a\right] = \mathbf{Pr}\left[Y \geq a\right] = \mathbf{Pr}\left[Y^2 \geq a^2\right] &\leq \frac{\mathbf{E}\left[Y^2\right]}{a^2} \\ &= \frac{\mathbf{E}\left[(X - \mathbf{E}\left[X\right])^2\right]}{a^2} = \frac{\mathbf{Var}\left[X\right]}{a^2}.\end{aligned}$$

□

Example 7 (Coupon Collector Revisited) We apply Chebyshev's inequality to the coupon collector problem. Assuming the notation before, we have

$$\mathbf{Pr}\left[X \geq nH_n + t\right] \leq \mathbf{Pr}\left[|X - \mathbf{E}\left[X\right]| \geq t\right] \leq \frac{\mathbf{Var}\left[X\right]}{t^2}.$$

Recall that the variable X_i indicates the number of draws to get a new coupon while there are i coupons in hands. For distinct i and j, X_i and X_j are independent. Then

$$\mathbf{Var}\left[X\right] = \mathbf{Var}\left[\sum_{i=0}^{n-1} X_i\right] = \sum_{i=0}^{n-1} \mathbf{Var}\left[X_i\right].$$

For $i \in \{0, 1, \ldots, n-1\}$, $X_i \sim \mathtt{Geom}\left(\frac{n-i}{n}\right)$, so we have

$$\mathbf{Var}\left[X_i\right] = \frac{1 - \frac{n-i}{n}}{\left(\frac{n-i}{n}\right)^2} = \frac{i \cdot n}{(n-i)^2} \leq \frac{n^2}{(n-i)^2}.$$

It remains to bound $\sum_{i=0}^{n-1} \frac{1}{(n-i)^2} = \sum_{i=1}^{n} \frac{1}{i^2}$. Note that

$$\sum_{i=1}^{n} \frac{1}{i^2} \leq 1 + \int_1^\infty \frac{\mathrm{d}x}{x^2} = 2.$$

Therefore, we have $\mathbf{Var}\left[X\right] \leq 2n^2$ and $\mathbf{Pr}\left[X \geq nH_n + t\right] \leq \frac{2n^2}{t^2}$. The probability that we need to draw the coupon for more than $\sqrt{200}n + nH_n$ times is less than 0.01. [8]

[8] The bound obtained by Chebyshev's inequality is much tighter than that via Markov's inequality where in order to obtain the same confidence, one needs to choose $t = \Theta(n \log n)$.

8.3 Vanilla Chernoff Bound

If we apply Markov inequality to

$$\mathbf{Pr}\left[f(X) \geq f(t)\right]$$

with $f(x) = e^{\alpha x}$ where $\alpha > 0$, then the bound amounts to bound $\mathbf{E}\left[e^{\alpha X}\right]$ which is the *moment generating function* of X.

When the random variable X can be written as the sum of independent Bernoulli variables, its moment generating function is easy to estimate and we obtain sharp concentration bounds.

Theorem 8.3 (Chernoff Bound) . Let $X_1, \dots, X_n$ be independent random variables such that $X_i \sim Ber(p_i)$ for each $i = 1, 2, \dots, n$. Let $X = \sum_{i=1}^{n} X_i$ and denote $\mu \triangleq \mathbf{E}\left[X\right] = \sum_{i=1}^{n} p_i$, we have

$$\mathbf{Pr}\left[X \geq (1+\delta)\mu\right] \leq \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu}$$

If $0 < \delta < 1$, then we have

$$\mathbf{Pr}\left[X \leq (1-\delta)\mu\right] \leq \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mu}$$

Proof. We only prove the upper tail bound and the proof of lower tail bound is similar. For every $\alpha > 0$, we have

$$\mathbf{Pr}\left[X \geq (1+\delta)\mu\right] = \mathbf{Pr}\left[e^{\alpha X} \geq e^{\alpha(1+\delta)\mu}\right] \leq \frac{\mathbf{E}\left[e^{\alpha X}\right]}{e^{\alpha(1+\delta)\mu}}.$$

Therefore, we need to estimate the moment generating function $\mathbf{E}\left[e^{\alpha X}\right]$. Since $X = \sum_{i=1}^{n} X_i$ is the sum of independent Bernoulli variables, we have

$$\mathbf{E}\left[e^{\alpha X}\right] = \mathbf{E}\left[e^{\alpha \sum_{i=1}^{n} X_i}\right] = \mathbf{E}\left[\prod_{i=1}^{n} e^{\alpha X_i}\right] = \prod_{i=1}^{n} \mathbf{E}\left[e^{\alpha X_i}\right].$$

Since $X_i \sim \mathrm{Ber}(p_i)$, we can compute $\mathbf{E}\left[e^{\alpha X_i}\right]$ directly:

$$\mathbf{E}\left[e^{\alpha X_i}\right] = p_i e^{\alpha} + (1-p_i) = 1 + (e^{\alpha} - 1)p_i \leq e^{((e^{\alpha}-1)p_i)}.$$

Therefore,

$$\mathbf{E}\left[e^{\alpha X}\right] \leq \prod_{i=1}^{n} e^{((e^{\alpha}-1)p_i)} = e^{((e^{\alpha}-1)\sum_{i=1}^{n} p_i)} = e^{((e^{\alpha}-1)\mu)}.$$

Therefore,

$$\mathbf{Pr}\left[X \leq (1+\delta)\mu\right] \leq \frac{\mathbf{E}\left[e^{\alpha x}\right]}{e^{\alpha(1+\delta)\mu}} \leq \left(\frac{e^{(e^{\alpha}-1)}}{e^{(\alpha(1+\delta))}}\right)^{\mu}$$

Note that above holds for any $\alpha > 0$. Therefore, we can choose α so as to minimize $\frac{e^{(e^{\alpha}-1)}}{e^{(\alpha(1+\delta))}}$. To this end, we let $\left(\frac{e^{(e^{\alpha}-1)}}{e^{(\alpha(1+\delta))}}\right)' = 0$. This gives $\alpha = \log(1+\delta)$. Therefore

$$\mathbf{Pr}\left[X \leq (1+\delta)\mu\right] \leq \left(\frac{e^{(e^{\alpha}-1)}}{e^{(\alpha(1+\delta))}}\right)^{\mu} = \left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{\mu}.$$

□

The following form of Chernoff bound is more convenient to use (but weaker):

Corollary 8.1 For any $0 < \delta < 1$,

$$\mathbf{Pr}\left[X \geq (1+\delta)\mu\right] \leq \exp\left\{\left(-\frac{\delta^2}{3}\mu\right)\right\}$$

$$\mathbf{Pr}\left[X \leq (1-\delta)\mu\right] \leq \exp\left\{\left(-\frac{\delta^2}{2}\mu\right)\right\}$$

Proof. We only prove the upper tail. It suffices to verify that for $0 < \delta < 1$, we have

$$\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}} \leq \exp\left\{\left(-\frac{\delta^2}{3}\right)\right\}$$

Taking logarithm of both sides, this is equivalent to

$$\delta - (1+\delta)\ln(1+\delta) \leq -\frac{\delta^2}{3}$$

Let $f(\delta) = \delta - (1+\delta)\ln(1+\delta) + \frac{\delta^2}{3}$ and note that

$$f'(\delta) = -\ln(1+\delta) + \frac{2}{3}\delta, \quad f''(\delta) = -\frac{1}{1+\delta} + \frac{2}{3}.$$

Then for $0 < \delta < 1/2$, $f''(\delta) < 0$, and for $1/2 < \delta < 1$, $f''(\delta) > 0$. Therefore, $f'(\delta)$ first decreases and then increases in $[0,1]$. Also note that $f'(0) = 0$, $f'(1) < 0$ and $f'(\delta) \le 0$ when $0 \le \delta \le 1$. Therefore $f(\delta) \le f(0) = 0$. □

Example 8 (Tossing p-coins) . Consider a p-coin that we get a head with probability p when tossing it. If we toss a p-coin n times, the average number of heads is pn.We want to determine the value δ such that with high probability (say 99%), the total number of heads is in the interval of $[(1-\delta)pn, (1+\delta)pn]$. We use Chernoff bound to determine δ.

Let X denote the total number of heads, and $X_i \sim \mathrm{Ber}(p)$ be the indicator of whether the i-th toss gives a head. Then by Chernoff bound, we have

$$\mathbf{Pr}\left[|X - pn| \ge \delta \cdot pn\right] \le 2\exp\left\{\left(-\frac{\delta^2}{3} \cdot pn\right)\right\} \le 0.01$$

So if p is a constant, it suffices to choose

$$\delta = \Omega\left(\frac{1}{\sqrt{n}}\right).$$

9 Discrete Markov Chain

9.1 Markov Chain

Definition 9.1 (Discrete Markov Chain) Suppose there is a sequence of random variables

$$X_0, X_1, \ldots, X_t, X_{t+1}, \ldots$$

where the $Ran(X_t) \subseteq \Omega$ for some countable Ω. Then we call $\{X_t\}$ a discrete Marcov chain if $\forall t \ge 1$ the distribution of X_t is only related to X_{t-1}, that is $\forall a_0, a_1, \ldots, a_t \in \Omega$,

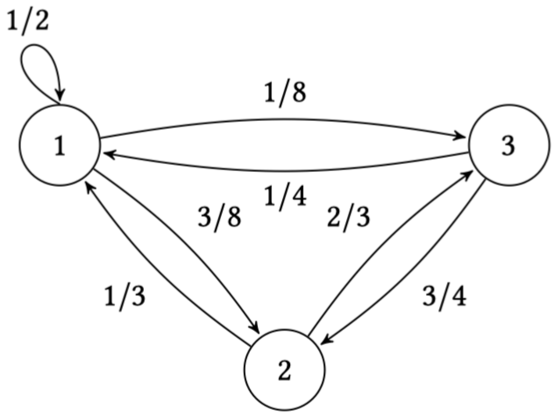
$$\mathbf{Pr}\left[X_t = a_t | X_{t-1} = a_{t-1}, \ldots, X_1 = a_1, X_0 = a_0\right] = \mathbf{Pr}\left[X_t = a_t | X_{t-1} = a_{t-1}\right].$$

Example 9 (Random Walk on $\mathbb{Z}$) . Consider the random walk on $\mathbb{Z}$. One starts at 0 and in each round, he tosses a fair coin to determine the direction of moving: with probability 50% to the left and 50% to the right. If we use X_t to denote his position at time t, then we have $X_0 = 0$ and for every $t > 0$, $X_t = X_{t-1} + 1$ with probability 50% and $X_t = X_{t-1} - 1$ with probability 50%. This is a simple Markov chain, since the position at time t only depends on the position at time $t-1$.

In this lecture, we consider the situation that the state space $\Omega = [n]$ is finite. Then a (time-homogeneous) Markov chain can be characterized by a $n \times n$ matrix $P = (p_{ij})_{i,j\in[n]}$ where $p_{ij} = \mathbf{Pr}\left[X_{t+1} = j \mid X_t = i\right]$ for all $t \ge 0$.

In general, a Markov chain can be equivalently viewed as a random walk on a weighted directed graph where the edge weight from i to j means the probability of moving to vertex j when one is standing at vertex i.

Example 10 (Finite State Random Walk) The following three vertex directed graph corresponds to the Markov chain with transition matrix $P = (p_{ij}) = \begin{bmatrix} 1/2 & 3/8 & 1/8 \\ 1/3 & 0 & 2/3 \\ 1/4 & 3/4 & 0 \end{bmatrix}$. We sometimes call the graph the *transition graph* of P.



At any time $t \geq 0$, we use μ_t to denote the distribution of X_t meaning

$$\mu_t(i) \triangleq \mathbf{Pr}[X_t = i].$$

By the law of total probability, $\mu_{t+1}(j) = \sum_i \mu_t(i) \cdot p_{ij}$, we have $\mu_t^\mathsf{T} P = \mu_{t+1}^\mathsf{T}$. As a result, we have $\mu_t^\mathsf{T} = \mu_0^\mathsf{T} P^t$. This is a useful formula as we can compute the distribution at any time given the initial distribution and the transition matrix.

Sometimes, we will simply denote the transition matrix P as the Markov chain for convenience.

9.2 Stationary Distribution

Definition 9.2 (Stationary Distribution) . A distribution π is a stationary distribution of P if it remains unchanged in the Markov chain as time progresses, i.e.,

$$\pi^\mathsf{T} P = \pi^\mathsf{T}.$$

One of the major algorithmic applications of Markov chains is the *Markov chain Monte Carlo (MCMC)* method. It is a general method for designing an algorithm to sample from a certain distribution π. The idea of MCMC is

- First design a Markov Chain of which the stationary distribution is the desired π;
- Simulate the chain from a certain initial distribution for a number of steps and output the state.

Therefore, we hope that the distribution μ_t is close to π when t is large enough.

Example 11 (Card Shuffling) Consider a naive "top-to-random" card shuffle: Suppose we have n cards, every time we take the top card of the deck and insert it into the deck at one of the n distinct possible places uniformly at random. Thus, there are $n!$ possible permutations and $p_{ij} > 0$ only if the i^{th} permutation can come to the j^{th} through one step "top-to-random" shuffle.

Performing the shuffle repeatedly is a Markov chain. It is not difficult to verify that the uniform distribution $\left(\frac{1}{n!}, \frac{1}{n!}, \ldots, \frac{1}{n!}\right)^T$ over all $n!$ permutations is a stationary distribution.

One of the main purposes of the course is to understand the MCMC method. Therefore, the following four basic questions regarding stationary distributions are important.

- Does each Markov chain have a stationary distribution?
- If a Markov chain has a stationary distribution, is it unique?
- If the chain has a unique stationary distribution, does μ_t always converge to it from any μ_0?
- If μ_t always converges to the stationary distribution, what is the rate of convergence?

10 Fundamental Theorem of Markov Chains

10.1 The Existence of Stationary Distribution

We will show that, for every finite Markov chain P, there exists some π such that $\pi^\mathsf{T} P = \pi^\mathsf{T}$. Observe that this is equivalent to "1 is an eigenvalue of P^T with a nonnegative eigenvector ($P^\mathsf{T}\pi = \pi$)".

We use the following lemma and theorem in linear algebra.

Lemma 10.1 Every eigenvalue of nonnegative matrix P is no larger than the maximum row sum of P.

Proof. Let λ be a eigenvalue of P and x is the corresponding eigenvector. We have

$$\|\lambda x\|_\infty = \|Px\|_\infty \leq \|P\|_\infty \cdot \|x\|_\infty.$$

Note that $\|\lambda x\|_\infty = |\lambda| \|x\|_\infty$ and $\|x\|_\infty > 0$. Thus, we have $\lambda \leq |\lambda| \leq \|P\|_\infty$, that is λ is no larger than the maximum row sum of nonnegative matrix P. □

Theorem 10.1 (Perron-Frobenius Theorem) Each nonnegative matrix A has a nonnegative real eigenvalue with spectral radius $\rho(A) = a$, and a has a corresponding nonnegative eigenvector.

[9]

We will prove the Perron-Frobenius theorem in Section 10.3.

Since P is a stochastic matrix, we have

$$P \cdot \mathbf{1} = \mathbf{1}.$$

[9] Let $A = (a_{ij})_{i\in[n], j\in[m]}$. We say A is nonnegative (resp. positive) if every $a_{ij} \geq 0$ (resp. > 0).

Thus, P has an eigenvalue 1. Since every eigenvalue of P is no larger than the row sum, 1 is the largest eigenvalue. Also, $P^\top$ shares the same characteristic polynomial with P, which implies the eigenvalues of $P^\top$ and P are the same. As a result, $\rho(P^\top)$ also equals to 1. According to Perron-Frobenius theorem, there exists a nonnegative eigenvector π such that

$$P^\top \pi = \pi,$$

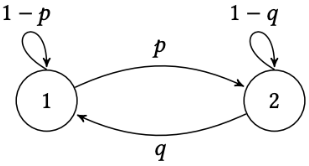
which is equivalent to

$$\pi^\top P = \pi^\top.$$

It then follows that $\frac{\pi}{\|\pi\|_1}$ is a stationary distribution of P.

10.2 Uniqueness and Convergence

Consider the following Markov chain with two states.



Clearly, the transition matrix of this Markov chain is

$$P = \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix}$$

It is easy to verify that

$$\pi = \left(\frac{q}{p+q}, \frac{p}{p+q} \right)^\top$$

is a stationary distribution of P.

We are going to check whether starting from any μ_0, the distribution μ_t will always converge to π, i.e.,

$$\lim_{t\to\infty} \left\| \mu_0^\top P^t - \pi^\top \right\| = 0.$$

In our example, the distribution has only two dimensions and the sum of the two components equals to 1, so we only need to check whether the first dimension converges, i.e.,

$$\left| \mu_0^\top P^t(1) - \pi(1) \right| \to 0.$$

Now we define

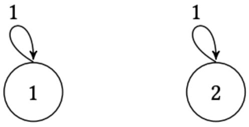
$$\begin{aligned} \Delta_t &\triangleq | \mu_t(1) - \pi(1) | \\ &= \left| \mu_{t-1}^T \cdot P(1) - \pi(1) \right| \\ &= \left| (1-p) \cdot \mu_{t-1}(1) + q \cdot (1 - \mu_{t-1}(1)) - \frac{q}{p+q} \right| \\ &= \left| (1-p-q) \cdot \mu_{t-1}(1) + q \cdot \left(1 - \frac{1}{p+q} \right) \right| \\ &= |1-p-q| \cdot \Delta_{t-1} \end{aligned}$$

Therefore, we can see that $\Delta_t \to 0$ except in the two cases:

- $p = q = 0$,
- $p = q = 1$.

In fact, the two cases prevent convergence for different reasons.

Let us first consider the case when $p = q = 0$. The Markov chain looks like:



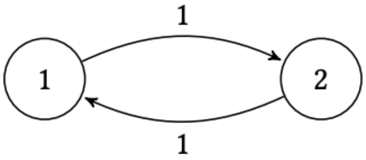
The transition graph is disconnected, so it can be partitioned into two disjoint components. Since each component is still a Markov chain, each of them has its own stationary distribution. Notice that any convex combination of these small distributions is a stationary distribution for the whole Markov chain. It immediately follows that in this case the stationary distribution is not unique. It gives a negative answer to the second question.

This observation motivates us to define the following:

Definition 10.1 (Irreducibility). A finite Markov chain is *irreducible* if its transition graph is strongly connected.

If the transition graph of P is not strongly connected, we say P is *reducible*.

When $p = q = 1$, the Markov chain looks like this:



This transition graph is bipartite. It is easy to see that $(\frac{1}{2}, \frac{1}{2})$ is the unique stationary distribution of it. However, for $\mu_0 = (1, 0)$, one can see that μ_t ocsillates between "left" and "right". Therefore, the answer to the third question is no.

This phenomenon is captured by the following notion:

Definition 10.2 (Aperiodicity). A Markov chain is *aperiodic* if for any state v, it holds that

$$\gcd \{|c| \mid c \in C_v\} = 1,$$

where C_v denotes the set of the directed cycles containing v in the transition graph.

Otherwise, we say the chain *periodic*.

We have the following important theorem.

Theorem 10.2 (Fundamental theorem of Markov chains). If a finite Markov chain $P \in \mathbb{R}^{n\times n}$ is irreducible and aperiodic, then it has a unique stationary distribution $\pi \in \mathbb{R}^n$. Moreover, for any distribution $\mu \in \mathbb{R}^n$,

$$\lim_{t\to\infty} \mu^\top P^t = \pi^\top.$$

10.3 Proof of Perron-Frobenius Theorem

Most proofs in the section are from [Mey00]. We first prove the Perron-Frobenius theorem for positive matrices. Then we use this theorem and Lemma 10.2 to prove Theorem 10.1.

In the following statement, we use $|\cdot|$ to denote a matrix or vector of absolute values, i.e., $|A|$ is the matrix with entries $|a_{ij}|$. We say a vector or matrix is larger than $\mathbf{0}$ if all its entries are larger than 0 and denote it by $A > \mathbf{0}$. We define the operation $\geq$, $\leq$ and $<$ for vectors and matrices similarly.

Theorem 10.3 (Perron-Frobenius Theorem for Positive Matrices) Each positive matrix $A > 0$ has a positive real eigenvalue $\rho(A)$, and $\rho(A)$ has a corresponding positive eigenvector.

Proof. We first prove that $\rho(A) > 0$. If $\rho(A) = 0$, then all the eigenvalues of A is 0 which is equivalent to that A is nilpotent. This is impossible since every $a_{ij} > 0$. Thus $\rho(A) > 0$ for positive matrix A.

Assume that λ is the eigenvalue of A that $|\lambda| = \rho(A)$. Then we have

$$|\lambda||x| = |\lambda x| = |Ax| \leq |A||x| = A|x|.$$

Then we show that $|\lambda||x| < A|x|$ is impossible. Let $z = A|x|$ and $y = z - \rho(A)|x|$. Assume that $y \neq \mathbf{0}$, We have that $Ay > \mathbf{0}$. There must exist some $\epsilon > 0$ such that $Ay > \epsilon\rho(A)\cdot z$ or equivalently, $\frac{A}{(1+\epsilon)\rho(A)}z > z$. Successively multiply both sides of $\frac{A}{(1+\epsilon)\rho(A)}z > z$ by $\frac{A}{(1+\epsilon)\rho(A)}$ and we have

$$\left(\frac{A}{(1+\epsilon)\rho(A)}\right)^k z > \cdots > \frac{A}{(1+\epsilon)\rho(A)} z > z, \quad \text{for } k = 1, 2, \ldots.$$

Note that $\lim_{k\to\infty}\left(\frac{A}{(1+\epsilon)\rho(A)}\right)^k \to \mathbf{0}$ because $\rho\left(\frac{A}{(1+\epsilon)\rho(A)}\right) = \frac{\rho(A)}{(1+\epsilon)\rho(A)} < 1$. Then, in the limit, $z < \mathbf{0}$. This conflicts the fact that $z > \mathbf{0}$. The assumption that $y \neq \mathbf{0}$ is invalid

Thus we have $y = \mathbf{0}$ which means $\rho(A)$ is a positive eigenvalue of A and $|x|$ is the corresponding eigenvector. Since $\rho(A)|x| = A|x| > 0$, we have $|x| > 0$. □

Lemma 10.2 For $A, B \in \mathbb{C}^{n\times n}$, if $|A| \leq B$, then $\rho(A) \leq \rho(B)$.

Proof. By spectral radius formula, we have that for any sub-multiplicative norm $\|\cdot\|$, $\rho(A) = \lim_{k\to\infty}\left\|A^k\right\|^{\frac{1}{k}}$ and $\rho(B) = \lim_{k\to\infty}\left\|B^k\right\|^{\frac{1}{k}}$.

Note that since $|A| \le B$, we have $|A|^k \le B^k$ for $k \in \mathbb{N}\backslash\{0\}$. Then $\left\|A^k\right\|_\infty \le \left\| |A|^k \right\|_\infty \le \left\|B^k\right\|_\infty$ and sequentially $\left\|A^k\right\|_\infty^{\frac{1}{k}} \le \left\|B^k\right\|_\infty^{\frac{1}{k}}$. Thus, $\rho(A) \le \rho(B)$. □

Theorem 10.4 (Theorem 10.1 restated). Each nonnegative matrix A has a nonnegative real eigenvalue with spectral radius $\rho(A) = a$, and a has a corresponding nonnegative eigenvector.

Proof. Construct a matrix sequence $\{A_k\}_{k=1}^\infty$ by letting $A_k = A + \frac{\mathbf{E}}{k}$ where $\mathbf{E}$ is the matrix of all 1's. Let $a_k = \rho(A_k) > 0$ and $x_k > \mathbf{0}$ is the corresponding eigenvector.[10] Without loss of generality, let $\|x_k\|_1 = 1$. Since $\{x_k\}_{k=1}^\infty$ is bounded, by BolzanoWeierstrass theorem, there exists a subsequence of $\{x_k\}_{k=1}^\infty$ in $\mathbb{R}^n$ that is convergent. Denote this convergent subsequence by $\left\{x_{k_i}\right\}_{i=1}^\infty$ and $\left\{x_{k_i}\right\}_{i=1}^\infty \to z$ where $z \ge 0$ and $z \ne 0$ (for each x_{k_i} satisfies that $\left\|x_{k_i}\right\|_1 = 1$). Since $\{A_k\}_{k=1}^\infty$ is monotone decreasing, by Lemma 10.2, we have that $a_1 \ge \cdots \ge a_k \ge a$. Sequence $\{a_k\}_{k=1}^\infty$ is nonincreasing and bounded, so $\lim_{k\to\infty} a_k \to a^*$ exists and $\lim_{i\to\infty} a_{k_i} \to a^* \ge a$. Then we have

$$Az = \lim_{i\to\infty} A_{k_i} x_{k_i} = \lim_{i\to\infty} a_{k_i} x_{k_i} = a^* z.$$

Thus, a^* is an eigenvalue of A and $a^* \le a$. Then we have $a^* = a$. So A has a nonnegative real eigenvalue a and z is the corresponding nonnegative eigenvetor. □

11 Fundamental Theorem of Markov Chains

Recall the fundamental theorem of Markov chains for *finite* chains we introduced in the last lecture.

Theorem 11.1 (Fundamental theorem of Markov chains) If a finite Markov chain $P \in \mathbb{R}^{n\times n}$ is irreducible and aperiodic, then it has a unique stationary distribution $\pi \in \mathbb{R}^n$. Moreover, for any distribution $\mu \in \mathbb{R}^n$,

$$\lim_{t\to\infty} \mu^\top P^t = \pi^\top.$$

Today we give a proof of the theorem. To this end, we first study the properties of the transition matrix P of an irreducible and aperiodic chain. Then we introduce the notion of *coupling*, a powerful technique to analyze stochastic processes.

Claim 11.1 Let $P \in \mathbb{R}^{n\times n}$ be an irreducible and aperiodic Markov chain. It holds that

$$\exists\, t^* : \forall\, i,j \in [n] : \quad P^{t^*}(i,j) > 0\,.$$

We use Lemma 11.1 to prove Claim 11.1.

Lemma 11.1 Let $c_1, c_2, \ldots, c_s$ be a group of positive integers satisfying $\gcd(c_1, \ldots, c_s) = 1$. For any sufficiently large integer b, there exists $y_1, y_2, \ldots, y_s \in \mathbb{N}$ such that[a]

$$c_1 y_1 + c_2 y_2 + \cdots c_s y_s = b.$$

[a] That is, there exists some $b_0 > 0$ such that for any $b > b_0$, the diophantine equation $c_1 y_1 + c_2 y_2 + \cdots + c_s y_s = b$ always has non-negative solutions

Proof. By Bézout's identity there exists $x_1, x_2, \ldots, x_s \in \mathbb{Z}$ such that

$$c_1 x_1 + c_2 x_2 + \cdots c_s x_s = 1\,.$$

We apply induction on s. The case $s = 1$ trivially holds. Assume $s \ge 2$ and the lemma holds for smaller s. Let $g = \gcd(c_1, \ldots, c_{s-1})$. By induction hypothesis, we know that

$$\frac{a_1}{g} \cdot x_1 + \frac{a_2}{g} \cdot x_2 + \cdots + \frac{a_{s-1}}{g} \cdot x_{s-1} = b' \iff a_1 \cdot x_1 + a_2 \cdot x_2 + \cdots + a_{s-1} x_{s-1} = g \cdot b'$$

has non-negative solutions for sufficiently large b'. Therefore, we only need to prove that the equation

$$g \cdot b' + a_s \cdot x_s = b \tag{3}$$

has nonegative solution (b', x_s) with sufficiently large b' when b is sufficiently large. In other words, we need to prove for any $b_0 > 0$, eq. (3) has nonegative solution with $b' > b_0$ for any sufficiently large b.

Note that $\gcd(g, a_s) = 1$, we can find integers (y, x) such that

$$g \cdot y + a_s \cdot x = 1 \iff g \cdot (by) + a_s \cdot (bx) = b.$$

[10] The existance of such x_k is guaranteed by Theorem 10.3.

Noting that for any $k \in \mathbb{Z}_{\geq 0}$, we have $g \cdot (by + ka_s) + a_s \cdot (bx - kg) = b$. We need $by + ka_s > b_0$ and $bx - kg \geq 0$, which are equivalent to

$$\frac{bx}{g} \geq k > \frac{b_0 - by}{a_s}.$$

We can always find such an integer k if $b \geq g(b_0 + a_s)$.

□

Proof. [Proof of Claim 11.1]

The property of irreducibility implies that

$$\forall\, i, j : \exists\, t : \quad P^t(i, j) > 0\,.$$

Suppose that there are s loops of length $c_1, c_2, \ldots, c_s$ starting from and ending at state i. Then by aperiodicity we have

$$\gcd(c_1, c_2, \ldots, c_s) = 1\,.$$

For any sufficiently large m and any pair of states (i, j), by Lemma 11.1 and irreducibility, there exists a path from i to j with exactly m steps. Thus, there exist $t^* > 0$ such that for any state pair (i, j), $P^{t^*}(i, j) > 0$. Furthermore, for any $t > t^*$, $P^t(i, j) > 0$ for any $i, j \in \Omega$.

□

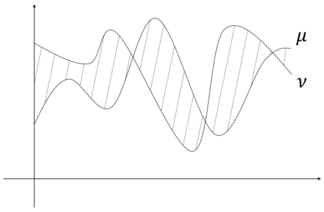
11.1 Coupling

To measure how close the two distributions are, we need to define the distance between them.

Definition 11.1 (Total Variation Distance) . The total variation distance between two distributions μ and ν on a countable state space Ω is given by

$$D_{TV}(\mu, \nu) = \frac{1}{2}\sum_{x\in\Omega} |\mu(x) - \nu(x)|.$$

We can look at the following figure of two distributions on the sample space. The total variation distance is half the area enclosed by the two curves.



This figure gives us the intuition of the following proposition which states that the total variation distance can be equivalently viewed in another way.

Proposition 11.1 We define $\mu(A) = \sum\limits_{x\in A} \mu(x)$, $\nu(A) = \sum_{x\in A} \nu(x)$, then we have

$$D_{TV}(\mu, \nu) = \max_{A\subseteq\Omega} |\mu(A) - \nu(A)|.$$

The coupling of two distributions is simply a joint distribution of them.

Definition 11.2 (Coupling) . Let μ and ν be two distributions on the same space Ω. Let ω be a distribution on the space $\Omega \times \Omega$. If $(X, Y) \sim \omega$ satisfies $X \sim \mu$ and $Y \sim \nu$, then ω is called a coupling of μ and ν.

[11]

We now give a toy example about how to construct different couplings on two fixed distributions. There are two coins: the first coin has probability $\frac{1}{2}$ for head in a toss and $\frac{1}{2}$ for tail, and the second coin has probability $\frac{1}{3}$ and $\frac{2}{3}$ respectively. We now construct two couplings as follows.

prob \ y x	HEAD	TAIL
HEAD	1/3	1/6
TAIL	0	1/2

prob \ y x	HEAD	TAIL
HEAD	1/6	1/3
TAIL	1/6	1/3

[11] In other words, the marginal probabilities of the disjoint distribution ω are μ and ν respectively. A special case is when x and y are independently. However, in many applications, we want x and y to be correlated while keeping their respect marginal probabilities correct.

The table defines a joint distribution and the sum of a certain row/column equal to the corresponding marginal probability. It is clear that both table are couplings of the two coins. Among all the possible couplings, sometimes we are interested in the one who is "mostly coupled".

Lemma 11.2 (Coupling Lemma) . Let μ and ν be two distributions on a sample space Ω. Then for any coupling ω of μ and ν it holds that,

$$\mathbf{Pr}_{(X,Y)\sim\omega}\left[X \neq Y\right] \geq D_{TV}(\mu, \nu).$$

And furthermore, there exists a coupling ω^* of μ and ν such that

$$\mathbf{Pr}_{(X,Y)\sim\omega^*}\left[X \neq Y\right] = D_{TV}(\mu, \nu).$$

Proof. For finite Ω, designing a coupling is equivalent to filling a $\Omega \times \Omega$ matrix in the way that the marginals are correct. Clearly we have

$$\begin{aligned}\mathbf{Pr}\left[X = Y\right] &= \sum_{t\in\Omega} \mathbf{Pr}\left[X = Y = t\right] \\ &\leq \sum_{t\in\Omega} \min\left\{\mu(t), \nu(t)\right\}.\end{aligned}$$

Thus,

$$\begin{aligned}\mathbf{Pr}\left[X \neq Y\right] &\geq 1 - \sum_{t\in\Omega} \min\left(\mu(t), \nu(t)\right) \\ &= \sum_{t\in\Omega} \left(\mu(t) - \min\left\{\mu(t), \nu(t)\right\}\right) \\ &= \max_{A\subseteq\Omega}\left\{\mu(A) - \nu(A)\right\} \\ &= D_{\mathrm{TV}}(\mu, \nu).\end{aligned}$$

To construct ω^* achieving the equality, for every $t \in \Omega$, we let $\mathbf{Pr}_{(X,Y)\sim\omega^*}\left[X = Y = t\right] = \min\left\{\mu(t), \nu(t)\right\}$. □ [12]

11.2 Proof of FTMC

Proof. We already know that P has a stationary distribution π. What we would like to show is that for all starting distribution μ_0, it holds that

$$\lim_{t\to\infty} D_{\mathrm{TV}}(\mu_t, \pi) = 0\,,$$

where $\mu_t^\intercal = \mu_0^\intercal P^t$.

Suppose that $\{X_t\}$ and $\{Y_t\}$ are two identical Markov chains starting from different distribution, where $Y_0 \sim \pi$ while X_0 is generated from an arbitrary distribution μ_0.

Now we have two sequence of random variables:

$$\begin{array}{ccccccccccccc}
\mu_0 & & \mu_1 & & & & & & \mu_t & & & & \\
\wr & & \wr & & & & & & \wr & & & & \\
X_0 & \longrightarrow & X_1 & \longrightarrow & X_2 & \longrightarrow & \cdots & \longrightarrow & X_t & \longrightarrow & X_{t+1} & \longrightarrow & \cdots \\
 & & & & & & & & & & & & \\
Y_0 & \longrightarrow & Y_1 & \longrightarrow & Y_2 & \longrightarrow & \cdots & \longrightarrow & Y_t & \longrightarrow & Y_{t+1} & \longrightarrow & \cdots \\
\wr & & \wr & & & & & & \wr & & & & \\
\pi & & \pi & & & & & & \pi & & & &
\end{array}$$

The coupling lemma establishes the connection between the distance of distributions and the discrepancy of random variables. To show that $D_{\mathrm{TV}}(\mu_t, \pi) \longrightarrow 0$, it is sufficient to construct a coupling ω_t of μ_t and π and then compute $\mathbf{Pr}_{(X_t,Y_t)\sim\omega_t}\left[X_t \neq Y_t\right]$.

Here we give a simple coupling. Let $(X_t, Y_t) \sim \omega_t$ and we construct ω_{t+1}. If $X_t = Y_t$ for some $t \geq 0$, then let $X_{t'} = Y_{t'}$ for all $t' > t$, otherwise X_{t+1} and Y_{t+1} are independent. Namely, $\{X_t\}$ and $\{Y_t\}$ are two independent Markov chains until X_t and Y_t reach the same state for some $t \geq 0$, and once they meet together then they move together forever. The coupling lemma tells us that $D_{\mathrm{TV}}(\mu_t, \pi) \leq \mathbf{Pr}_{(X_t,Y_t)\sim\omega_t}\left[X_t \neq Y_t\right]$.

Let t^* be the same t^* with Claim 11.1. Let α be a positive constant such that $P^{t^*}(i, j) \geq \alpha > 0$ for any state pair (i, j). Define event B as $\{\exists t < t^*, X_t = Y_t\}$. We have that

$$\mathbf{Pr}\left[X_{t^*} = Y_{t^*}\right] = \mathbf{Pr}\left[X_{t^*} = Y_{t^*} \wedge B\right] + \mathbf{Pr}\left[X_{t^*} = Y_{t^*} \wedge \bar{B}\right] \tag{4}$$

[12] The coupling lemma provides a way to upper bound the distance between two distributions: For any two distributions μ and ν and any coupling ω of μ and ν, an upper bound for $\mathbf{Pr}_{(X,Y)\sim\omega}\left[X \neq Y\right]$ is an upper bound for $D_{TV}(\mu, \nu)$. This is a quite useful approach to bound the total variation distance.

Suppose $\{X'_t\}$ and $\{Y'_t\}$ are two independent Markov chains with transition matrix P and $X'_0 \sim \mu_0$ and $Y'_0 \sim \pi$. The only difference between $(\{X'_t\}, \{Y'_t\})$ and $(\{X_t\}, \{Y_t\})$ is that $\{X'_t\}$ and $\{Y'_t\}$ are independent all the time. Then

$$\begin{aligned}
&\mathbf{Pr}\left[X_{t^*} = Y_{t^*} = 1 \wedge \bar{B}\right] \\
=&\mathbf{Pr}\left[X'_{t^*} = Y'_{t^*} = 1 \wedge \bar{B}\right] \\
=&\mathbf{Pr}\left[X'_{t^*} = 1\right] \cdot \mathbf{Pr}\left[Y'_{t^*} = 1\right] \\
&- \sum_{t=0}^{t^*-1} \sum_{z\in[n]} \mathbf{Pr}\left[X'_t = Y'_t = z \wedge \forall s < t, X'_s \neq Y'_s\right] \\
&\qquad \cdot \mathbf{Pr}\left[X'_{t^*} = 1 \,\middle|\, X'_t = z\right] \\
&\qquad\qquad \cdot \mathbf{Pr}\left[Y'_{t^*} = 1 \,\middle|\, Y'_t = z\right].
\end{aligned}$$

Note that

$$\begin{aligned}
&\mathbf{Pr}\left[X_{t^*} = Y_{t^*} \wedge B\right] \\
\geq&\mathbf{Pr}\left[X_{t^*} = Y_{t^*} = 1 \wedge B\right] \\
=&\sum_{t=0}^{t^*-1} \sum_{z\in[n]} \mathbf{Pr}\left[X_t = Y_t = z \wedge \forall s < t, X_s \neq Y_s\right] \cdot \mathbf{Pr}\left[X_{t^*} = 1 \mid X_t = z\right] \\
=&\sum_{t=0}^{t^*-1} \sum_{z\in[n]} \mathbf{Pr}\left[X'_t = Y_t = z \wedge \forall s < t, X'_s \neq Y'_s\right] \cdot \mathbf{Pr}\left[X'_{t^*} = 1 \,\middle|\, X'_t = z\right].
\end{aligned}$$

Thus, Equation (4)$\geq \mathbf{Pr}\left[X'_{t^*} = 1\right] \cdot \mathbf{Pr}\left[Y'_{t^*} = 1\right] \geq \alpha^2$.

By the coupling and the Markov property, we have

$$\begin{aligned}
\mathbf{Pr}\left[X_{2t^*} \neq Y_{2t^*}\right] &= \mathbf{Pr}\left[X_{2t^*} \neq Y_{2t^*} | X_{t^*} = Y_{t^*}\right] \mathbf{Pr}\left[X_{t^*} = Y_{t^*}\right] \\
&\quad + \mathbf{Pr}\left[X_{2t^*} \neq Y_{2t^*} | X_{t^*} \neq Y_{t^*}\right] \mathbf{Pr}\left[X_{t^*} \neq Y_{t^*}\right] \\
&\leq \mathbf{Pr}\left[X_{2t^*} \neq Y_{2t^*} | X_{t^*} \neq Y_{t^*}\right] \mathbf{Pr}\left[X_{t^*} \neq Y_{t^*}\right] \\
&\leq (1-\alpha^2)^2.
\end{aligned}$$

Then we have $\mathbf{Pr}\left[X_{kt^*} \neq Y_{kt^*}\right] \leq (1-\alpha^2)^k$ by recursion. It yields that

$$\mathbf{Pr}\left[X_t \neq Y_t\right] = \sum_{x_0, y_0 \in [n]} \mu_0(x_0) \cdot \pi(y_0) \cdot \mathbf{Pr}\left[X_t \neq Y_t | X_0 = x_0, Y_0 = y_0\right] \to 0$$

as $t \to \infty$. □

12 Mixing Time

We are ready to study the convergence rate of Markov chains. We start with the notion of mixing time. For any $\varepsilon > 0$, the mixing time of a Markov chain P up to error ε is the minimum step t such that if we run the Markov chain from any initial distribution, its total variation distance to the stationary distribution is at most ε. Formally,

$$\tau_{\text{mix}}(\varepsilon) := \max_{\mu_0} \min_t D_{\text{TV}}(\mu_t, \pi) \leq \varepsilon.$$

Recalling in our proof of FTMC using the coupling argument, we obtain the following inequality

$$D_{\text{TV}}(\mu_t, \pi) \leq \mathbf{Pr}_{(X_t, Y_t)\sim\omega_t}\left[X_t \neq Y_t\right].$$

Therefore, if we can construct a coupling ω_t such that for two arbitrary initial distributions, $\mathbf{Pr}_{(X_t, Y_t)\sim\omega_t}\left[X_t \neq Y_t\right] \leq \varepsilon$, then $\tau_{\text{mix}}(\varepsilon) \leq t$.

Example 12 (Random walk on hypercube) . Consider the random walk on the n-cube. The state space $\Omega = \{0, 1\}^n$, and there is an edge between two state x and y iff $\|x - y\|_1 = 1$. We start from a point $X_0 \in \Omega$. In each step,

- With probability $\frac{1}{2}$ do nothing.
- Otherwise, pick $i \in [n]$ uniformly at random and flip $X(i)$.

It's equivalent to the following process:

- Pick $i \in [n], b \in \{0, 1\}$ uniformly at random.
- Change $X(i)$ to b.

Now we analyze the mixing time of the process using coupling. We apply the following simple coupling rule:

- We couple two walks X_t and Y_t by choosing the same i, b in every step.

Once a position $i \in [n]$ has been picked, $X_t(i)$ and $Y_t(i)$ will be the same forever. Therefore, the problem again reduces to the coupon collector problem.

For $t \geq n \log n + cn$, the probability that the i^{th} dimension is not chosen is

$$\left(1 - \frac{1}{n}\right)^{n \log n + cn} \leq \frac{e^{-c}}{n}.$$

Then the probability that there exists at least one dimension which is not chosen is no larger than e^{-c}. We want this value to be less than ϵ. Then we choose $c > \log \frac{1}{\epsilon}$. Thus,

$$\tau_{\text{mix}}(\varepsilon) \leq n \log \frac{n}{\varepsilon}.$$

Let's modify the process a bit by changing $\frac{1}{2}$ into $\frac{1}{n+1}$, i.e. w.p. $\frac{1}{n+1}$ do nothing, to make the lazy walk more active. Note that we add the lazy move in order to make the chain aperiodic.

Now in this case, we describe another coupling of X_t, Y_t. Without loss of generality, we can reorder the entries of two vectors so that all disagreeing entries come first. Namely there exists an index k such that $X_t(i) \neq Y_t(i)$ if $1 \leq i \leq k$, and $X_t(i) = Y_t(i)$ for $i > k$. Our coupling is as follows:

- If $k = 0$, Y acts the same as X.
- If $k = 1$, Y acts the same as X except when X flips the first entry, Y does nothing and vice versa.
- For $k > 2$, we distinguish between whether X flip indices in $[k]$:
  - If X did nothing or flipped one of $i > k$: Y acts the same.
  - If X flipped $1 \leq i \leq k$: Y flips $(i \bmod k) + 1$, i.e. $1 \mapsto 2, 2 \mapsto 3, \cdots, k-1 \mapsto k, k \mapsto 1$.

It's clear that the above is indeed a coupling. In fact, this coupling acts like a doubled speed coupon collector, since in the case $k > 2$ we can always collect two coupons at a time when lady luck is smiling. It is therefore conceivable that

$$\tau_{\text{mix}} \leq \frac{1}{2} n \log n + O(n).$$

Example 13 (Shuffling cards) . Given a deck of n cards, consider the following rule of shuffling

- pick a card uniformly at random;
- put the card on the top.

The shuffling rule can be viewed as a random walk on all $n!$ permutations of the n cards and it is easy to verify that the uniform distribution is the stationary distribution. Let us design a coupling for this Markov chain. That is, let X_t and Y_t be decks of cards, and we construct X_{t+1} and Y_{t+1} by

- picking the same random card and put it on the top.

[13]

This is clearly a coupling, and once some card, say $\heartsuit K$ has been picked, then $\heartsuit K$ in two decks will be always at the same location. Therefore, if we ask in how many rounds T, $X_T = Y_T$, then the question is equivalent to the coupon collector problem again. So we have,

$$\tau_{\text{mix}}(\varepsilon) \leq n \log \frac{n}{\varepsilon}.$$

13 Reversible Markov Chains

A Markov chain P over state space $[n]$ is *(time) reversible* if there exists some distribution π satisfying

$$\forall i, j \in [n], \ \pi(i)P(i,j) = \pi(j)P(j,i).$$

This family of identities is called *detailed balance conditions.* Moreover, the distribution π must be a stationary distribution of P. To see this, note that

$$\pi^{\mathsf{T}} P(j) = \sum_{i \in [n]} \pi(i)P(i,j) = \sum_{i \in [n]} \pi(j)P(j,i) = \pi(j).$$

[13] Note that we are picking the "same card", not the card at the same location. That is, we draw a random card from X_t, say $\heartsuit K$, and then we pick $\heartsuit K$ in Y_t as well.

The name *reversible* comes from the fact that for any sequence of variables $X_0, X_1, \dots, X_t$ following the chain which start from the stationary distribution, the distribution of $(X_0, X_1, \dots, X_{t-1}, X_t)$ is identical to the distribution of $(X_t, X_{t-1}, \dots, X_1, X_0)$, namely for all $x_0, x_2, \dots x_t \in [n]$,

$$\begin{aligned}
&\mathbf{Pr}_{X_0\sim\pi}\left[X_0 = x_0, X_1 = x_1, \dots, X_t = x_t\right] \\
=\ &\pi(x_0)P(x_0, x_1)\cdots P(x_{t-1}, x_t) \\
=\ &\pi(x_t)P(x_t, x_{t-1})\cdots P(x_1, x_0) \\
=\ &\mathbf{Pr}_{X_0\sim\pi}\left[X_0 = x_t, X_1 = x_{t-1}, \dots, X_t = x_0\right]
\end{aligned}$$

We will study reversible chains since their transition matrices are essentially *symmetric* in some sense, so many powerful tools in linear algebra apply. We will also see that reversible chains are general enough for most of our (algorithmic) applications. You can verify that the the random walks on the hypercube is reversible Markov chains with respect to uniform distribution.

Recall the two conditions of FTMC: irreducibility and aperiodicity. Since the transition graph is undirected if we only consider the connectivity, irreducibility is equivalent to the connectivity of the transition graph. Aperiodicity, on the other hand, is equivalent to that the graph is *not* bipartite.

14 The Metropolis Algorithm

Given a distribution π over a state space Ω, how can we design a Markov chain P so that π is the stationary distribution of P? The *Metropolis algorithm* provides a way to achieve the goal as long as the transition graph G is connected and undirected.

Let Δ be the maximum degree of the transition graph except selfloop (that is $\Delta \triangleq \max_{u\in[n]} \sum_{v\neq u\in[n]} \mathbb{1}[(u,v)\in E]$). We describe the following process to construct a transition matrix P: Choose $k \in [\Delta+1]$ uniformly at random. For any $i \in [n]$, let $\{j_1, j_2, \dots, j_d\}$ be the d neighbours of i. We consider the transition at state i:

- If $d + 1 \le k \le \Delta + 1$, do nothing.
- If $k \le d$,
  - propose to move from i to j_k.
  - accept the proposal with probability $\min\left\{\frac{\pi(j_k)}{\pi(i)}, 1\right\}$.

Then the transition matrix is, for $i, j \in [n]$,

$$P(i,j) = \begin{cases} \frac{1}{\Delta+1}\min\left\{\frac{\pi(j)}{\pi(i)}, 1\right\}, & \text{if } i \neq j; \\ 1 - \sum_{k\neq i} P(i,k), & \text{if } i = j. \end{cases}$$

We can verify that P is reversible with respect to π:

$$\begin{aligned}
&\forall i, j \in \Omega: \\
&\pi(i)P(i,j) = \pi(i)\cdot\frac{1}{\Delta+1}\min\left\{\frac{\pi(j)}{\pi(i)}, 1\right\} = \frac{\min\{\pi(i), \pi(j)\}}{\Delta+1} = \pi(j)P(j,i).
\end{aligned}$$

[14]

Example 14 We give a toy example to show how the algorithm works. Consider a graph with 3 vertices $\{a, b, c\}$. There are undirected edges between (a, b), (b, c) and (a, c) and selfloops for each vertex. In this situation, $\Delta = 2$. If we want to design a transition matrix P with stationary distribution $(\frac{1}{2}, \frac{1}{3}, \frac{1}{6})$, by Metropolis algorithm we have

$$\begin{aligned}
P(a,b) &= \frac{1}{2+1}\cdot\frac{2}{3} = \frac{2}{9}, \\
P(a,c) &= \frac{1}{2+1}\cdot\frac{1}{3} = \frac{1}{9}, \\
P(a,a) &= 1 - \frac{1}{9} - \frac{2}{9} = \frac{2}{3}.
\end{aligned}$$

15 Sampling Proper Colorings

Let's consider the problem of sampling proper colorings. Given a graph $G = (V, E)$, we want to color the vertices using q colors under the condition that no two adjacent vertices share the same color. More formally, a coloring of G is a mapping $c : V \mapsto [q]$, and we call it *proper* iff $\forall \{u, v\} \in E, c(u) \neq c(v)$. The problem is NP-hard in general. However, for $q > \Delta$ there's

[14] The advantage of the Metropolis algorithm is that we do not need to know π in order to implement the algorithm. We only need to know the quantity $\frac{\pi(j)}{\pi(i)}$, which is much easier to compute in many applications.

always at least one suitable solution and can be easily obtained by a greedy algorithm, where Δ is the maximum degree of the graph.

Let $\mathscr{C}$ be the set of all proper colorings. We want to sample uniformly on $\mathscr{C}$. Consider the following Markov chain (assume that the start state is a proper coloring):

- Pick $v \in V$ and $c \in [q]$ uniformly at random.
- Recolor v with c if the modified coloring is still proper.

The chain is aperiodic since selfloops exist in the walk. For $q \geq \Delta + 2$, the chain is irreducible.

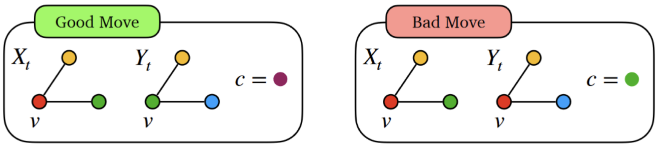
We verify that this Markov chain is reversible, that is, there exist a distribution π that for any $\sigma, \sigma' \in \mathscr{C}$, $\pi(\sigma)P(\sigma, \sigma') = \pi(\sigma')P(\sigma', \sigma)$.

- If $\sigma = \sigma'$, it is obvious that $P(\sigma, \sigma') = P(\sigma', \sigma)$.
- If σ and σ' differ at more than 1 vertices, then $P(\sigma, \sigma') = P(\sigma', \sigma) = 0$.
- If σ and σ' differ at exactly 1 vertex, then $P(\sigma, \sigma') = \frac{1}{n} \cdot \frac{1}{q} = P(\sigma', \sigma)$.

Thus, it is reversible with respect to the uniform distribution and furthermore, uniform distribution on $\mathscr{C}$ is the stationary distribution of the Markov chain defined above.

Suppose X_t, Y_t are two proper colorings. We define the distance $d(X_t, Y_t)$ as their Hamming distance, i.e. the number of vertices colored differently in two colorings. Our coupling of two chains is that we always choose the same v, c in each step. The distance between two colorings can change at most 1 since only v is affected. The possible changes can be divided into two cases:

- Good move: $X_t(v) \neq Y_t(v)$, and both change into c successfully. This will decrease distance by 1.
- Bad move: $X_t(v) = Y_t(v)$, and exactly one change succeeds. This will increase distance by 1.



Consider the probabilities of two types of moves. For good moves, w.p. $\frac{d(X_t,Y_t)}{n}$, $X_t(v) \neq Y_t(v)$, and there are at least $q - 2\Delta$ choices of c to make it a good move. So

$$\begin{aligned}\mathbf{Pr}\left[d(X_{t+1}, Y_{t+1}) = d(X_t, Y_t) - 1\right] &= \mathbf{Pr}\left[(v, c) \text{ is a good move}\right] \\ &\geq \frac{d(X_t, Y_t)}{n} \cdot \frac{q - 2\Delta}{q}.\end{aligned}$$

For bad moves, there exists a neighbor w of v such that its color is different in two colorings, and in one coloring w is of color c. Note that there are at most 2Δ choices of c to make it a bad move. So we have

$$\begin{aligned}\mathbf{Pr}\left[d(X_{t+1}, Y_{t+1}) = d(X_t, Y_t) + 1\right] &= \mathbf{Pr}_{(v,c) \in V \times [q]}\left[(v, c) \text{ is a bad move}\right] \\ &\leq \frac{d(X_t, Y_t)}{n} \cdot \frac{2\Delta}{q}.\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbf{E}\left[d(X_{t+1}, Y_{t+1}) | (X_t, Y_t)\right] &= d(X_t, Y_t) + \mathbf{Pr}\left[d(X_{t+1}, Y_{t+1}) = d(X_t, Y_t) + 1\right] \\ &\quad - \mathbf{Pr}\left[d(X_{t+1}, Y_{t+1}) = d(X_t, Y_t) - 1\right] \\ &\leq d(X_t, Y_t) + \frac{d(X_t, Y_t)}{n} \cdot \frac{2\delta}{q} - \frac{d(X_t, Y_t)}{n} \cdot \frac{q - 2\Delta}{q} \\ &\leq d(X_t, Y_t) \left(1 - \frac{q - 4\Delta}{nq}\right).\end{aligned}$$

In the case $q > 4\Delta$,

$$D_{\mathrm{TV}}(X_{t+1}, Y_{t+1}) \leq \mathbf{E}\left[d(X_{t+1}, Y_{t+1})\right] \leq \left(1 - \frac{1}{nq}\right)^t n \leq \varepsilon.$$

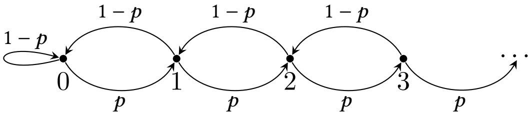
The mixing time is therefore bounded by

$$\tau_{\mathrm{mix}}(\varepsilon) \leq nq \log \frac{n}{\varepsilon}.$$

16 Countably Infinite Markov Chains

We have proved that finite Markov chain must have a stationary distribution using Perron-Frobenius Theorem. However, when the Markov chain has infinite states, even it's countable infinite, there is something going wrong.

Consider the following random walk on $\mathbb{N}$. The state space is $\mathbb{N}$ and at each state i, go to $i+1$ w.p. p and go to $i-1$ w.p $1-p$ (if $i=0$, w.p. $1-p$ stay put).



Let π be the stationary distribution of this Markov chain (if there exists a stationary distribution). We have that

$$\begin{aligned}
&\pi(0) = \pi(0)(1-p) + \pi(1)(1-p) && \Longrightarrow \pi(1) = \frac{p}{1-p}\pi(0),\\
&\pi(1) = \pi(0)p + \pi(2)(1-p) && \Longrightarrow \pi(2) = \frac{p}{1-p}\pi(1),\\
&\dots\\
&\pi(i) = \pi(i-1)p + \pi(i+1)(1-p) && \Longrightarrow \pi(i+1) = \frac{p}{1-p}\pi(i).\\
&\dots
\end{aligned}$$

Note that π is a distribution, so $\sum_{i=0}^{\infty}\pi(i) = 1$. Then, we have

- If $p < \frac{1}{2}$, that is, $\frac{p}{1-p} < 1$, then $\sum_{i=0}^{\infty}\left(\frac{p}{1-p}\right)^i \pi(0) = 1$. By direct calculation we have $\pi(0) = \frac{1-2p}{1-p}$ and $\pi(i) = \left(\frac{p}{1-p}\right)^i \frac{1-2p}{1-p}$ for $i \in \mathbb{N}$.
- If $p > \frac{1}{2}$, then $\frac{p}{1-p} > 1$. When $i \to \infty$, if $\pi(0) \neq 0$, $\pi(i) \to \infty$. This yields that $\pi(0) = \pi(1) = \cdots = \pi(i) = \cdots = 0$. The Markov chain doesn't have a stationary distribution in this case.
- If $p = \frac{1}{2}$, $\frac{p}{1-p} = 1$. Then $\pi(0) = \pi(1) = \cdots = \pi(i) = \cdots$ and $\sum_{i=0}^{\infty}\pi(0) = 1$. This yields that $\pi(0) = 0$ and there is no stationary distribution in this case.

16.1 Recurrence

Definition 16.1 [a] For $i \in \Omega$, let $T_i > 0$ be the first hitting time of state i. Let $\mathbf{P}_i = \mathbf{Pr}\left[\cdot|X_0 = i\right]$. We say a state i is *recurrent* if $\mathbf{P}_i[T_i < \infty] = 1$, o.w. we say the state is *transient*.

[a] $T_i \triangleq \min\{t > 0 | X_t = i\}$.

[15]

Let $N_i \triangleq \sum_{t=0}^{\infty}\mathbb{1}[X_t = i]$, then we have the following propositions.

Proposition 16.1 If i is recurrent, then $\mathbf{P}_i[N_i = \infty] = 1$.

Proof. Assume that $\mathbf{P}_i[N_i = \infty] < 1$. Then there exists $\Omega' \subseteq \hat{\Omega}$ such that $N_i < \infty$ on Ω' and $\mathbf{P}_i[\Omega'] > 0$. This means that with probability larger than 0, we will never reach state i after the last time we visit it. This is in contradiction with the fact that i is recurrent. □

Proposition 16.2 If i is recurrent and there exists a finite path from i to j, then

- $\mathbf{P}_i[T_j < \infty] = 1$.
- $\mathbf{P}_j[T_i < \infty] = 1$.
- j is recurrent.

Proof.

- Let $q \triangleq \mathbf{P}_i[\text{reach } j \text{ before returning to } i]$. Since there is a finite path from i to j, we have $q > 0$ and $\mathbf{P}_i[\text{visit } i\ n \text{ times before reac}$ $(1-q)^n$.

[15] Recall the probability space of a stochastic process. One can view the outcomes of the probability space is the set of infinite sequence of real numbers between $[0,1]$, namely $\hat{\Omega} = [0,1]^{\mathbb{N}}$. The sigma-algebra can be defined in a way similar to the problem 1 in our first homework. Therefore, the random variable T_i is therefore a function $\hat{\Omega} \to \mathbb{R}$.

Assume that $\mathbf{P}_i[T_j = \infty] = \alpha > 0$. [16]Then we have $\mathbf{P}_i[T_j = \infty | N_i = \infty] = \alpha$ since $\mathbf{P}_i[N_i = \infty] = 1$. Let T_i^n be the n^{th} time that the chain visits state i. Then

$$\forall n > 0,\ \mathbf{P}_i[T_j > T_i^n | N_i = \infty] \geq \mathbf{P}_i[T_j = \infty | N_i = \infty] = u$$

On the otherhand, we have $\lim_{n\to\infty} \mathbf{P}_i[T_j > T_i^n | N_i = \infty] = \lim_{n\to\infty} \mathbf{P}_i[T_j > T_i^n] = \lim_{n\to\infty}(1-q)^n = 0$. This is a contradiction. Thus, $\mathbf{P}_i[T_j = \infty] = 0$.

- If $\mathbf{P}_j[T_i = \infty] = p > 0$, then we have that $\mathbf{P}_i[T_i = \infty] \geq q \cdot p > 0$. This is in contradiction with the fact that i is recurrent.
- If $\mathbf{P}_j[T_j = \infty] = r > 0$, then $\mathbf{P}_i[T_j = \infty] \geq q \cdot r > 0$. This is in contradiction with the first item of this proposition.

□

17 Recurrence and Positive Recurrence

Recall that we say a state i is *recurrent* if $\mathbf{P}_i[T_i < \infty] = 1$ or equivalently $\mathbf{E}_i[N_i] = \infty$. Otherwise, we say the state is *transient*. A transient state j will be visited for finite times with probability 1. [17] From Proposition 3 of last lecture, we know that *recurrence* is a class property, that is, given a recurrent state i, all the other states that i can reach in finite steps are also recurrent. We are only concerned with irreducible Markov chains in this lecture. So we may say a Markov chain is recurrent or transient in the future.

Example 15 (Drunk person and drunk bird) Imagine a random walk on a grid that we pick a direction uniformly at random at each time step. Can we go back to the original point with probability 1? Or equivalently, is this Markov chain recurrent or transient?

First we consider the one-dimensional grid. Let $X_0 = 0$ and $X_{t+1} = X_t + \Delta$ where Δ is uniformly at random picked from $\{-1, 1\}$. Then,

$$\mathbf{E}_0[N_0] = \mathbf{E}_0[\sum_{t=0}^{\infty} \mathbb{1}[X_t = 0]] = \sum_{t=0}^{\infty} \mathbf{P}_0[X_t = 0] = \sum_{m=0}^{\infty} \mathbf{P}_0[X_{2m} = 0].$$

[18] where the last equality follows from the fact that we can not go back within exactly odd steps. Then let's compute $\mathbf{P}_0[X_{2m} = 0]$ using the Stirling's formula. For $m \geq 1$,

$$\mathbf{P}_0[X_{2m} = 0] = \frac{\binom{2m}{m}}{2^{2m}} \approx \frac{\sqrt{4\pi m}\left(\frac{2m}{e}\right)^{2m}}{2\pi m \left(\frac{m}{e}\right)^{2m}} \cdot 2^{-2m} = \frac{1}{\sqrt{\pi m}}.$$

Thus, $\mathbf{E}_0[N_0] = \sum_{m=0}^{\infty} \mathbf{P}_0[X_{2m} = 0] \approx 1 + \sum_{m=1}^{\infty} \frac{1}{\sqrt{\pi m}}$ which is divergent. This indicates that the Markov chain for random walk on one-dimensional grid is recurrent.

For d-dimensional grid, we regard the game as independently pick Δ_i u.a.r. from $\{-1, 1\}$ for $i \in [d]$ at each time step and walk to $X_{t+1} = X_t + (\Delta_1, \Delta_2, \ldots, \Delta_d)$. So we have that $\mathbf{P}_i[X_{2m} = \mathbf{0}] = (\mathbf{P}_i[X_{2m}(1) = 0])^d \approx \left(\frac{1}{\sqrt{\pi m}}\right)^d$. We know that $1 + \sum_{m=1}^{\infty} \left(\frac{1}{\sqrt{\pi m}}\right)^d$ is divergent if and only if $d \leq 2$. Thus, only if the dimension of the grid is 1 or 2, the random walk is recurrent.

Definition 17.1 (Positive recurrence) If a state i is recurrent and $\mathbf{E}_i[T_i] < \infty$, we say it is *positive recurrent*. If the state is recurrent but with $\mathbf{E}_i[T_i] = \infty$, then we say it is *null recurrent*.

Now we give some examples to distinguish the concept of null recurrence and positive recurrence.

Example 16 (Drunk person) We have proved that the Markov chain of drunk person is recurrent. We can further verify that it is null recurrent by noticing that

$$\mathbf{E}_i[T_i] = \sum_{t=1}^{\infty} \mathbf{P}_i[\mathbb{1}[X_t = i]] \cdot t \approx \sum_{m=1}^{\infty} \left(\frac{1}{\sqrt{\pi m}}\right)^2 \cdot 2m = \infty.$$

Example 17 Recall the random walk on $\mathbb{N}$ we talked about in the last lecture: at each state i, go to $i+1$ w.p. p and go to $i-1$ w.p $1-p$ (if $i = 0$, w.p. $1-p$ stay at 0). Then the following statements hold:

- When $p > \frac{1}{2}$, the Markov chain is transient.
- When $p = \frac{1}{2}$, it is null recurrent.
- When $p < \frac{1}{2}$, it is positive recurrent.

[16] $\mathbf{P}_i[T_j = \infty] = \mathbf{P}_i[T_j = \infty \mid N_i = \infty] \cdot \mathbf{P}_i[N_i = \infty] + \mathbf{P}_i[T_j = \infty \mid N_i < \infty] \cdot \mathbf{P}_i[N_i < \infty] = \mathbf{P}_i[T_j = \infty \mid N_i = \infty]$.

[17] Here we follow the notations of the last lecture, that is: $X_0, X_1, \ldots, X_t, \ldots$ is a sequence of variables that follows the Markov chain P. $T_i \triangleq \inf\{t > 0 : X_t = i\}$, $N_i \triangleq \sum_{t=0}^{\infty} \mathbb{1}[X_t = i]$, $\mathbf{P}_i[\cdot] = \mathbf{Pr}[\cdot | X_0 = i]$ and $\mathbf{E}_i[\cdot] = \mathbf{E}[\cdot | X_0 = i]$.

[18] Stirling's formula: $n! = \sqrt{2\pi n}\left(\frac{n}{e}\right)^n (1 + o(1))$.

18 Laws of Large Numbers

$X_1, X_2, \ldots$ is an infinite sequence of independent and identically distributed Lebesgue integrable random variables with expected value $\mathbf{E}[X_1] = \mathbf{E}[X_2] = \cdots = \mu < \infty$. Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ be the sample average. Then we have the following two laws of large numbers.

Theorem 18.1 (Weak law of large numbers or Khinchin's law) The sample average converge in probability towards the expected value:

$$\bar{X}_n \xrightarrow{p} \mu \quad \textit{when } n \to \infty.$$

That is, for any positive value ε,

$$\lim_{n\to\infty} \mathbf{Pr}\left[\left|\bar{X}_n - \mu\right| < \epsilon\right] = 1.$$

Theorem 18.2 (Strong law of large numbers or Kolmogorov's law) The sample average converges almost surely or with probability 1 to the expected value:

$$\bar{X}_n \xrightarrow{a.s.} \mu \quad \textit{when } n \to \infty.$$

That is, [a]

$$\mathbf{Pr}\left[\lim_{n\to\infty} \bar{X}_n \to \mu\right] = 1.$$

[a] Let $(\Omega, \mathcal{F}, P)$ be the probability space. Here $\bar{X}_n \to \mu$ means $\exists M \in \mathcal{F}$ satisifying

- P(M)=1;
- $\forall \omega \in M$, $\bar{X}_n(\omega) \overset{n\to\infty}{\longrightarrow} \mu$.

As the name of the laws shows, *convergence in probability* is weaker than *convergence with probability* 1. Consider a sequence of independent random variables $X_1, X_2, \ldots$ that X_n is 1 with probability $\frac{1}{n}$ and X_n is 0 with probability $1 - \frac{1}{n}$. Then the sequence converges to 0 in probability but not with probability 1 since we cannot find an $M \in \mathcal{F}$ with measure 1 such that $\bar{X}_n(\omega) \overset{n\to\infty}{\longrightarrow} 0$ for every $\omega \in M$.

Theorem 18.3 (Strong law of large numbers for Markov chains) If there is a finite path from state i to j, then

$$\mathbf{P}_i\left[\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^n \mathbb{1}[X_t = j] = \frac{1}{\mathbf{E}_j[T_j]}\right] = 1.$$

Proof. If j is transient, then the random process will visit j for finite times with probability 1. Thus $\mathbf{P}_i\left[\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^n \mathbb{1}[X_t = j] = \frac{}{\mathbf{E}_j}\right.$ $\mathbf{P}_i\left[\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^n \mathbb{1}[X_t = j] = 0\right] = 1$.

If j is recurrent, we first prove the theorem for $i = j$. We call a loop from j to j a cycle (we visit j only at the beginning and end of the loop). Denote C_r as the length of the r^{th} cycle during the process. Let $S_k = \sum_{r=1}^k C_r$. Let k_n be the number of cycles before the $n+1$ step, that is, $k_n = \max\{k | S_k \le n\}$. Then we have $S_{k_n} \le n < S_{k_n+1}$ and consequently $\frac{S_{k_n}}{k_n} \le \frac{n}{k_n} < \frac{S_{k_n+1}}{k_n}$. Note that with probability 1, $k_n \to \infty$ when $n \to \infty$. We have with probability 1 that

$$\lim_{k\to\infty} \frac{S_k}{k} \le \lim_{n\to\infty} \frac{n}{k_n} < \lim_{k\to\infty} \frac{S_{k+1}}{k}.$$

Note that $S_k = \sum_{r=1}^k C_r$ where each C_r is an i.i.d random variable with mean $\mathbf{E}_j\left[T_j\right]$. So by SLLN (Theorem 18.2), we have $\lim_{k\to\infty} \frac{S_k}{k} = \mathbf{E}_j\left[T_j\right]$ and $\lim_{k\to\infty} \frac{S_{k+1}}{k} = \lim_{k\to\infty} \frac{S_{k+1}}{k+1} \cdot \frac{k+1}{k} = \mathbf{E}_j\left[T_j\right]$. As a result, with probability 1,

$$\mathbf{E}_j\left[T_j\right] = \lim_{n\to\infty} \frac{n}{k_n} = \lim_{n\to\infty} \frac{n}{\sum_{t=1}^n \mathbb{1}[X_t = j]}.$$

If j is recurrent and $i \neq j$, let $T_{i\to j}$ be the first time visiting j. Then we have $\frac{S_{k_n} + T_{i\to j}}{k_n} \le \frac{n}{k_n} < \frac{S_{k_n+1} + T_{i\to j}}{k_n}$. Since $\mathbf{P}_i[T_j < \infty] = 1$, $\mathbf{P}_i[\lim_{k\to\infty} \frac{T_{i\to j}}{k} = 0] = 1$. The remaining proof is the same with the situation that $i = j$. □

Corollary 18.1 Let P be the transition function of an irreducible Markov chain where $P^t(i,j) = \mathbf{Pr}[X_t = j | X_0 = i]$. Then for any states i, j,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^n P^t(i,j) = \frac{1}{\mathbf{E}_j[T_j]}.$$

Proof. By the strong law of large numbers for Markov chains, there exists a set $M \in \mathcal{F}$ such that $P(M) = 1$ and

$\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} \mathbb{1}[X_t(\omega)=j] = \frac{1}{\mathbf{E}_j[T_j]}$ for any $\omega \in M$. Then,

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n} P^t(i,j) &= \lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\mathbf{E}_i\left[\mathbb{1}[X_t=j]\right]\\ &= \lim_{n\to\infty}\mathbf{E}_i\left[\frac{1}{n}\sum_{t=1}^{n}\mathbb{1}[X_t=j]\right]\\ &= \mathbf{E}_i\left[\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\mathbb{1}[X_t=j]\right]\\ &= \frac{1}{\mathbf{E}_j[T_j]},\end{aligned}$$

[19] where the third equation follows from the bounded convergence theorem. □

19 Fundamental Theorem

First we introduce some abbreviations to simplify the expression:

- Aperiodicity:[A],
- Irreducibility:[I],
- Recurrence:[R],
- Positive Recurrence: [PR],
- Has a stationary distribution:[S],
- Has a unique stationary distribution:[U],
- Convergence:[C],
- Finiteness:[F].

The finite FTMC can be written as: [F]+[A]+[I]⇒[S]+[U]+[C]. For infinite Markov chains, the theorem need to be modified as: [PR]+[A]+[I]⇒[S]+[U]+[C]. We will first prove the existence and uniqueness of the stationary distribution in this lecture.(i.e. [S] and [U])

Theorem 19.1 [I]+[PR]⇒[S]+[U].

Proof. [Proof of [U]] Let $\mathcal{S}$ be the set of states. Assume π is a stationary distribution of the Markov chain, i.e.,

$$\forall j \in \mathcal{S},\ \forall t \geq 0,\ \sum_{i\in\mathcal{S}}\pi(i)P^t(i,j) = \pi(j).$$

This yields that for $n \geq 1$,

$$\frac{1}{n}\sum_{i\in\mathcal{S}}\pi(i)\sum_{t=1}^{n}P^t(i,j) = \pi(j).$$

[20] Taking $n \to \infty$ and applying the dominated convergence theorem, we have

$$\pi(j) = \sum_{i\in\mathcal{S}}\pi(i)\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}P^t(i,j) = \sum_{i\in\mathcal{S}}\pi(i)\cdot\frac{1}{\mathbf{E}_j[T_j]} = \frac{1}{\mathbf{E}_j[T_j]}.$$

□

Proof. [Proof of [S]] Then we prove the above π is a stationary distribution.

$\mathcal{S}$ is finite. We first assume $\mathcal{S}$ is finite, so that we can safely exchange the order of taking limitation and summation in the calculations below.

$$\begin{aligned}\sum_{j\in\mathcal{S}}\pi(j) &= \sum_{j\in\mathcal{S}}\frac{1}{\mathbf{E}_j[T_j]} = \sum_{j\in\mathcal{S}}\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}P^t(i,j)\\ &= \lim_{n\to\infty}\sum_{j\in\mathcal{S}}\frac{1}{n}\sum_{t=1}^{n}P^t(i,j) = \lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\sum_{j\in\mathcal{S}}P^t(i,j) = 1.\end{aligned}$$

[19] Bounded Convergence Theorem: If $X_n \xrightarrow{a.s.} X$ and $\mathbf{E}[X] < \infty$, then $\mathbf{E}[X_n] \to \mathbf{E}[X]$.

[20] Dominated Convergence Theorem: If $\int_S |f_n| < \infty$, then $\lim_{n\to\infty}\int_S f_n = \int_S \lim_{n\to\infty} f_n$.

This indicates that π is a legal distribution. We then verify that π is indeed the stationary distribution.

Note that $P^{t+1}(i,j) = \sum_{k\in\mathcal{S}} P^t(i,k)P(k,j)$. Then

$$\begin{aligned}\frac{1}{\mathbf{E}_j[T_j]} &= \lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}P^t(i,j) = \lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}P^{t+1}(i,j)\\ &= \lim_{n\to\infty}\sum_{k\in\mathcal{S}}P(k,j)\frac{1}{n}\sum_{t=1}^{n}P^t(i,k) = \sum_{k\in\mathcal{S}}P(k,j)\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}P^t(i,k)\\ &= \sum_{k\in\mathcal{S}}P(k,j)\cdot\frac{1}{\mathbf{E}_k[T_k]}.\end{aligned}$$

That is,

$$\pi(j) = \sum_{k\in\mathcal{S}}P(k,j)\pi(k).$$

It is worth noting that [PR] is equivalent to [I] when $\mathcal{S}$ is finite.

$\mathcal{S}$ is infinite. When $\mathcal{S}$ is (countably) infinite, we consider every finite subset A of $\mathcal{S}$. Then

$$\begin{aligned}\sum_{j\in A}\pi(j) &= \sum_{j\in A}\frac{1}{\mathbf{E}_j[T_j]} = \sum_{j\in A}\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}P^t(i,j)\\ &= \lim_{n\to\infty}\sum_{j\in A}\frac{1}{n}\sum_{t=1}^{n}P^t(i,j) = \lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\sum_{j\in A}P^t(i,j) < 1.\end{aligned}$$

Therefore

$$\sum_{j\in\mathcal{S}}\pi(j) = \sup_{\text{finite } A\subseteq\mathcal{S}}\sum_{j\in A}\pi(j) =: C \le 1.$$

Since $[PR]$, we know that $C \neq 0$. In the following, we will prove that π/C is a stationary distribution. Then $C = 1$ follows from the uniqueness of the stationary distribution we just proved.

For every finite $A \subseteq \mathcal{S}$, we have

$$\sum_{k\in A}P(k,j)\cdot\frac{1}{\mathbf{E}_k\left[T_k\right]} \le \frac{1}{\mathbf{E}_j\left[T_j\right]}.$$

Therefore,

$$\sum_{k\in\mathcal{S}}P(k,j)\cdot\frac{1}{\mathbf{E}_k\left[T_k\right]} = \sup_{\text{finite } A\subseteq\mathcal{S}}\sum_{k\in A}P(k,j)\cdot\frac{1}{\mathbf{E}_k\left[T_k\right]} \le \frac{1}{\mathbf{E}_j\left[T_j\right]}.$$

We show that indeed the equality holds. Assume for a contradiction that

$$\sum_{k\in\mathcal{S}}P(k,j)\cdot\frac{1}{\mathbf{E}_k\left[T_k\right]} < \frac{1}{\mathbf{E}_j\left[T_j\right]}.$$

Summing the both sides over all $j \in \mathcal{S}$, we obtain

$$\sum_{k\in\mathcal{S}}\frac{1}{\mathbf{E}_k\left[T_k\right]} < \sum_{j\in\mathcal{S}}\frac{1}{\mathbf{E}_j\left[T_j\right]},$$

which is a contradiction. As a result, we know

$$\sum_{k\in\mathcal{S}}P(k,j)\cdot\frac{1}{\mathbf{E}_k\left[T_k\right]} = \frac{1}{\mathbf{E}_j\left[T_j\right]},$$

and $\hat{\pi}(j) = \frac{1}{C\cdot\mathbf{E}_j[T_j]}$ is a stationary distribution. By the uniqueness of the distribution, we have $C = 1$.

□

20 Galton-Watson Process

The model was formulated by F. Galton in the study of the survival and extinction of family names. In the nineteenth century, there was concern amongst the Victorians that aristocratic surnames were becoming extinct. In 1873, Galton originally posed the question regarding the probability of such an event, and later H. W. Watson replied with a solution.

Using more modern terms, the process can be defined formally as follows:

Definition 20.1 (Galton-Watson Process) Suppose that all the individuals reproduce independently of each other and have the same offspring distribution. More precisely, let G_t denote the number of individuals of t-th generation:

- We start from the zero generation. For convenience, let $G_0 = 1$.
- Each individual of generation t gives birth to a random number of children of generation $t+1$. That is, $\forall t \geq 0$ and $i \in [G_t]$, let $X_{t,i}$ denote the number of children of the i-th individual in the t-th generation. Then $\{X_{t,i}\}$ is an array of i.i.d. random variables with $\mathbf{Pr}\left[X_{t,i} = k\right] = p_k$.
- All individuals of generation $t+1$ are children of individuals of generation t:[a]

$$G_{t+1} = \sum_{i=1}^{G_t} X_{t,i}$$

[a] It is clear that the process $\{G_t\}_{t\geq 0}$ is a Markov chain.

Denote by ρ the probability of extinction, namely

$$\rho \triangleq \mathbf{Pr}\,[\text{extinction}] = \mathbf{Pr}\left[\cup_{t\geq 1}\{G_t = 0\}\right].$$

Then the question is to determine the value of ρ. First we consider two trivial situations:

- When $p_0 = 0$, it is clear that there will be offspring and $\rho = 0$.
- When $p_0 > 0$ and $p_0 + p_1 = 1$, we can verify that $\rho = 1$. We know that

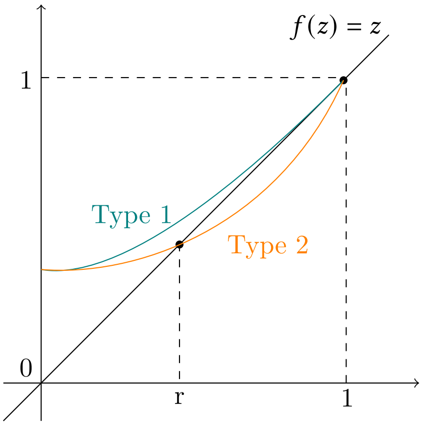
$$\mathbf{E}\left[G_{t+1} | G_t\right] = p_1 \cdot G_t.$$

Compute the expectation of both sides, we have $\mathbf{E}\left[G_{t+1}\right] = p_1 \mathbf{E}\left[G_t\right]$. This yields that when $t \to \infty$, $\mathbf{Pr}\left[G_t \geq 1\right] \leq \mathbf{E}\left[G_t\right] = p_1^t \mathbf{E}\left[G_0\right] \to 0$.

Then we assume that $p_0 > 0$ and $p_0 + p_1 < 1$. By the independence of each individual and the Markov property, we can calculate ρ as follows:

$$\begin{aligned}\rho &= \sum_{k=0}^{\infty} \mathbf{Pr}\left[\text{extinction} \wedge G_1 = k\right] \\ &= \sum_{k=0}^{\infty} \mathbf{Pr}\left[\text{extinction} | G_1 = k\right] p_k \\ &= \sum_{k=0}^{\infty} \rho^k p_k. \qquad (5)\end{aligned}$$

Let $\psi(z) \triangleq \sum_{k=0}^{\infty} p_k z^k$. Then Equation (5) yields that ρ is a fixed point of ψ, i.e., $\psi(\rho) = \rho$. By direct calculation we know ψ is an increasing and convex function on $[0,1]$ with $\psi(0) = p_0$ and $\psi(1) = 1$. Then there can be two types of ψ depending on whether $\psi'(1)$ is larger than 1 as the following figure shows.



When $\psi'(1) = \sum_{k=1}^{\infty} k p_k = \mathbf{E}\left[X_{t-i}\right] \leq 1$, $z = 1$ is the only fixed point of ψ which corresponds to the Type 1 in the figure. That is to say, when $\mathbf{E}\left[X_{t-i}\right] \leq 1$, we have $\rho = 1$.

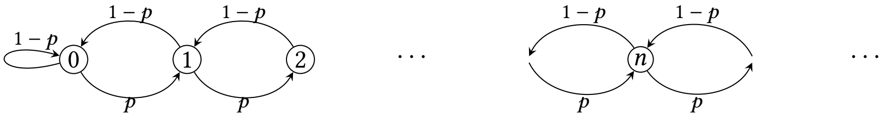
When $\mathbf{E}\left[X_{t-i}\right] > 1$ (Type 2), although there are two fixed points of ψ: r and 1, we claim that $\rho = r$ rather than 1 by showing that $\rho \leq r$. Let $q_t \triangleq \mathbf{Pr}\left[G_t = 0\right]$. Then $q_t \leq q_{t+1} < 1$ since $G_t = 0$ can always yields $G_{t+1} = 0$. We induct on t to show that $q_t \leq r$:

- When $t = 0$, it is obvious that $q_0 = 0 < r$.
- Assume that $q_t \leq r$. Since $q_{t+1} = \sum_{k=0}^{\infty} p_k q_t^k = \psi(q_t)$ and ψ is an increasing function, $q_{t+1} = \psi(q_t) \leq \psi(r) = r$.

We know that $\rho = \lim_{t\to\infty} q_t$ and $q_t \leq r$ for all $t \geq 0$. Thus $\rho \leq r$. However, we have shown that ρ is a fixed point of ψ. So $\rho = r$ when $\mathbf{E}[X_{t-i}] > 1$. In conclusion, $\rho = 1$ iff $\mathbf{E}[X_{t-i}] \leq 1$.

21 1-D Random Walk

Consider the following one-dimensional random walk:



Let X_t be the position at time step t. Let $T_{i\to j}$ be the first hitting time of state j starting from i, that is, $T_{i\to j} = \min\{t > 0 | X_t = j \wedge X_0 = i\}$. Define event $\mathscr{A}$ = [the first step is to the right]. Then we consider the problem that when will this Markov chain be recurrent. Note that

$$\begin{aligned}
\mathbf{Pr}[T_{0\to 0} < \infty] &= \mathbf{Pr}\left[T_{0\to 0} < \infty | \bar{\mathscr{A}}\right] \mathbf{Pr}\left[\bar{\mathscr{A}}\right] + \mathbf{Pr}[T_{0\to 0} < \infty | \mathscr{A}] \, \mathbf{Pr}[\mathscr{A}] \\
&= (1-p)\cdot 1 + p \cdot \mathbf{Pr}[T_{1\to 0} < \infty], \qquad (6) \\
\mathbf{Pr}[T_{1\to 0} < \infty] &= \mathbf{Pr}\left[T_{1\to 0} < \infty | \bar{\mathscr{A}}\right] \mathbf{Pr}\left[\bar{\mathscr{A}}\right] + \mathbf{Pr}[T_{1\to 0} < \infty | \mathscr{A}] \, \mathbf{Pr}[\mathscr{A}] \\
&= (1-p)\cdot 1 + p \cdot \mathbf{Pr}[T_{2\to 0} < \infty], \qquad (7) \\
\mathbf{Pr}[T_{2\to 0} < \infty] &= \mathbf{Pr}[T_{2\to 1} < \infty \wedge T_{1\to 0} < \infty] \\
&= \mathbf{Pr}[T_{2\to 1} < \infty] \cdot \mathbf{Pr}[T_{1\to 0} < \infty] \\
&= \mathbf{Pr}[T_{1\to 0} < \infty]^2 . \qquad (8)
\end{aligned}$$

Let $y \triangleq \mathbf{Pr}[T_{1\to 0} < \infty]$ for brevity. Combine Equation (7) and Equation (8), we have $y = 1 - p + py^2$ which then yields $y = 1$ or $y = \frac{1-p}{p}$. By Equation (6), $\mathbf{Pr}[T_{0\to 0} < \infty] = 1$ or $2 - 2p$.

- When $p < \frac{1}{2}$, $2 - 2p$ is meaningless as a probability. So $\mathbf{Pr}[T_{0\to 0} < \infty] = 1$ and the Markov chain is recurrent.
- When $p = \frac{1}{2}$, $2 - 2p = 1$. The Markov chain is also recurrent in this situation.
- When $p > \frac{1}{2}$, we verify that $\mathbf{Pr}[T_{0\to 0} < \infty] < 1$, and therefore $\mathbf{Pr}[T_{0\to 0} < \infty] = 2 - 2p$. Let $\{\Delta_k\}_{k=0}^{\infty}$ be a sequence of i.i.d. random variables with

  $$\Delta_k = \begin{cases} +1, & \text{w.p. } p \\ -1, & \text{w.p. } 1-p \end{cases}.$$

  Given a sufficiently large $n \in \mathbb{N}$, we can walk to n from 0 in n steps (i.e. $X_n = n$) with probability $p^n > 0$. Assume that we have arrived at n, consider the probability that we go back to 0 from n in exactly k steps. Apparently, this probability is zero when $k < n$. For every $k \geq n$, we upper bound the probability $\mathbf{Pr}[T_{n\to 0} = k]$:

  $$\begin{aligned}
  \mathbf{Pr}[T_{n\to 0} = k] &\leq \mathbf{Pr}\left[\sum_{t=1}^{k} \Delta_t = -n\right] \\
  &\leq \mathbf{Pr}\left[\sum_{t=1}^{k} \Delta_t - \mathbf{E}\left[\sum_{t=1}^{k} \Delta_t\right] \leq -n - \mathbf{E}\left[\sum_{t=1}^{k} \Delta_t\right]\right] \\
  &\leq \exp\left\{-\frac{2k\left(\frac{n+(2p-1)k}{k}\right)^2}{4}\right\}.
  \end{aligned}$$

  where the third inequality follows from the Hoeffding's inequality.

  Then we calculate the probability that we can go back from n to 0. By union bound,

  $$\begin{aligned}
  \mathbf{Pr}[T_{n\to 0} < \infty] &= \mathbf{Pr}\left[\bigcup_{k\geq n} [T_{n\to 0} = k]\right] \\
  &\leq \sum_{k=n}^{\infty} \mathbf{Pr}[T_{n\to 0} = k] \\
  &\leq \exp\{-(2p-1)n\} \sum_{k=n}^{\infty} \exp\left\{-\frac{n^2}{2k} - \frac{(2p-1)^2 k}{2}\right\}.
  \end{aligned}$$

Note that

$$\begin{aligned}\sum_{k=n}^{\infty} \exp\left\{-\frac{n^2}{2k}\right\} \cdot \exp\left\{-\frac{(2p-1)^2 k}{2}\right\} &\le \sum_{k=n}^{\infty} \exp\left\{-\frac{(2p-1)^2 k}{2}\right\} \\ &= \frac{\exp\left\{-\frac{(2p-1)^2}{2} n\right\}}{1-\exp\left\{-\frac{(2p-1)^2}{2}\right\}}\end{aligned}$$

Thus,

$$\mathbf{Pr}\left[T_{n\to 0} < \infty\right] \le \frac{\exp\left\{-\frac{(2p-1)^2}{2} n - (2p-1)n\right\}}{1-\exp\left\{-\frac{(2p-1)^2}{2}\right\}}. \tag{9}$$

We can find a sufficiently large constant n such that $\mathbf{Pr}\left[T_{n\to 0} < \infty\right] < 1$ since the RHS of Equation (9) is exponentially small with regard to n. So for sufficiently large n, the probability that we walk to n and never come back to 0 is larger than $p^n \cdot \mathbf{Pr}\left[T_{n\to 0} = \infty\right] > 0$. Thus, this Markov chain is transient.

Now we verify that the Markov chain is positive recurrent when $p < \frac{1}{2}$ and null recurrent when $p = \frac{1}{2}$. Note that

$$T_{0\to 0} = \mathbb{1}[\bar{\mathscr{A}}] \cdot 1 + \mathbb{1}[\mathscr{A}](1 + T_{1\to 0}) \tag{10}$$

$$T_{1\to 0} = \mathbb{1}[\bar{\mathscr{A}}] \cdot 1 + \mathbb{1}[\mathscr{A}](1 + T_{2\to 0}) \tag{11}$$

$$T_{2\to 0} = T_{2\to 1} + T_{1\to 0} = 2T_{1\to 0}. \tag{12}$$

Taking the expectation of Equation (11) and combining with Equation (12), we have

$$\mathbf{E}\left[T_{1\to 0}\right] = 1 - p + p(1 + 2\mathbf{E}\left[T_{1\to 0}\right]),$$

which yields $\mathbf{E}\left[T_{1\to 0}\right] = \frac{1}{1-2p}$. Take the expectation of Equation (10), we get $\mathbf{E}\left[T_{0\to 0}\right] = \frac{1-p}{1-2p}$. Thus:

- When $p = \frac{1}{2}$, $\mathbf{E}\left[T_{0\to 0}\right] = \infty$ and the Markov chain is null recurrent.
- When $p < \frac{1}{2}$, $\mathbf{E}\left[T_{0\to 0}\right] < \infty$ and the Markov chain is positive recurrent.

22 2-SAT

SAT is the problem of determining whether a CNF formula has satisfying assignments. k-SAT is the special cases of SAT that the clauses of the CNF formula consist of exact k literals. For example,

$$\phi = (x \vee y) \wedge (y \vee \bar{z}) \wedge (\bar{x} \vee z)$$

is a 2-CNF formula and $x = y = z = 1$ is one of its satisfying assignments. SAT is **NP**-complete and we have k-SAT $\in$ **NP** for $k \ge 3$. [21] One can use an algorithm for finding strongly connected components to solve 2- SAT problem in linear time. Nevertheless, we introduce a simple randomized algorithm that can also solve this problem in polynomial-time with high probability.

Let ϕ be a 2-CNF formula and $V = \{v_1, v_2, ..., v_n\}$ be its set of variables The algorithm runs as follows:

- Pick an arbitrary assignment $\sigma_0 : V \to \{\text{true,false}\}$.
- For $t = 0, 1, 2, \ldots, 100n^2$:

  If σ_t satisfies ϕ, output σ_t;

  Else, pick an arbitrary unsatisfying clause, say $c = x \vee y$. Choose from $\{x, y\}$ uniformly at random and flip the assignment of the chosen literal. Let σ_{t+1} be the flipped assignment.
- Output "ϕ is not satisfiable".

Claim 22.1 This algorithm outputs the correct answer with probability at least $1 - \frac{1}{100}$.

Proof. It is clear that if a 2-SAT instance has no solution then our algorithm will always give the correct answer. So we consider the probability that our algorithm outputs no solution conditioned on that the instance indeed has a satisfying assignment.

Our algorithm produces $100n^2 + 1$ assignments $\sigma_0, \sigma_1, \ldots, \sigma_{100n^2}$. We claim that with probability at least $1 - \frac{1}{100}$, some of σ_k for $k \in \left\{0, \ldots, 100n^2 + 1\right\}$ is a satisfying assignment. The argument here, at first glance, is a bit weird. We fix an *arbitrary*

[21] We will extend the algorithm to solving 3-SAT in the homework!

$\sigma : V \to \{\mathtt{true}, \mathtt{false}\}$ satisfying assignment. We in fact prove the following: For large enough k, conditioned on the event that none of $\sigma_0, \sigma_1, \dots, \sigma_k$ is a satisfying assignment, $\sigma_{k+1} = \sigma$ holds with high probability.

Let $\{X_t\}_{t=0}^{100n^2}$ be a random variable sequence that

$$X_t \triangleq |\{v \in V : \sigma_t(v) = \sigma(v)\}|.$$

[22]

First we verify that $\mathbf{Pr}\left[X_{t+1} = X_t + 1 \mid \sigma_t\right] \geq \frac{1}{2}$ [23] and $\mathbf{Pr}\left[X_{t+1} = X_t - 1 \mid \sigma_t\right] \leq \frac{1}{2}$. WLOG assume we chose the clause $c = x \vee y$ in round t. Since c is not satisfied by σ_t, we have $\sigma_t(x) = \sigma_t(y) = \mathtt{false}$. Similarly, $x \vee y$ is satisfying under σ, so there are three possible assignments of $\sigma(x)$ and $\sigma(y)$:

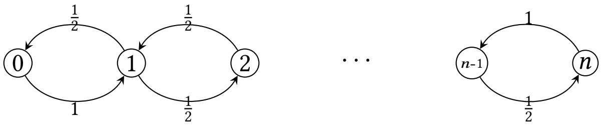
- If $\sigma(x) = \mathtt{true}$ and $\sigma(y) = \mathtt{false}$, $\mathbf{Pr}\left[X_{t+1} = X_t + 1 \mid \sigma_t\right] = \mathbf{Pr}\left[\text{flip } x\right] = \frac{1}{2}$ and $\mathbf{Pr}\left[X_{t+1} = X_t - 1 \mid \sigma_t\right] = \mathbf{Pr}\left[\text{flip } y\right] = \frac{1}{2}$.
- If $\sigma(x) = \mathtt{false}$ and $\sigma(y) = \mathtt{true}$, we have $\mathbf{Pr}\left[X_{t+1} = X_t + 1 \mid \sigma_t\right] = \mathbf{Pr}\left[X_{t+1} = X_t - 1 \mid \sigma_t\right] = \frac{1}{2}$ similarly.
- If $\sigma(x) = \mathtt{true}$ and $\sigma(y) = \mathtt{true}$, $\mathbf{Pr}\left[X_{t+1} = X_t + 1 \mid \sigma_t\right] = \mathbf{Pr}\left[\text{flip } x \text{ or } y\right] = 1$.

Thus we have $\mathbf{Pr}\left[X_{t+1} = X_t + 1 \mid \sigma_t\right] \geq \frac{1}{2}$ on condition that none of $\sigma_0, \sigma_1, \dots, \sigma_t$ is a satisfying assignment.

Consider the 1-D random walk $\{Y_t\}_{t \geq 0}$ on $[n] \cup \{0\}$ that $Y_0 = X_0$ and for $Y_t \notin \{0, 1\}$

$$Y_{t+1} = \begin{cases} Y_t + 1, & \text{w.p. } \frac{1}{2} \\ Y_t - 1, & \text{w.p. } \frac{1}{2} \end{cases}.$$

If $Y_t = 0$, $Y_{t+1} = Y_t + 1$ w.p. 1 and if $Y_t = n$, then $Y_{t+1} = Y_t - 1$ w.p. 1.



Then we have[24]

$$\begin{aligned} \mathbf{Pr}\left[\text{the algorithm is correct}\right] &\geq \mathbf{Pr}\left[\exists t \in [0, 100n^2] \ s.t. X_t = n\right] \\ &\geq \mathbf{Pr}\left[\exists t \in [0, 100n^2] \ s.t. Y_t = n\right]. \end{aligned} \tag{13}$$

Assume that $Y_0 = X_0 = i$. Let $T_{i \to n}$ be the first hitting time of n from i. Then

$$T_{i \to n} = \sum_{k=i}^{n-1} T_{k \to k+1}.$$

For $i > 0$, we have[25]

$$\begin{aligned} T_{i \to i+1} &= \mathbb{1}[\mathscr{A}] + \mathbb{1}[\bar{\mathscr{A}}](1 + T_{i-1 \to i+1}) \\ &= \mathbb{1}[\mathscr{A}] + \mathbb{1}[\bar{\mathscr{A}}](1 + T_{i-1 \to i} + T_{i \to i+1}) \end{aligned}$$

Taking the expectation of both sides, we have $\mathbf{E}\left[T_{i \to i+1}\right] = 2 + \mathbf{E}\left[T_{i-1 \to i}\right]$. Note that $T_{0 \to 1} = 1$, then

$$\mathbf{E}\left[T_{i \to n}\right] = \sum_{k=i}^{n-1} \mathbf{E}\left[T_{k \to k+1}\right] = \sum_{k=i}^{n-1} 2k + 1 = n^2 - i^2 \leq n^2.$$

Then we apply the Markov's inequality to give a lower bound for $\mathbf{Pr}\left[\exists t \in [0, 100n^2] \ s.t. Y_t = n\right]$:

$$\begin{aligned} 1 - \mathbf{Pr}\left[\exists t \in [0, 100n^2] \ s.t. Y_t = n\right] &= \mathbf{Pr}\left[T_{Y_0 \to n} > 100n^2\right] \\ &\leq \frac{\mathbf{E}\left[T_{Y_0 \to n}\right]}{100n^2} \leq \frac{1}{100}. \end{aligned}$$

By Equation (13), we know that $\mathbf{Pr}\left[\text{the algorithm is correct}\right]$ is lower bounded by $1 - \frac{1}{100}$. □

[22] Note that $\{X_t\}_{t=0}^{100n^2}$ is not a Markov chain since it only contains partial information of σ_t and we cannot determine the distribution of X_{t+1} given X_t.

[23] Let Y be a random variable. Then function $\mathbf{Pr}\left[\cdot \mid Y\right] : \mathrm{Ran}(Y) \to \mathbb{R}$ is defined by $\mathbf{Pr}\left[\cdot \mid Y\right] = \mathbf{E}\left[\mathbb{1}[\cdot] \mid Y\right]$. Note that $\mathbf{Pr}\left[\cdot \mid Y\right]$ is a random variable. Here we slightly abuse the notations and denote the event "$\forall a \in \mathrm{Ran}(Y), \mathbf{Pr}\left[\cdot \mid Y = a\right] \geq \frac{1}{2}$" as $\mathbf{Pr}\left[\cdot \mid Y\right] \geq \frac{1}{2}$.

[24] The second inequality can be verified by constructing a coupling which satisfies $Y_t \geq X_t$ for all $t \geq 0$. The existence of such coupling is guaranteed by $\mathbf{Pr}\left[X_{t+1} = X_t + 1 \mid \sigma_t\right] \geq \mathbf{Pr}\left[Y_{t+1} = Y_t + 1\right]$. Specifically, if there is one false and one true in $\{\sigma(x), \sigma(y)\}$, then Y_{t+1} moves the same as X_{t+1}. If $\sigma(x) = \sigma(y)$ =true, then Y_{t+1} moves $+1$ or -1 uniformly at random.

[25] Recall $\mathscr{A}$ = [the first step is to the right].

23 Martingale

Consider a fair gambling game in which the expected gain in each round is zero. As a result, regardless of how much one bets in each round, the money in expectation remains the same. The balances after each round form a *martingale*.

Definition 23.1 (Martingale) Let $\{X_n\}_{n\geq 0}$ and $\{Z_n\}_{n\geq 0}$ be two sequences of random variables. Let $Z_n = \sum_{t=0}^n X_t$.[a] We say $\{Z_n\}_{n\geq 0}$ is a martingale w.r.t. $\{X_n\}_{n\geq 0}$ if

$$\mathbf{E}\left[Z_{n+1} \mid X_0, X_1 \ldots, X_n\right] = Z_n.$$

[a] Consider the problem of fair gambling where X_n is the gain of n-th round and $Z_n = \sum_{t=1}^n X_n$. $\{Z_n\}_{n\geq 0}$ is not necessarily a Markov chain. The value X_n may depend on information before round $n-1$.

Sometimes we say a single sequence $\{X_n\}_{n\geq 0}$ is a martingale if it is a martingale w.r.t. itself. Formally, if for every $n \geq 0$, it holds that

$$\mathbf{E}\left[X_{n+1} \mid X_0, \ldots, X_n\right] = X_n.$$

For convenience, from now on we use $\overline{X}_{i,j} = \left(X_i, X_{i+1} \ldots, X_j\right)$ to simplify the notations. The conditional expectation $\mathbf{E}\left[Z_{n+1} \,\middle|\, \overline{X}_{0,n}\right]$ is equivalent to $\mathbf{E}\left[Z_{n+1} \,\middle|\, \sigma(\overline{X}_{0,n})\right]$ where $\sigma(\overline{X}_{0,n})$ is the σ-algebra generated by $X_0, \ldots, X_n$. The motivates us to define martingale in a more general way.

Definition 23.2 (Martingale (defined by filtration)) Let $\{\mathcal{F}_n\}_{n\geq 0}$ be a sequence of σ-algebras. We call such σ-algebra sequence a filtration if it satisfies

$$\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}_n \subseteq \mathcal{F}_{n+1} \subseteq \cdots.$$

Given a filtration $\{\mathcal{F}_n\}_{n\geq 0}$, let $\{Z_n\}_{n\geq 0}$ be a stochastic process that Z_n is $\mathcal{F}_n$-measurable for every $n \geq 0$. Then we say $\{Z_n\}_{n\geq 0}$ is a martingale w.r.t. $\{\mathcal{F}_n\}_{n\geq 0}$ if for every $n \geq 0$

$$\mathbf{E}\left[Z_{n+1} | \mathcal{F}_n\right] = Z_n.$$

[26]

Example 18 (1-D Random Walk) Consider a random walk on $\mathbb{Z}$ starting from 0. The probability to the left and the probability to the right are both $\frac{1}{2}$ at each step. Denote the action at the n-th step by a uniform random variable $X_n \in \{-1, +1\}$. Let $S_n = \sum_{k=0}^n X_k$. Then we can verify $\{S_n\}_{n\geq 0}$ is a martingale w.r.t. $\{X_n\}_{n\geq 0}$ (or w.r.t. $\{S_n\}_{n\geq 0}$) by noticing that

$$\mathbf{E}\left[S_{n+1} \,\middle|\, \overline{X}_{0,n}\right] = \mathbf{E}\left[S_n + X_{n+1} \,\middle|\, \overline{X}_{0,n}\right] = S_n + \mathbf{E}\left[X_{n+1} \,\middle|\, \overline{X}_{0,n}\right] = S_n.$$

More generally, if $\mathbf{E}\left[X_k \,\middle|\, \overline{X}_{0,n}\right] = \mu$, we define $Y_k = X_k - \mu$ and $S'_n \triangleq \sum_{k=0}^n Y_k = S_n - (n+1)\mu$. Then S'_n is a martingale w.r.t. $\{Y_n\}_{n\geq 0}$.

Example 19 Consider a sequence of random variables $\{X_n\}_{n\geq 0}$ where $\mathbf{E}\left[X_n \,\middle|\, \overline{X}_{0,n-1}\right] = 1$ for all $n \geq 1$. Let $P_n = \prod_{k=0}^n X_k$. Then we can verify $\{P_n\}_{n\geq 0}$ is a martingale w.r.t. $\{X_n\}_{n\geq 0}$ by verifying that

$$\mathbf{E}\left[P_{n+1} \,\middle|\, \overline{X}_{0,n}\right] = \mathbf{E}\left[P_n \cdot X_{n+1} \,\middle|\, \bar{X}_{0,n}\right] = P_n \cdot \mathbf{E}\left[X_{n+1} \,\middle|\, \overline{X}_{0,n}\right] = P_n.$$

Example 20 (Galton-Watson Process) Recall the Galton-Watson process we discussed in the last lecture. Suppose that all the individuals reproduce independently of each other and have the same offspring distribution. Let G_t be the number of individuals of the t-th generation. Each individual of generation t gives birth to a random number of children of generation $t+1$. Denote by $X_{t,k}$ the number of children of the k-th individual in the t-th generation. Assume that $X_{t,k}$ are i.i.d. and let $\mu \triangleq \mathbf{E}\left[X_{t,k}\right]$. Then we have $G_{t+1} = \sum_{k=1}^{G_t} X_{t,k}$. Thus,

$$\mathbf{E}\left[G_{t+1} \mid G_t\right] = \mathbf{E}\left[\sum_{k=1}^{G_t} X_{t,k} \,\middle|\, G_t\right] = G_t \cdot \mathbf{E}\left[X_{t,1}\right] = \mu G_t.$$

Define $M_t = \mu^{-t} G_t$. Then

$$\mathbf{E}\left[M_{t+1} \mid G_t\right] = \mu^{-t-1}\mathbf{E}\left[G_{t+1} \mid G_t\right] = \mu^{-t} G_t = M_t.$$

That is, $\{M_t\}_{t\geq 0}$ is a martingale w.r.t. $\{G_t\}_{t\geq 0}$.

Example 21 (Pólya's urn) Suppose there are some white balls and black balls in an urn. All of these balls are identical except the colors. Consider the following stochastic process: each round we pick a ball uniformly at random and observe its

[26] If $\mathbf{E}\left[Z_{n+1} | \mathcal{F}_n\right] \leq Z_n$ in Definition 23.2, we call $\{Z_n\}_{n\geq 0}$ a supermartingale w.r.t. $\{\mathcal{F}_n\}_{n\geq 0}$. Similarly, if $\mathbf{E}\left[Z_{n+1} | \mathcal{F}_n\right] \geq Z_n$, we call it a submartingale.

color; then we return the ball, and add an additional ball of the same color into the urn. We repeat the process, and our goal is to study the sequence of colors of the selected balls. [27]

W.l.o.g. assume that we start from only one white ball and one black ball in the urn, and the index of rounds starts from 2. Then after round n, there are exactly n balls in the urn. Let X_n be the number of black balls after round n, and $Z_n = \frac{X_n}{n}$ is the ratio of black balls after round n. Clearly $Z_2 = \frac{1}{2}$. Then we have

$$\begin{aligned} \mathbf{E}\left[Z_{n+1} \,\middle|\, \bar{X}_{2,n}\right] &= \frac{1}{n+1}\mathbf{E}\left[X_{n+1} \,\middle|\, \overline{X}_{2,n}\right] \\ &= \frac{1}{n+1}\left(Z_n(X_n+1) + (1-Z_n)X_n\right) = \frac{Z_n + X_n}{n+1} = Z_n. \end{aligned}$$

That is, $\{Z_n\}_{n\geq 2}$ is a martingale w.r.t. $\{X_n\}_{n\geq 2}$.

24 Optional Stopping Theorem

24.1 Stopping Time

If $\{Z_n\}_{n\geq 0}$ is a martingale w.r.t. $\{\mathcal{F}_n\}_{n\geq 0}$, $\mathbf{E}\left[Z_{n+1} \mid \mathcal{F}_n\right] = Z_n$. Take the expectation of both sides, we have

$$\mathbf{E}\left[Z_{n+1}\right] = \mathbf{E}\left[Z_n\right] = \cdots = \mathbf{E}\left[Z_1\right] = \mathbf{E}\left[Z_0\right].$$

That is, for any fixed $t \geq 0$, $\mathbf{E}\left[Z_t\right] = \mathbf{E}\left[Z_0\right]$. However, when t is a random variable (we denote it by τ), does $\mathbf{E}\left[Z_\tau\right] = \mathbf{E}\left[Z_0\right]$ still holds? The answer is obviously no in general. Consider the one-dimensional random walk with $Z_0 = 1$ and τ is the first time that $Z_t = 100$. We have $\mathbf{E}\left[Z_\tau\right] = 100 \neq \mathbf{E}\left[Z_0\right]$. To determine under which condition we could conclude $\mathbf{E}\left[Z_\tau\right] = \mathbf{E}\left[Z_0\right]$, lets formalize the notion of *stopping time* first.

Definition 24.1 (Stopping Time) Let $\tau \in \mathbb{N} \cup \{\infty\}$ be a random variable. We say τ is a stopping time defined on a filtration $\{\mathcal{F}_n\}_{n\geq 0}$ if for any $t \geq 0$, $\mathbb{1}\left[\tau \leq t\right]$ is $\mathcal{F}_t$-measurable.

[28]

Then we introduce the Optional Stopping Theorem (OST) to show the sufficient conditions that $\mathbf{E}\left[Z_\tau\right] = \mathbf{E}\left[Z_0\right]$.

24.2 Optional Stopping Theorem

Theorem 24.1 (Optional Stopping Theorem) Suppose that $\{X_t\}_{t\geq 0}$ is a martingale with respect to a filtration $\{\mathcal{F}_t\}_{t\geq 0}$ and τ is a stopping time with respect to the same filtration. Then $\mathbf{E}\left[X_\tau\right] = \mathbf{E}\left[X_0\right]$ if at least one of the following holds

1. τ is bounded almost surely, that is, $\exists n \in \mathbb{N}$ such that $\mathbf{Pr}\left[\tau \leq n\right] = 1$.

2. $\mathbf{Pr}\left[\tau < \infty\right] = 1$ and $\exists M \in \mathbb{N}$ such that $|X_t| \leq M$ for all $t \leq \tau$.

3. $\mathbf{E}\left[\tau\right] < \infty$ and $\exists c \in \mathbb{N}$ such that $\mathbf{E}\left[|X_{t+1} - X_t| |\mathcal{F}_t\right] \leq c$ for all $t \leq \tau$.

Before proving Theorem 24.1, we see some applications of OST.

Example 22 (Sex Ratio) Recall the problem of sex ratio we mentioned in the first lecture. Consider the following reproduction strategies

1. Every family keeps having children until they give birth to a boy.

2. Every family keeps having children until their sons are more than their daughters.

3. Every family keeps having children until their sons are more than their daughters or the number of kids is not less than 5.

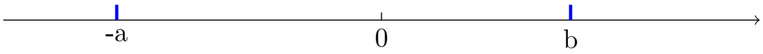
Assume that the natural birth sex ratio is uniform and every family only gives birth to a child at a time. Fix a family. Let $X_t = \{-1, +1\}$ denote whether the t-th child is a boy, and $Z_n = \sum_{t=1}^{n} X_t$ denote the number of boys more than girls. This process is a random walk on $\mathbb{Z}$ starting from 0. Let $\tau_1 = \min\{t \geq 0 : X_t = +1\}$, $\tau_2 = \min\{t \geq 0 : Z_t = 1\}$ and $\tau_3 = \min\{\tau_2, 5\}$. Obviously, τ_1, τ_2 and τ_3 are stopping times corresponding to the 3 strategies.

It is not hard to verify that τ_1 satisfies the third condition in Theorem 24.1 and τ_3 satisfies all the 3 conditions while τ_2 does not satisfy any conditions in Theorem 24.1. Thus, we have that the sex ratio is always 1: 1 using the first or third strategy. However, if the family adopts strategy 2, since boys are always more than girls when they stop having kids, the sex ratio is unbalanced.

[27] Example 21 shows that X_n does not have to be i.i.d..

[28] For a counter example, in a game, let $\tau' \triangleq$"the last time I win 5 in streak". Then τ is not a stopping time.

Example 23 (1-D Random Walk with Two Absorbing Barriers) Let $a, b > 0$ be two integers. A man starts the random walk from 0 and stops when he arrives at $-a$ or b. Let τ be the time when the man first reaches $-a$ or b. We want to compute the expected value of τ, that is, the average stopping time of the walk.



Let $X_n \in \{-1, +1\}$ be a uniform random variable, $Z_n = \sum_{k=1}^{n} X_k$ and $Z_0 = 0$. Then $\tau = \min\{t : Z_t = -a \text{ or } Z_t = b\}$. We know that $\{Z_t\}_{t\geq 1}$ is a martingale w.r.t. $\{X_t\}_{t\geq 0}$ and τ is a stopping time.

First we verify that $\mathbf{E}[Z_\tau] = \mathbf{E}[Z_0]$ by showing $\{Z_t\}$ satisfies the second condition in Theorem 24.1. Note that $|Z_n|$ is bounded. So, in order to apply OST, we should prove that $\mathbf{Pr}[\tau < \infty] = 1$. Since the probability of ending within the next $a + b$ steps is at least 2^{-a-b} no matter where the current position is, the random walk ends in finite steps with probability 1. Furthermore, if we divide the time into consecutive periods in this manner, in expected finite time, we can meet some period with the event happened, that is, $\mathbf{E}[\tau] < \infty$. So $\{Z_t\}$ also satisfies the third condition. It then follows that $\mathbf{E}[Z_\tau] = \mathbf{E}[Z_0] = 0$. That is

$$-a \cdot \mathbf{Pr}[\text{ending at } -a] + b \cdot (1 - \mathbf{Pr}[\text{ending at } -a]) = 0.$$

This yields $\mathbf{Pr}[\text{ending at } -a] = \frac{b}{a+b}$ and $\mathbf{Pr}[\text{ending at } b] = \frac{a}{a+b}$.

Then we define $\{Y_t\}_{t\geq 0}$ by $Y_t = Z_t^2 - t$. Note that

$$\begin{aligned}
\mathbf{E}\left[Y_{t+1} \,\middle|\, \overline{X}_{1,t}\right] &= \mathbf{E}\left[Z_{t+1}^2 - (t+1) \,\middle|\, \overline{X}_{1,t}\right] \\
&= \mathbf{E}\left[(Z_t + X_{t+1})^2 - (t+1) \,\middle|\, \overline{X}_{1,t}\right] \\
&= \mathbf{E}\left[Z_t^2 + X_{t+1}^2 + 2Z_t \cdot X_{t+1} - (t+1) \,\middle|\, \overline{X}_{1,t}\right] \\
&= Z_t^2 + 2Z_t\mathbf{E}\left[X_{t+1} \,\middle|\, \overline{X}_{1,t}\right] + \mathbf{E}\left[X_{t+1}^2 \,\middle|\, \overline{X}_{1,t}\right] - (t+1) \\
&= Z_t^2 + 0 + 1 - (t+1).
\end{aligned}$$

That is, $\{Y_t\}_{t\geq 0}$ is a martingale w.r.t. $\{X_t\}_{t\geq 1}$. Note that $\{Y_t\}$ satisfy the third condition of OST. Then $\mathbf{E}[Y_\tau] = \mathbf{E}[Y_0] = 0$. Thus,

$$\mathbf{E}\left[Z_\tau^2 - \tau\right] = 0 \Rightarrow \mathbf{E}[\tau] = \mathbf{E}\left[Z_\tau^2\right] = a^2 \cdot \frac{b}{a+b} + b^2 \cdot \frac{a}{a+b} = ab.$$

Example 24 (Pattern Matching) Suppose that there is a $\{H, T\}$-string P of length ℓ (H for "head" and T for "tail"). We flip a coin consecutively until the last ℓ results form exactly the same string as P. How many times do we flip the coin?

Note that if we flip the coin N times and observe the string S consisting of N results. No matter which pattern we choose, by the linearity of expectation, the expected number of occurrence [29] is

$$\mathbf{E}[\# \text{ of occurrence of } P \text{ in } S] = \sum_{i=1}^{n-\ell+1} \mathbf{E}\left[\mathbb{1}[S_{i,i+1,\ldots,i+\ell-1} = P]\right] = (n - l + 1) \cdot 2^{-\ell}.$$

However, if we would like to compute the first time that pattern P occurs, the pattern itself has an impact on the expected time. Intuitively, lets consider two patterns HT and HH. Assume that the first flipping result is H. Then we consider what happens if the second result fails. Suppose that the desired pattern is HT and H appears. Although we fail, we obtain an H. However, if the desired pattern is HH and the second flipping result is T, then we obtain nothing and the first two flips are a waste. So we should believe that the expected times of the first occurrence of HT is smaller than HH.

We now use the optional stopping theorem to solve this problem. Let $P = p_1p_2 \ldots p_\ell$. For every $n \geq 0$, assume that before $n+1$-th flipping there is a new gambler G_{n+1} coming with 1 unit of money to bet that the following ℓ result (i.e., the $n+1$-th to $n+\ell$-th results) are exactly the same as P. At the $n+k$-th flipping, G_{n+1} will bet that the result is p_k by an all in strategy, that is, if the $n+k$-th result is p_k then G_{n+1} will have twice as much money as before; otherwise they will lose all. Suppose that the patter P = HTHTH and the flipping results are HTHHTHTH. The following table shows the total money of each gambler after flipping.

Gambler	H	T	H	H	T	H	T	H	Money	
1	H	T	H	T					0	1→2→4→8→0
2		H							0	1→0
3			H	T					0	1→2→0
4				H	T	H	T	H	32	1→2→4→8→16→32
5					H				0	1→0
2						H	T	H	8	1→2→4→8
5							H		0	1→0
5								H	2	1→2

[29] That means the expected number of substrings exactly the same as P in the resulting string S.

Let X_t be the result of t-th flipping, $M_i(t)$ denote the money that G_i has after t-th flipping, and $Z_t \triangleq \sum_{i=1}^{t}(M_i(t)-1)$ be the total income of all gamblers after t-th flipping. It is easy to verify that $\{M_i(t)\}_{t\geq 0}$ is a martingale with respect to $\{X_t\}$ since

$$\mathbf{E}\left[M_i(t+1) \,\middle|\, \bar{X}_{0,t}\right] = \frac{1}{2}\cdot 2M_i(t) + \frac{1}{2}\cdot 0 = M_i(t).$$

Then by the linearity of expectation we conclude that $\{Z_t\}_{t\geq 0}$ is a martingale with respect to the flipping results $\{X_t\}$ since $\mathbf{E}[M_i(t)] = 0$. Let τ be the stopping time defined by the first time that some gambler wins, namely, the first time that P occurs in the flipping results. Applying Condition 2 of OST we obtain that $\mathbf{E}[Z_\tau] = \mathbf{E}[Z_0] = 0$. Sequentially we have $\mathbf{E}\left[\sum_{i=1}^{\tau} M_i(\tau) - \tau\right] = 0$ and $\mathbf{E}[\tau] = \sum_{i=1}^{\tau}\mathbf{E}[M_i(\tau)]$.

Note that $M_i(t) = 0$ for $i \leq \tau - \ell$ and $M_i(t) = 2^{\tau-i+1}\chi_{\tau-i+1}$ for $i > \tau - \ell$ where χ_j is defined by

$$\chi_j = \mathbb{1}[p_1 p_2 \dots p_j = p_{\ell-j+1}\dots p_{\ell-1}p_\ell].$$

Hence,

$$\mathbf{E}[\tau] = \sum_{i=\tau-\ell+1}^{\tau}\mathbf{E}[M_i(\tau)] = \sum_{i=1}^{\ell} 2^i \chi_i.$$

Recall the example of HH and HT. If P is HH, $\mathbf{E}[\tau] = 2 + 4 = 6$. If P is HT, $\mathbf{E}[\tau] = 4$. This confirms our hypothesis that $\mathbf{E}[\tau]$ for HH is larger than $\mathbf{E}[\tau]$ for HT.

25 Proof of OST

Theorem 25.1 (Optional Stopping Theorem (restated)) Suppose that $\{X_t\}_{t\geq 0}$ is a martingale with respect to a filtration $\{\mathcal{F}_t\}_{t\geq 0}$ and τ is a stopping time with respect to the same filtration. Then $\mathbf{E}[Z_\tau] = \mathbf{E}[Z_0]$ if at least one of the following holds

1. τ is bounded almost surely, that is, $\exists n \in \mathbb{N}$ such that $\mathbf{Pr}[\tau \leq n] = 1$.
2. $\mathbf{Pr}[\tau < \infty] = 1$ and $\exists M \in \mathbb{N}$ such that $|X_t| \leq M$ for all $t \leq \tau$.
3. $\mathbf{E}[\tau] < \infty$ and $\exists c \in \mathbb{N}$ such that $\mathbf{E}[|X_{t+1} - X_t||\mathcal{F}_t] \leq c$ for all $t \leq \tau$.

Proof. First we show that for every $n \in \mathbb{N}$, $\mathbf{E}\left[X_{\min\{n,\tau\}}\right] = \mathbf{E}[X_0]$. Define $Z_n = X_{\min\{n,\tau\}} = X_0 + \sum_{i=0}^{n-1}(X_{i+1} - X_i)\mathbb{1}[\tau > i]$. We verify that $\{Z_n\}_{n\geq 0}$ is a martingale. By definition

$$\mathbf{E}[Z_{n+1} \mid \mathcal{F}_n] = \mathbf{E}[Z_n + (X_{n+1} - X_n)\mathbb{1}[\tau > n] \mid \mathcal{F}_n] = Z_n + \mathbb{1}[\tau > n]\left(\mathbf{E}[X_{n+1} \mid \mathcal{F}_n] - X_n\right) = Z_n.$$

So we have $\mathbf{E}\left[X_{\min\{n,\tau\}}\right] = \mathbf{E}[Z_n] = \mathbf{E}[Z_0] = \mathbf{E}[X_0]$.

Therefore, this motivates us to decompose X_τ into two terms:

$$\forall n \in \mathbb{N},\; X_\tau = X_{\min\{n,\tau\}} + \mathbb{1}[\tau > n]\cdot(X_\tau - X_n).$$

Taking expectation and letting n tend to infinity, we obtain

$$\mathbf{E}[X_\tau] = \mathbf{E}[X_0] + \lim_{n\to\infty}\mathbf{E}[\mathbb{1}[\tau > n]\cdot(X_\tau - X_n)].$$

Therefore, we only need to verify that each of the three conditions in the statement can guarantee $\lim_{n\to\infty}\mathbf{E}[\mathbb{1}[\tau > n]\cdot(X_\tau - X_n)] = 0$.

1. If τ is bounded almost surely, then clearly $\mathbf{E}[\mathbb{1}[\tau > n]\cdot(X_\tau - X_n)] = 0$ for sufficiently large n.
2. In this case,

$$\begin{aligned}\mathbf{E}[\mathbb{1}[\tau > n]\cdot(X_\tau - X_n)] &\leq \mathbf{E}[\mathbb{1}[\tau > n]\cdot(|X_\tau| + |X_n|)] \\ &\leq 2M\cdot\mathbf{E}[\mathbb{1}[\tau > n]] \\ &= 2M\cdot\mathbf{Pr}[\tau > n] \to 0 \text{ as } n \to \infty.\end{aligned}$$

3. In order to apply our bounds on the gap between consecutive X_t, we write

$$\begin{aligned}\mathbb{1}[\tau > n]\cdot(X_\tau - X_n) &= \sum_{t=n}^{\tau-1}(X_{t+1} - X_t) \\ &\leq \sum_{t=n}^{\tau-1}|X_{t+1} - X_t| \\ &= \sum_{t=n}^{\infty}|X_{t+1} - X_t|\cdot\mathbb{1}[\tau > t].\end{aligned}$$

Taking expectation and applying the *Fubini's theorem*, we obtain

$$\begin{aligned}\mathbf{E}\left[\mathbb{1}[\tau > n] \cdot (X_\tau - X_n)\right] &\le \sum_{t=n}^{\infty} \mathbf{E}\left[|X_{t+1} - X_t| \cdot \mathbb{1}[\tau > t]\right] \\ &= \sum_{t=n}^{\infty} \mathbf{E}\left[\mathbf{E}\left[|X_{t+1} - X_t| \cdot \mathbb{1}[\tau > t] \mid \mathcal{F}_t\right]\right] \\ &= \sum_{t=n}^{\infty} \mathbf{E}\left[\mathbf{E}\left[|X_{t+1} - X_t| \mid \mathcal{F}_t\right] \cdot \mathbb{1}[\tau > t]\right] \\ &\le c \sum_{t=n}^{\infty} \mathbf{E}\left[\mathbb{1}[\tau > t]\right],\end{aligned}$$

where the last equality is due to the fact that $\mathbb{1}[\tau > t]$ is $\mathcal{F}_t$-measurable.

□

26 Hoeffding's Inequality

Recall the Chernoff bound we discussed in Lecture 2.

Theorem 26.1 (Chernoff Bound) . Let $X_1, \dots, X_n$ be independent random variables such that $X_i \sim Ber(p_i)$ for each $i = 1, 2, \dots, n$. Let $X = \sum_{i=1}^n X_i$ and denote $\mu \triangleq \mathbf{E}[X] = \sum_{i=1}^n p_i$, we have

$$\mathbf{Pr}\left[X \ge (1+\delta)\mu\right] \le \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^\mu \le \exp\left\{\left(-\frac{\delta^2}{3}\mu\right)\right\}.$$

If $0 < \delta < 1$, then we have

$$\mathbf{Pr}\left[X \le (1-\delta)\mu\right] \le \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^\mu \le \exp\left\{\left(-\frac{\delta^2}{2}\mu\right)\right\}.$$

One of annoying restrictions of Chernoff bound is that each X_i needs to be a Bernoulli random variable. We first relax this requirement by introducing Hoeffding's inequality which allows X_i to follow any distribution, provided its value is almost surely bounded.

Theorem 26.2 (Hoeffding's Inequality) Let $X_1, \dots, X_n$ be independent random variables where each $X_i \in [a_i, b_i]$ for certain $a_i \le b_i$ with probability 1. Let $X = \sum_{i=1}^n X_i$ and $\mu \triangleq \mathbf{E}[X] = \sum_{i=1}^n \mathbf{E}[X_i]$, then

$$\mathbf{Pr}\left[|X - \mu| \ge t\right] \le 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

for all $t \ge 0$.

It is instructive to compare Hoeffding and Chernoff when X_i's are independent Bernoulli variables. Formally, let $X_1, \dots, X_n$ be i.i.d. random variables where $X_i \sim \mathtt{Ber}(p)$ for all $i = 1, \dots, n$. Set $X = \sum_{i=1}^n X_i$ and denote $\mathbf{E}[X] = np$ by μ. By Hoeffding's inequality, we have

$$\mathbf{Pr}\left[|X - \mu| \ge t\right] \le 2\exp\left(-\frac{2t^2}{n}\right).$$

By Chernoff Bound, we have

$$\mathbf{Pr}\left[|X - \mu| \ge t\right] \le 2\exp\left(-\frac{t^2}{3pn}\right).$$

Comparing the exponent, it is easy to see that for $p > 1/6$, Hoeffding's inequality is tighter up to a certain constant factor. However, for smaller p, Chernoff bound is significantly better than Hoeffding's inequality.

Before proving Theorem 26.2 in Section 28, we see a practical application of Hoeffding's inequality.

Example 25 (Meal Delivery) During the quarantine of our campus, the professors deliver meals for students using their private cars or trikes. Then a practical problem is how to estimate the amount of meals on a trike conveniently (See the news).

Assume there are n boxes of meal on the trike ($n \ge 200$ and is unknown for us). Let X_i be the weight of the i-th box of meal. Assume that $X_1, X_2, \dots, X_n$ are i.i.d. random variables, each $X_i \in [250, 350]$ (unit: gram) and $\mu = \mathbf{E}[X_i] = 300$. Let S be the total weight of the meal boxes on the trike, that is, $S = \sum_{i=1}^n X_i$. We can weigh the meal boxes and use $\hat{n} = \frac{S}{\mu}$ as an estimator for n. Now we compute how accurate this estimator is.

Let $\delta \ge 0$ be a constant. By Hoeffding's inequality,

$$\mathbf{Pr}\left[|\hat{n} - n| > \delta n\right] = \mathbf{Pr}\left[|S - \mu n| > \delta \mu n\right] \le 2\exp\left\{-\frac{2\delta^2\mu^2 n^2}{\sum_{i=1}^n (350 - 250)^2}\right\}. \qquad (14)$$

Plugging $\mu = 300$, $\delta = 0.05$ and $n \geq 200$ into Equation (14), by direct calculation, we have

$$\mathbf{Pr}\left[\hat{n} \in [0.95n, 1.05n]\right] \geq 1 - 2.4682 \times 10^{-4}.$$

27 Concentration on Martingale

We consider the balls-in-a-bag problem. There are g green balls and r red balls in a bag. These balls are the all same except for the color. We want to estimate the ratio $\frac{r}{r+g}$ by drawing balls. There are two scenarios.

- Draw balls with replacement. Let $X_i = \mathbf{1}[\text{the } i\text{-th ball is red}]$. Let $X = \sum_{i=1}^{n} X_i$. Then clearly each $X_i \sim \mathrm{Ber}\left(\frac{r}{r+g}\right)$ and $\mathbf{E}[X] = n \cdot \frac{r}{r+b}$.

  Since all X_i's are independent, we can directly apply Hoeffding's inequality and obtain

  $$\mathbf{Pr}\left[|X - \mathbf{E}[X]| \geq t\right] \leq 2\exp\left(-\frac{2t^2}{n}\right).$$

- Draw balls without replacement. Again we let $Y_i = \mathbf{1}[\text{the } i\text{-th ball is red}]$, then unlike the case of drawing with replacement, variables in $\{Y_i\}$ are dependent. Let $Y = \sum_{i=1}^{n} Y_i$. We first calculate $\mathbf{E}[Y]$.

  For every $i \geq 1$, $\mathbf{E}[Y_i]$ is the probability that the i-th draw is a red ball. Note that drawing without replacement is equivalent to first drawing a uniform permutation of $r + g$ balls and drawing each ball one by one in that order. Therefore, the probabilty of $Y_i = 1$ is $\frac{r \cdot (r+g-1)!}{(r+g)!} = \frac{r}{r+g}$. So we have $\mathbf{E}[Y] = n \cdot \frac{r}{r+g}$.

  However, since $\{Y_i\}$ are dependent, we cannot apply Hoeffding's inequality directly. This motivate us to generalize it by removing the requirement of independence.

27.1 Azuma-Hoeffding's Inequality

Theorem 27.1 (Azuma-Hoeffding's Inequality) If $\{S_n\}_{n\geq 0}$ where $S_k = \sum_{i=0}^{k} X_i$ is a martingale w.r.t. $\{X_n\}_{n\geq 0}$ with $X_i \in [a_i, b_i]$ with probability 1, then

$$\mathbf{Pr}\left[|S_n - S_0| \geq t\right] \leq 2\exp\left(-\frac{2t^2}{\sum\limits_{i=1}^{n}(b_i - a_i)^2}\right).$$

The proof of this theorem is in Section 28. The requirement of martingale in Theorem 27.1 seems to be even harder to satisfy than the requirement of independence. However, in many cases, we can construct a doob martingale to apply the Azuma-Hoeffding's inequality.

Definition 27.1 (Doob Martingale, Doob Sequence) Let $X_1, \ldots, X_n$ be a sequence of (unnecessarily independent) random variables and $f(\overline{X}_{1,n}) = f(X_1, \ldots, X_n) \in \mathbb{R}$ be a function. For $i \geq 0$, Let $Z_i \triangleq \mathbf{E}\left[f(\overline{X}_{1,n}) \,\middle|\, \overline{X}_{1,i}\right]$. Then we call $\{Z_n\}_{n\geq 0}$ a Doob martingale or a Doob sequence.

It is easy to verify that $\{Z_n\}_{n\geq 0}$ in Definition 27.1 is indeed a martingale w.r.t. $\{X_n\}$ by seeing

$$\mathbf{E}\left[Z_i \,\middle|\, \overline{X}_{1,i-1}\right] = \mathbf{E}\left[\mathbf{E}[f(\overline{X}_{1,n}) \,\middle|\, \overline{X}_{1,i}] \,\middle|\, \overline{X}_{1,i-1}\right] = \mathbf{E}\left[f(\overline{X}_{1,n}) \,\middle|\, \overline{X}_{1,i-1}\right] = Z_{i-1}.$$

Let $\mathcal{F} = \sigma(\overline{X}_{1,i})$. We can see that Z_i is $\mathcal{F}_i$ measurable by definition. Moreover, we know that $Z_0 = \mathbf{E}\left[f(\overline{X}_{1,n})\right]$ and $Z_n = f(\overline{X}_{1,n})$.

Recall the balls-in-a-bag problem we discussed above. Define $f : \mathbb{R}^n \to \mathbb{R}$ by letting $f(y_1, y_2, \ldots, y_n) = \sum_{i=1}^{n} y_i$. Then in the drawing without replacement scenario, $Y = \sum_{i=1}^{n} Y_i = f(Y_1, Y_2, \ldots, Y_n)$. Now we construct the Doob martingale for f.

Let $Z_i = \mathbf{E}\left[f(\overline{Y}_{1,n}) \,\middle|\, \overline{Y}_{1,i}\right]$. We know that $Z_0 = \mathbf{E}\left[f(\overline{Y}_{1,n})\right] = \mathbf{E}[Y] = n \cdot \frac{r}{r+g}$ and $Z_n = f(\overline{Y}_{1,n})$. Let $X_0 \triangleq Z_0$ and $X_i \triangleq Z_i - Z_{i-1}$ for $i \geq 1$. Then Z_n can be represented as $Z_n = \sum_{i=0}^{n} X_i$. In order to apply Azuma-Hoeffding, we need to bound the *width* of the martingale, i.e. $|X_i| = |Z_i - Z_{i-1}|$.

By definition,

$$Z_i - Z_{i-1} = \mathbf{E}\left[f(\overline{Y}_{1,n}) \,\middle|\, \overline{Y}_{1,i}\right] - \mathbf{E}\left[f(\overline{Y}_{1,n}) \,\middle|\, \overline{Y}_{1,i-1}\right].$$

If we use S_i to denote the number of red balls among the first i balls, namely $S_i = \sum_{j=1}^{i} Y_j$, then

$$\mathbf{E}\left[f(\overline{Y}_{1,n}) \,\middle|\, \overline{Y}_{1,i}\right] = \mathbf{E}\left[f(\overline{Y}_{1,n}) \,\middle|\, S_i\right] = S_i + (n - i) \cdot \frac{r - S_i}{g + r - i}.$$

Therefore $S_i = S_{i-1} + Y_i$ and

$$\begin{aligned} Z_i - Z_{i-1} &= \left(S_i + (n-i) \cdot \frac{r - S_i}{g + r - i}\right) - \left(S_{i-1} + (n - i + 1) \cdot \frac{r - S_{i-1}}{g + r - i + 1}\right) \\ &= \frac{g + r - n}{g + r - i} \left(Y_i + \frac{S_{i-1} - r}{g + r - i + 1}\right). \end{aligned}$$

Note that $r \geq S_{i-1}$ and $g \geq (i-1) - S_{i-1}$, we have

$$\begin{aligned} Z_i - Z_{i-1} &\leq \frac{g + r - n}{g + r - i} \left(1 + \frac{S_{i-1} - r}{g + r - i + 1}\right) \leq \frac{g + r - n}{g + r - i} \leq 1, \\ Z_i - Z_{i-1} &\geq \frac{g + r - n}{g + r - i} \left(\frac{S_{i-1} - r}{g + r - i + 1}\right) \geq -\frac{g + r - n}{g + r - i} \geq -1. \end{aligned}$$

Therefore $-1 \leq X_i \leq 1$ and we can apply Azuma-Hoeffding to $Z_n - Z_0$ to obtain

$$\mathbf{Pr}\left[|Y - \mathbf{E}[Y]| \geq t\right] \leq 2\exp\left(-\frac{t^2}{2n}\right).$$

27.2 McDiarmids Inequality

The Doob sequence we used in the balls-in-a-bag example is a very powerful and general tool to obtain concentration bounds. For a model defined by n random variables $X_1, \ldots, X_n$ and any quantity $f(X_1, \ldots, X_n)$ that we want to estimate, we can apply the Azuma-Hoeffding inequality to the Doob sequence of f. As shown in the previous example, the quality of the bound relies on the *width* of the martingale, that is, the volume of $|Z_i - Z_{i-1}|$. To determine the width of each $|Z_i - Z_{i-1}|$ is relatively easy if the function f and the variables $\{X_i\}_{1 \leq i \leq n}$ enjoy certain nice properties.

Definition 27.2 (c-Lipschitz Function) A function $f(x_1, \cdots, x_n)$ satisfies c-Lipschitz condition if

$$\begin{aligned} &\forall i \in [n], \forall x_1, \cdots, x_n, \forall y_i : \\ &|f(x_1, \cdots, x_i, \cdots, x_n) - f(x_1, \cdots, y_i, \cdots, x_n)| \leq c. \end{aligned}$$

The McDiarmid's inequality is the application of Azuma-Hoeffding inequality to Lipschitz f and independent $\{X_i\}$.

Theorem 27.2 (McDiarmid's Inequality) Let f be a function on n variables satisfying c-Lipschitz condition and $X_1, \cdots, X_n$ be n independent variables. Then we have

$$\mathbf{Pr}\left[|f(X_1, \cdots, X_n) - \mathbf{E}[f(X_1, \cdots, X_n)]| \geq t\right] \leq 2e^{-\frac{2t^2}{nc^2}}.$$

Proof. We use f and $\{X_i\}_{i \geq 1}$ to define a Doob martingale $\{Z_i\}_{i \geq 1}$:

$$\forall i : Z_i = \mathbf{E}\left[f(\overline{X}_{1,n}) \,\middle|\, \overline{X}_{1,i}\right].$$

Then

$$Z_i - Z_{i-1} = \mathbf{E}\left[f(\overline{X}_{1,n}) \,\middle|\, \overline{X}_{1,i}\right] - \mathbf{E}\left[f(\overline{X}_{1,n}) \,\middle|\, \overline{X}_{1,i-1}\right].$$

Next we try to determine the width of $Z_i - Z_{i-1}$. Clearly

$$Z_i - Z_{i-1} \geq \inf_x \left\{\mathbf{E}\left[f(\overline{X}_{1,n}) \,\middle|\, \overline{X}_{1,i-1}, X_i = x\right] - \mathbf{E}\left[f(\overline{X}_{1,n}) \,\middle|\, \overline{X}_{1,i-1}\right]\right\},$$

and

$$Z_i - Z_{i-1} \leq \sup_y \left\{\mathbf{E}\left[f(\overline{X}_{1,n}) \,\middle|\, \overline{X}_{1,i-1}, X_i = y\right] - \mathbf{E}\left[f(\overline{X}_{1,n}) \,\middle|\, \overline{X}_{1,i-1}\right]\right\}.$$

The gap between the upper bound and the lower bound is

$$\sup_{x,y} \left\{\mathbf{E}\left[f(\overline{X}_{1,n}) \,\middle|\, \overline{X}_{1,i-1}, X_i = y\right] - \mathbf{E}\left[f(\overline{X}_{1,n}) \,\middle|\, \overline{X}_{1,i-1}, X_i = x\right]\right\}.$$

For every x, y and $\sigma_1,\dots,\sigma_{i-1}$,

$$\begin{aligned}
&\mathbf{E}\left[f(\overline{X}_{1,n}) \,\middle|\, \bigwedge_{1\le j\le i-1} X_j=\sigma_j, X_i=y\right] - \mathbf{E}\left[f(\overline{X}_{1,n}) \,\middle|\, \bigwedge_{1\le j\le i-1} X_j=\sigma_j, X_i=x\right]\\
=&\sum_{\sigma_{i+1},\dots,\sigma_n}\left(\mathbf{Pr}\left[\bigwedge_{i+1\le j\le n} X_j=\sigma_j \,\middle|\, \bigwedge_{1\le j\le i-1} X_j=\sigma_j, X_i=y\right]\cdot f(\sigma_1,\dots,\sigma_{i-1},y,\sigma_{i+1},\dots,\sigma_n)\right.\\
&\left.\quad-\mathbf{Pr}\left[\bigwedge_{i+1\le j\le n} X_j=\sigma_j \,\middle|\, \bigwedge_{1\le j\le i-1} X_j=\sigma_j, X_i=x\right]\cdot f(\sigma_1,\dots,\sigma_{i-1},x,\sigma_{i+1},\dots,\sigma_n)\right)\\
\overset{(\heartsuit)}{=}&\sum_{\sigma_{i+1},\dots,\sigma_n}\mathbf{Pr}\left[\bigwedge_{i+1\le j\le n} X_j=\sigma_j\right]\cdot\left(f(\sigma_1,\dots,\sigma_{i-1},y,\sigma_{i+1},\dots,\sigma_n)-f(\sigma_1,\dots,\sigma_{i-1},x,\sigma_{i+1},\dots,\sigma_n)\right)\\
\overset{(\clubsuit)}{\le}&\ c.
\end{aligned}$$

where ($\heartsuit$) uses independence of $\{X_i\}$ and ($\clubsuit$) uses the c-Lipsichitz property of f.

Applying Azuma-Hoeffding, we have

$$\mathbf{Pr}\left[|Z_n-Z_0|\ge t\right]=\mathbf{Pr}\left[|f(X_1,\cdots,X_n)-\mathbf{E}\left[f(X_1,\cdots,X_n)\right]|\ge t\right]\le 2e^{-\frac{2t^2}{nc^2}}.$$

□

Then we examine two applications of McDiarmid's inequality.

Example 26 (Pattern matching) Let $P\in\{0,1\}^k$ be a fixed string. For a random string $X\in\{0,1\}^n$, what is the expected number of occurrences of P in X?

The expectation of occurrence times can be easily calculated using the linearity of expectations. We define n independent random variables $X_1,\cdots,X_n$, where X_i denotes i-th character of X. Let $Y=f(X_1,\cdots,X_n)$ be the number of occurrences of P in X. Note that there are at most $n-k+1$ occurrences of P in X, and we can enumerate the first position of each occurrence. By the linearity of expectation, we have

$$\mathbf{E}\left[f\right]=\frac{n-k+1}{2^k}.$$

We can then use McDarmid's inequality to show that f is well-concentrated. To see this, we note that variables in $\{X_i\}$ are independent and the function f is k-Lipschitz: If we change one bit of X, the number of occurrences changes at most k. Therefore

$$\mathbf{Pr}\left[|Z_n-Z_0|\ge t\right]=\mathbf{Pr}\left[|f-\mathbf{E}\left[f\right]|\ge t\right]\le 2e^{-\frac{2t^2}{nk^2}}.$$

Another application of McDiarmid's Inequality is to establish the concentration of chromatic number for Erdős-Rényi random graphs $\mathcal{G}(n,p)$.

Example 27 (Chromatic Number of $\mathcal{G}(n,p)$) Recall the notation $\mathcal{G}(n,p)$ specifies a distribution over all undirected simple graphs with n vertices. In the model, each of the $\binom{n}{2}$ possible edges exists with probability p independently.

For a graph $G\sim\mathcal{G}(n,p)$, we use $\chi(G)$ to denote its chromatic number, the minimum number q so that G can be properly colored using q colors. There are different ways to represent G using random variables.

The most natural way is to introduce a variable X_e for every pair of vertices $e=\{u,v\}\subseteq V$ where $X_e=\mathbf{1}[\text{the edge } e \text{ exists in } G]$. Then $\{X_e\}$ are independent and the chromatic number can be written as a function $\chi(X_{e_1},X_{e_2},\dots,X_{e_{\binom{n}{2}}})$. It is easy to see that χ is 1-Lipschitz as removing to adding one edge can only change the chromatic number by at most one. So by McDarmid's inequality, we have

$$\mathbf{Pr}\left[|\chi-\mathbf{E}\left[\chi\right]|\ge t\right]\le 2e^{-2t^2\binom{n}{2}^{-1}}.$$

However, this bound is not satisfactory as we need to set $t=\Theta(n)$ in order to upper bound the RHS by a constant.

We can encode the graph G in a more efficient way while reserving the Lipschitz and the independence property. Suppose the vertex set of G is $\{v_1,\dots,v_n\}$. We define n random variables $Y_1,\cdots,Y_n$, where Y_i encodes the edges between v_i and $\{v_1,\cdots,v_{i-1}\}$. Once $Y_1,\cdots,Y_n$ are given, the graph is known, so the chromatic number can be written as a function $\chi(Y_1,\dots,Y_n)$. Since Y_i only involves the connections between v_i and $v_1,\cdots,v_{i-1}$, the n variables are independent.

It is also easy to see that if X_i changes, the chromatic number changes at most one. Hence χ is 1-Lipschitz as well. Applying McDiarmid's inequality we have

$$\mathbf{Pr}\left[|\chi-\mathbf{E}\left[\chi\right]|\ge t\right]\le 2e^{-\frac{2t^2}{n}}.$$

In this way, we only need $t=\Theta(\sqrt{n})$ to bound the RHS.

28 Proof

28.1 Proof of Theorem 26.2

First, we prove the following Hoeffding's lemma which will be the main technical ingredient to prove the inequality.

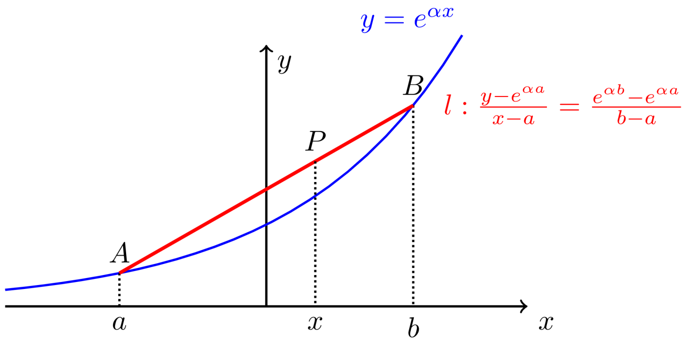
Lemma 28.1 Let X be a random variable with $\mathbf{E}[X] = 0$ and $X \in [a, b]$. Then it holds that

$$\mathbf{E}\left[e^{\alpha X}\right] \le \exp\left(\frac{\alpha^2(b-a)^2}{8}\right) \quad \text{for all } \alpha \in \mathbb{R}.$$

Proof.

We first find a linear function to upper bound $e^{\alpha x}$ so that we could apply the linearity of expectation to bound $\mathbf{E}\left[e^{\alpha X}\right]$. By the convexity of the exponential function and as illustrated in the figure below, we have

$$e^{\alpha x} \le \frac{e^{\alpha b} - e^{\alpha a}}{b-a}(x-a) + e^{\alpha a}, \quad \text{for all } a \le x \le b.$$


$y = e^{\alpha x}$
y
B
$l : \frac{y-e^{\alpha a}}{x-a} = \frac{e^{\alpha b}-e^{\alpha a}}{b-a}$
P
A
a
x
b
x

Thus,

$$\begin{aligned}
\mathbf{E}\left[e^{\alpha x}\right] &\le \frac{e^{\alpha b} - e^{\alpha a}}{b-a}(-a) + e^{\alpha a} = \frac{-a}{b-a}e^{\alpha b} + \frac{b}{b-a}e^{\alpha a} \\
&= e^{\alpha a}\left(\frac{b}{b-a} - \frac{a}{b-a}e^{\alpha(b-a)}\right) \\
&= e^{-\theta t}(1 - \theta + \theta e^{t}) \qquad \left(\theta = -\frac{a}{b-a}, t = \alpha(b-a)\right) \\
&\triangleq e^{g(t)},
\end{aligned}$$

where

$$g(t) = -\theta t + \log\left(1 - \theta + \theta e^{t}\right).$$

By Taylor's theorem, for every real t there exists a δ between 0 and t such that,

$$g(t) = g(0) + tg'(0) + \frac{1}{2}g''(\delta)t^2$$

Note that,

$$\begin{aligned}
g(0) &= 0; \\
g'(0) &= -\theta + \left.\frac{\theta e^{t}}{1 - \theta + \theta e^{t}}\right|_{t=0} \\
&= 0; \\
g''(\delta) &= \frac{\theta e^{t}(1 - \theta + \theta e^{t}) - \theta e^{t}}{(1 - \theta + \theta e^{t})^2} \\
&= \frac{(1-\theta)\theta e^{t}}{(1 - \theta + \theta e^{t})^2} \\
&= \frac{(1-\theta)\theta}{\theta^2 z + 2(1-\theta)\theta + \frac{(1-\theta)^2}{z}} && (z = e^{t}) \\
&\le \frac{(1-\theta)\theta}{2\theta(1-\theta) + 2(1-\theta)\theta} && (z > 0) \\
&= \frac{1}{4}.
\end{aligned}$$

Thus

$$g(t) \le 0 + t \cdot 0 + \frac{1}{2}t^2 \cdot \frac{1}{4} = \frac{1}{8}t^2 = \frac{1}{8}\alpha^2(b-a)^2.$$

Therefore, $\mathbf{E}\left[e^{\alpha x}\right] \le \exp\left(\frac{\alpha^2(b-a)^2}{8}\right)$ holds. □

Armed with Hoeffding's lemma, it is routine to prove Hoeffding's inequality.

Proof. [Proof of Theorem 26.2]

First note that we can assume $\mathbf{E}[X_i] = 0$ and therefore $\mu = 0$ (if not so, replace X_i by $X_i - \mathbf{E}[X_i]$). By symmetry, we only need to prove that $\mathbf{Pr}[X \geq t] \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$. Since

$$\mathbf{Pr}[X \geq t] \overset{\alpha \geq 0}{=} \mathbf{Pr}\left[e^{\alpha X} \geq e^{\alpha t}\right] \leq \frac{\mathbf{E}\left[e^{\alpha X}\right]}{e^{\alpha t}}$$

and

$$\mathbf{E}\left[e^{\alpha X}\right] = \mathbf{E}\left[e^{\alpha \sum_{i=1}^n X_i}\right] = \prod_{i=1}^n \mathbf{E}\left[e^{\alpha X_i}\right],$$

applying Hoeffding's lemma for each $\mathbf{E}\left[e^{\alpha X_i}\right]$ yields

$$\mathbf{E}\left[e^{\alpha X_i}\right] \leq \exp\left(\frac{\alpha^2 (b_i - a_i)^2}{8}\right).$$

Let $\alpha = \frac{4t}{\sum_{i=1}^n (b_i - a_i)^2}$, we have,

$$\begin{aligned} \mathbf{Pr}[X \geq t] \leq \frac{\prod_{i=1}^n \mathbf{E}\left[e^{\alpha X_i}\right]}{e^{\alpha t}} &\leq \exp\left(-\alpha t + \frac{\alpha^2}{8} \sum_{i=1}^n (b_i - a_i)^2\right) \\ &= \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right). \end{aligned}$$

□

28.2 Proof of Theorem 27.1

Now we will sketch a proof of the Azuma-Hoeffding, which is quite similar to our proof of the Hoeffding inequality.

Proof. [Proof of Theorem 27.1]

Recall when we were trying to prove the Hoeffding inequality, the most difficult part is to estimate the term

$$\mathbf{E}\left[e^{\alpha S_n}\right] = \mathbf{E}\left[\prod_{i=1}^n e^{\alpha X_i}\right].$$

In the case of Azuma-Hoeffding, we can use the property of martingales instead of independence to obtain a bound of this term. To see this, we have

$$\begin{aligned} \mathbf{E}\left[\prod_{i=1}^n e^{\alpha X_i}\right] &= \mathbf{E}\left[\mathbf{E}\left[\prod_{i=1}^n e^{\alpha X_i} \middle| \overline{X}_{n-1}\right]\right] \\ &= \mathbf{E}\left[\prod_{i=1}^{n-1} e^{\alpha X_i} \mathbf{E}\left[e^{\alpha X_n} \middle| \overline{X}_{n-1}\right]\right]. \end{aligned}$$

The bounds then follows by an induction argument and a conditional expectation version of Hoeffding lemma:

$$\mathbf{E}\left[e^{\alpha X_n} \middle| \overline{X}_{n-1}\right] \leq e^{-\frac{\alpha (b_i - a_i)^2}{8}}.$$

The proof is almost the same as our proof of Hoeffding lemma in the last lecture. □

29 Poisson Distribution

Example 28 Suppose that there exists a restaurant. The number of customers in the past five days are: 100, 120, 80, 75 and 110. To prepare ingredients in the right amount, we want to estimate the number of tomorrow's customers based on the information of the past several days. A natural idea is to use the average number (e.g., 97 in our instance) of the past. However, there are three out of the first five days that the restaurant do not prepare sufficient food if they adopt this in practice.

To analyze the distribution of the number of customers, we make some assumptions first. Assume that there are n isometric slots in a day. Every slot is sufficiently small s.t. at most one customer comes into the restaurant in a slot. The event "there is a customer coming in a slot" happens w.p. p and slots are independent of each other.

Let $X_i \triangleq \mathbf{1}[\text{there is a customer coming in the } i\text{-th slot}]$ for $i \in [n]$. Then we know $X_i \sim \mathrm{Ber}(p)$ and X_i's are mutually independent. Let $Z_n = \sum_{i=1}^n X_i$ and $\lambda = \mathbf{E}[Z_n] = pn$. Now lets compute the distribution of the number of customers Z_n. For any constant $k \in \mathbb{N}$,

$$\begin{aligned}
\mathbf{Pr}[Z_n = k] &= \tbinom{n}{k} p^k (1-p)^{n-k} \\
&= \frac{n(n-1)\cdots(n-k+1)}{k!} \cdot \left(\frac{\lambda}{n}\right)^k \cdot \left(1 - \frac{\lambda}{n}\right)^{n-k} \\
&= \frac{n(n-1)\cdots(n-k+1)}{n^k} \cdot \frac{\lambda^k}{k!} \cdot \left(1 - \frac{\lambda}{n}\right)^n \cdot \left(1 - \frac{\lambda}{n}\right)^{-k}. \qquad (15)
\end{aligned}$$

Note that λ and k are constants. Thus, when $n \to \infty$, Equation (15) equals to $\frac{\lambda^k}{k!} e^{-\lambda}$ and Z_n follows Poisson distribution with mean λ.

Definition 29.1 (Poisson Distribution) We say a random variable X follows Poisson distribution with mean λ or $X \sim Pois(\lambda)$, if for any $k \in \mathbb{Z}$,

$$\mathbf{Pr}[X = k] = \begin{cases} \frac{\lambda^k}{k!} e^{-\lambda} & if\ k \geq 0, \\ 0 & if\ k < 0. \end{cases}$$

Since we get the distribution of Z_n by taking the limit, we need to verify that it is a distribution and it's mean is indeed λ:

- We have $\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = 1$. Thus it is indeed a distribution.
- Since
$$\mathbf{E}[Z_n] = \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} = \lambda \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} e^{-\lambda} = \lambda \sum_{k=0}^{\infty} \frac{\lambda^k}{(k)!} e^{-\lambda} = \lambda,$$
the expectation of Z_n indeed equals to λ.

Then what is the distribution of two days customers? Let's examine the following property of Poisson distributions.

Proposition 29.1 Suppose $X_1 \sim Pois(\lambda_1)$ and $X_2 \sim Pois(\lambda_2)$ are two independent random variables. Then

$$X_1 + X_2 \sim Pois(\lambda_1 + \lambda_2).$$

Proof. For $n \geq 0$,

$$\begin{aligned}
\mathbf{Pr}[X_1 + X_2 = n] &= \sum_{m=0}^{n} \mathbf{Pr}[X_1 = m] \cdot \mathbf{Pr}[X_2 = n - m] \\
&= \sum_{m=0}^{n} \frac{\lambda_1^m}{m!} e^{-\lambda_1} \cdot \frac{\lambda_2^{n-m}}{(n-m)!} e^{-\lambda_2} \\
&= e^{-(\lambda_1 + \lambda_2)} \cdot \sum_{m=0}^{n} \binom{n}{m} \frac{\lambda_1^m \lambda_2^{n-m}}{n!} \\
&= e^{-(\lambda_1 + \lambda_2)} \frac{(\lambda_1 + \lambda_2)^n}{n!}
\end{aligned}$$

□

It is easy to extend the proposition to a sequence of independent Poissons and yield the following corollary.

Corollary 29.1 Suppose that $X_1, X_2, \ldots, X_n$ are n mutually independent random variables where $X_i \sim Pois(\lambda_i)$. Then

$$\sum_{i=1}^{n} X_i \sim Pois(\sum_{i=1}^{n} \lambda_i).$$

30 Poisson Process

30.1 Definition of Poisson Process

If we consider a period of time rather than a single day, e.g., from day t_1 to day t_2, then the number of customers follows $\mathrm{Pois}\left((t_2 - t_2)\lambda\right)$. Note that the time can be continuous. Thus, we introduce the notion of Poisson process.

Definition 30.1 A Poisson process $\{N(s) : s \geq 0\}$ with rate λ is a stochastic process that

1. $N(0) = 0$;
2. $\forall t, s \geq 0,\ N(t+s) - N(s) \sim Pois(\lambda t)$;
3. $\forall t_0 \leq t_1 \leq \cdots \leq t_n,\ N(t_1) - N(t_0),\ N(t_2) - N(t_1),\ \cdots,\ N(t_n) - N(t_{n-1})$ are mutually independent.

In fact, we can view the Poisson process in another way by considering the time gaps between arrivals. To see this, we first recall the exponential distribution.

Definition 30.2 The probability density function of the exponential distribution with rate $\lambda > 0$ is

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

The corresponding cumulative distribution function is

$$F(t) = \int_{-\infty}^{t} f(x)\,dx = \int_{0}^{t} \lambda e^{-\lambda x}\,dx = 1 - e^{-\lambda t}.$$

Then the following proposition gives another characterization of the Poisson process. [30]

Proposition 30.1 Suppose that $\tau_1, \tau_2, \ldots, \tau_n$ is a sequence of independent random variables that $\tau_i \sim Exp(\lambda)$. Let $T_n = \sum_{i=1}^{n} \tau_i$. For $s \geq 0$, let $N(s) = \max\{n \mid T_n \leq s\}$. Then $N(s)$ is a Poisson process with rate λ.

Before proving this proposition, we discuss some properties of the exponential distribution.

30.2 Properties of Exponential Distribution

Proposition 30.2 Let $X \sim Exp(\lambda)$. Then $\mathbf{E}[X] = \frac{1}{\lambda}$. [a]

[a] Since λ is the arriving rate, we can imagine that the average time between arrivals $\mathbf{E}[\tau_i]$ is the reciprocal of λ. This gives an intuition of Proposition 30.2.

Proof.

$$\begin{aligned} \mathbf{E}[X] &= \int_0^\infty t \cdot \lambda e^{-\lambda t} dt = \left(-te^{-\lambda t}\right)\Big|_0^\infty + \int_0^\infty e^{-\lambda t} dt \\ &= -\frac{1}{\lambda} e^{-\lambda t}\Big|_0^\infty = \frac{1}{\lambda}. \end{aligned}$$

□

Proposition 30.3 Let $X \sim Exp(\lambda)$. Then $\mathbf{Var}[X] = \frac{1}{\lambda^2}$.

Proof. Note that

$$\mathbf{Var}[X] = \mathbf{E}\left[X^2\right] - \mathbf{E}[X]^2 = \mathbf{E}\left[X^2\right] - \frac{1}{\lambda^2}.$$

And

$$\begin{aligned} \mathbf{E}\left[X^2\right] &= \int_0^\infty t^2 \cdot \lambda e^{-\lambda t} dt = \left(-t^2 e^{-\lambda t}\right)\Big|_0^\infty + \int_0^\infty 2te^{-\lambda t} dt^2 \\ &= 2\int_0^\infty t \cdot e^{-\lambda t} dt = \mathbf{E}[X] \cdot \frac{2}{\lambda} = \frac{2}{\lambda^2}. \end{aligned}$$

Thus we have $\mathbf{Var}[X] = \frac{1}{\lambda^2}$. □

Proposition 30.4 (Lack of Memory) Let $X \sim Exp(\lambda)$. Then for any $t, s > 0$,

$$\mathbf{Pr}[X > t + s \mid X > s] = \mathbf{Pr}[X > t].$$

[30] In Proposition 30.1, we can regard τ_i as the time gap between the arrival of the $i-1$-th and the i-th customer. The parameter λ can be understood as the coming rate. Then the CDF of τ_i $F(t) = 1 - e^{-\lambda t}$ is the probability that the i-th customer comes within time t after the arrival of the $i-1$-th person.

Proof.

$$\begin{aligned}\mathbf{Pr}\left[X > t+s \mid X > s\right] &= \frac{\mathbf{Pr}\left[X > t+s \wedge X > s\right]}{\mathbf{Pr}\left[X > s\right]} = \frac{\mathbf{Pr}\left[X > t+s\right]}{\mathbf{Pr}\left[X > s\right]} \\ &= \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t}.\end{aligned}$$

□

Proposition 30.5 (Exponential Races) Let $X_1 \sim Exp(\lambda_1)$ and $X_2 \sim Exp(\lambda_2)$ be two independent random variables. Then $Y \triangleq \min\{X_1, X_2\} \sim Exp(\lambda_1 + \lambda_2)$.

Proof. By the independence, we have

$$\mathbf{Pr}\left[Y > t\right] = \mathbf{Pr}\left[X_1 > t \wedge X_2 > t\right] = \mathbf{Pr}\left[X_1 > t\right] \cdot \mathbf{Pr}\left[X_2 > t\right] = e^{-(\lambda_1+\lambda_2)t}.$$

□

Proposition 30.5 describes the distribution of the earliest customer of two restaurants. And we can easily generalize this to the case of more restaurants.

Corollary 30.1 Let $X_1, X_2 \ldots, X_n$ be n mutually independent random variables where $X_i \sim Exp(\lambda_i)$. Then $Y \triangleq \min\{X_1, X_2, \ldots, X_n\}$ has an exponential distribution with rate $\sum_{i=1}^{n} \lambda_i$.

Now we consider the problem "who wins the race?". That is, the restaurants are racing to see who will first have a customer. We first assume that there are only two random variables. Let f_λ be the probability density function of exponential distribution with rate λ. Using the law of total probability, we can compute the probability that X_1 wins the race as follows:

$$\begin{aligned}\mathbf{Pr}\left[X_1 < X_2\right] &= \int_0^\infty f_{\lambda_1}(t)\mathbf{Pr}\left[X_2 \geq t\right] dt \\ &= \int_0^\infty \lambda_1 e^{-\lambda_1 t} e^{-\lambda_2 t} dt \\ &= \lambda_1 \int_0^\infty e^{-(\lambda_1+\lambda_2)t} dt = \frac{\lambda_1}{\lambda_1 + \lambda_2}.\end{aligned}$$

Thus, clearly, the probability that X_i wins the race among n random variables is $\frac{\lambda_i}{\sum_{j=1}^{n} \lambda_j}$.

30.3 Proof of Proposition 30.1

Proposition 30.6 (Proposition 30.1 restated) Suppose that $\tau_1, \tau_2, \ldots, \tau_n$ is a sequence of independent random variables that $\tau_i \sim Exp(\lambda)$. Let $T_n = \sum_{i=1}^{n} \tau_i$. For $s \geq 0$, let $N(s) = \max\{n \mid T_n \leq s\}$. Then $N(s)$ is a Poisson process with rate λ.

Proof. Note that $T_n = \sum_{i=1}^{n} \tau_i$ is the arrival time of the n-th customer. Let g_n be the probability density function of T_n. First we prove that the distribution of T_n follows the *Gamma distribution* (n, λ):

$$g_n(t) = \begin{cases} \lambda e^{-\lambda t} \cdot \frac{(\lambda t)^{n-1}}{(n-1)!} & t \geq 0, \\ 0 & t < 0. \end{cases}$$

We prove this by induction. Note that when $n = 1$, $T_1 = \tau_1 \sim \mathrm{Exp}(\lambda) = \Gamma(1, \lambda)$. Suppose that $T_n \sim \Gamma(n, \lambda)$ for some $n \geq 1$. By the independence of T_n and τ_{n+1}, for $t \geq 0$ we have

$$\begin{aligned}g_{n+1}(t) &= \int_0^t g_n(s) \cdot f_\lambda(t-s) ds \\ &= \int_0^t \lambda e^{-\lambda s} \cdot \frac{(\lambda s)^{n-1}}{(n-1)!} \cdot \lambda e^{-\lambda(t-s)} ds \\ &= \lambda e^{-\lambda t} \frac{\lambda^n}{(n-1)!} \int_0^t s^{n-1} ds \\ &= \lambda e^{-\lambda t} \frac{\lambda^n}{(n-1)!} \cdot \frac{t^n}{n} = \lambda e^{-\lambda t} \cdot \frac{(\lambda t)^n}{n!}.\end{aligned}$$

Then we compute the distribution of $N(t)$.

$$\begin{aligned}\mathbf{Pr}\left[N(t)=n\right] &= \mathbf{Pr}\left[T_n \geq t \wedge T_{n+1} > t\right] \\ &= \int_0^t g_n(s)\cdot \mathbf{Pr}\left[\tau_{n+1} > t-s\right] ds \\ &= \int_0^t \lambda e^{-\lambda s}\cdot \frac{(\lambda s)^{n-1}}{(n-1)!}\cdot e^{-\lambda(t-s)} ds \\ &= \lambda^n e^{-\lambda t}\frac{t^n}{n!}.\end{aligned}$$

Thus, $N(t) \sim \mathrm{Pois}(\lambda t)$. Then we verify that $\{N(t) : t \geq 0\}$ satisfies the three conditions in Definition 30.1. [31] First it is clear that $N(t) = 0$ when $t = 0$. By the lack of memory property, we know that $N(s+t) - N(s)$ follows the same distribution as $N(t) - N(0)$, which $\mathrm{Pois}(\lambda t)$. Furthermore, it is easy to see that $N(s+t) - N(s)$ is independent of $N(r)$ for all $r \leq s$ again by the lack of memory property. It implies that $N(t)$ has independent increments, and hence completes our proof of Proposition 30.1. □

30.4 Thinning

In the example of customers coming into the restaurant, sometimes we have a more detailed characterization of customers, such as the gender. We associate an i.i.d. random variable Y_i with i-th arrival, and then use the value of Y_i to label the arrival and separate the Poisson process into several. Suppose that $Y_i \in \mathbb{N}$ and let $p_j = \mathbf{Pr}\left[Y_i = j\right]$. For all $j \in \mathsf{Range}(Y_i)$, let $N_j(t)$ denote the number of arrivals with label j that have arrived by time t. Then $\left\{N_j(t)\right\}$ is called a thinning of a Poisson process. We have the following useful and surprising proposition.

Proposition 30.7 For each j, $\left\{N_j(t) : t \geq 0\right\}$ is a Poisson process with rate $p_j\lambda$. Moreover, the collections of processes $\left\{\left\{N_j(t) : t \geq 0\right\} : j \in \mathsf{Range}(Y)\right\}$ are mutually independent.

[32] *Proof.* For convenience we assume that $Y_i \in \{0,1\}$. Then the following calculation concludes the independence and the distribution of $N_j(t)$ at the same time.

$$\begin{aligned}\mathbf{Pr}\left[N_0(t)=j \wedge N_1(t)=k\right] &= \mathbf{Pr}\left[N_0(t)=j \wedge N(t)=k+j\right] \\ &= \mathbf{Pr}\left[N(t)=k+j\right]\cdot \mathbf{Pr}\left[N_0(t)=j \mid N(t)=k+j\right] \\ &= e^{-\lambda t}\frac{(\lambda t)^{j+k}}{(j+k)!}\cdot \binom{j+k}{j} p_0^j p_1^k \\ &= e^{-p_0\lambda t}\frac{(p_0\lambda t)^j}{j!}\cdot e^{-p_1\lambda t}\frac{(p_1\lambda t)^k}{k!}.\end{aligned}$$

Thus, when there are n labels, it easy to verify that $N_j(t) \sim \mathrm{Pois}(p_j\lambda)$ and they are mutually independent. □

Let's see an application of Poisson process.

Example 29 (Maximum Likelihood of Poisson Process) Suppose there are two editors reading a book of 300 pages. Editor A finds 100 typos in the book, and editor B finds 120 typos, 80 of which are in common.

Suppose that the author's typos follow a Poisson process with some unknown rate λ per page. The two editors catch typos with unknown probabilities of success p_A and p_B respectively. We want to know how many typos there actually are. We can estimate this by determining λ, p_A and p_B. Clearly, there are four types of typos:

Type 1 The typo is found by neither of the editors. This happens w.p. $q_1 = (1-p_A)(1-p_B)$.

Type 2 The typo is found only by editor A. This happens w.p. $q_2 = (1-p_A)p_B$.

Type 3 The typo is found only by editor B. This happens w.p. $q_3 = (1-p_B)p_A$.

Type 4 The typo is found by both editors. This happens w.p. $q_4 = p_A p_B$.

So the occurrence of type i typos follows an independent Poisson process with rate $q_i\lambda$. That is, letting X_1, X_2, X_3 and X_4 be the occurrence time of the corresponding type of typos in this book, then $X_i \sim \mathrm{Pois}(300q_i\lambda)$. Note that there are 20 typos of type 2, 40 typos of type 3 and 80 typos of type 4. We claim that the most likely values of the rates are

$$\begin{cases} 300(1-p_A)p_B\lambda = 20, \\ 300(1-p_B)p_A\lambda = 40, \\ 300p_Ap_B\lambda = 80. \end{cases}$$

[31] Imagine the difference between $N(s+t) - N(s)$ and $N(t) - N(0)$. In the $N(t) - N(0)$, we start to wait for the first customer at time 0, while in $N(s+t) - N(s)$, at time s, we might have waited for the first customer in the period for sometime. However, due to the lack of memory property of the waiting time, this is equivalent to start to wait at time s.

[32] Here is an example explains why this proposition is surprising. Assume that the customers coming into a restaurant is a Poisson process, and each customer is male or female independently with probability 1/2 and 1/2 respectively. In fact we can assume that we flip a coin to determine whether the arriving customer is male or female. So intuitively, one might think that a large number of men (such as 50) arriving in one hour would indicates a large volume of business and hence a larger than normal number of women arriving. However this proposition tells us that the number of men arriving and the number of women arriving are independent.

This yields that $p_A = \frac{2}{3}$, $p_B = \frac{4}{5}$ and $\lambda = \frac{1}{2}$.

It remains to prove that claim. Suppose $X \sim \mathrm{Pois}(\theta)$ with some unknown θ. Then given z, our goal is to find $\arg\max_\theta \mathbf{Pr}\left[X = z \mid X \sim \mathrm{Pois}(\theta)\right]$. Note that $\mathbf{Pr}\left[X = z \mid X \sim \mathrm{Pois}(\theta)\right] = e^{-\theta}\frac{\theta^z}{z!}$ and $\log e^{-\theta}\frac{\theta^z}{z!} = -\theta + z \cdot \log\theta$. So it is equivalent to find

$$\arg\max_\theta -\theta + z \cdot \log\theta \tag{16}$$

Let the derivation of Equation (16) equals to 0. We have $\theta = z$, that is, $\arg\max_\theta e^{-\theta}\frac{\theta^z}{z!} = z$.

31 Coupon Collector Problem with Non-Uniform Coupons

Recall the coupon collector problem we discussed in Lecture 2: If each box of a brand of cereals contains a coupon which is chosen from n different types uniformly at random, then we need to buy nH_n boxes in expectation to collect all kinds of coupons.

In this lecture, we generalize the setting by involving the non-uniformity. Suppose that each purchase gives a coupon of type j w.p. p_j for $j \in [n]$ and the coupon types contained in different boxes are independent. It is clear that $\sum_{j=1}^n p_j = 1$. Let N_j be the first time that we get type j. Then N_j follows the geometric distribution with parameter p_j. Let N be the number of purchases until all n types of coupons are collected, that is, $N = \max_{j\in[n]} N_j$. We would like to compute $\mathbf{E}[N]$ to see how many times of purchases is needed in expectation. However, it is not easy to compute the expected value of $\max_{j\in[n]} N_j$ since N_j's are not independent.

31.1 Coupon Collector Problem with Poisson Draw

We consider a similar case that the coupons are collected with Poisson draw. That is, each arrival of the Poisson process with rate 1 brings a coupon and the probability of the coupon being type j is p_j. Note that this process is different from the ordinary coupon collector problem since the arrival time is random.

Recall the thinning of Poisson process we discussed in the last lecture. Let $X_j(t)$ be the number of type j coupons we collect in time $[0, t]$ with Poisson draw. Then $\left\{X_j(t)\right\}$ is a thinning, that is, $\left\{X_j(t)\right\}$ is a Poisson process with rate p_j and $X_j(t)$ is independent with $X_i(t)$ for $i \neq j$. For $j \in [n]$, let $T_j \triangleq \min\left\{t \mid X_j(t) = 1\right\}$ be the first time that type j coupon appears. Obviously, T_j is the same as $\tau_j(1)$ [33] and $T_j \sim \mathrm{Exp}(p_j)$.

To ascertain the time of collecting all kinds of coupons, we need to compute $\mathbf{E}[T]$ where $T = \max_{j\in[n]} T_j$. This will be much easier since T_j is independent with each other. First we introduce a basal proposition in probability theory.

Proposition 31.1 Let X be a non-negative random variable.

- If X is discrete and $X \in \mathbb{N}$, then $\mathbf{E}[X] = \sum_{t=1}^\infty \mathbf{Pr}[X \geq t]$.
- If X is continuous, then $\mathbf{E}[X] = \int_0^\infty \mathbf{Pr}[X \geq t]\, dt$.

Proof.

- When X is discrete, we apply the double counting skill:

$$\begin{aligned}\mathbf{E}[X] &= \sum_{s=1}^\infty s\mathbf{Pr}[X = s] = \sum_{s=1}^\infty \sum_{t=1}^s \mathbf{Pr}[X = s] \\ &= \sum_{t=1}^\infty \sum_{s=t}^\infty \mathbf{Pr}[X = s] = \sum_{t=1}^\infty \mathbf{Pr}[X \geq t]\,.\end{aligned}$$

- When X is continuous,

$$\begin{aligned}\mathbf{E}[X] &= \mathbf{E}\left[\int_0^X 1\, dt\right] = \mathbf{E}\left[\int_0^\infty \mathbf{1}[X \geq t]\, dt\right] \\ &\overset{(\heartsuit)}{=} \int_0^\infty \mathbf{E}\left[\mathbf{1}[X \geq t]\right]\, dt = \int_0^\infty \mathbf{Pr}[X \geq t]\, dt,\end{aligned}$$

where $(\heartsuit)$ comes from the *Fubini's theorem.*

□

Note that for any $t \in \mathbb{R}_{\geq 0}$,

$$\mathbf{Pr}[T \geq t] = 1 - \mathbf{Pr}[T < t] = 1 - \prod_{j=1}^n \mathbf{Pr}\left[T_j < t\right] = 1 - \prod_{j=1}^n \left(1 - e^{-p_j t}\right).$$

[33] Here $\tau_j(1)$ denotes the time gap between the arrival of the customers with coupon j.

By the continuous version of Proposition 31.1, we have

$$\mathbf{E}[T] = \int_0^\infty \mathbf{Pr}[T \geq t]\, dt = \int_0^\infty 1 - \prod_{j=1}^n \left(1 - e^{-p_j t}\right) dt.$$

That is, we need a time of $\int_0^\infty \left(1 - e^{-p_j t}\right) dt$ in expectation to collect all kinds of coupons.

31.2 Ordinary Coupon Collector Problem

Then we link the result in Section 31.1 to the ordinary coupon collector problem by coupling. Specifically, let τ_i denote the time gap between the $i-1$-th and the i-th arrival. Imagine the ordinary version as one customer comes with a coupon in every slot, that is, the time gap is a constant. We couple the two process by letting the i-th arrival in the Poisson version carry the same type of coupon with the i-th arrival in the ordinary version.

Recall that N is the number of purchases until all n types of coupons are collected in the ordinary coupon collector problem. Then we have $T = \sum_{i=1}^N \tau_i$. Note that $\tau_i \sim \mathrm{Exp}(1)$ and $\mathbf{E}[\tau_i] = 1$. If N is a constant, we can deduce $\mathbf{E}[N] = \mathbf{E}\left[\sum_{i=1}^N \tau_i\right] = \mathbf{E}[T]$ directly. However, N is a random variable and thus the summation and expectation are not guaranteed to be exchangeable. To show the validity of $\mathbf{E}[N]\,\mathbf{E}[\tau_i] = \mathbf{E}\left[\sum_{i=1}^N \tau_i\right]$ in this case, we introduce the following theorem.

Theorem 31.1 (Wald's Equation) Let $X_1, X_2, \ldots$ be n i.i.d. random variables that $\mathbf{E}[|X_1|] < \infty$. Let T be a stopping time that $\mathbf{E}[T] < \infty$. Then we have $\mathbf{E}\left[\sum_{t=1}^T X_t\right] = \mathbf{E}[T]\,\mathbf{E}[X_1]$.

Proof. Let $Z_k = \sum_{i=1}^k (X_i - \mathbf{E}[X_i])$. Then $\{Z_k\}$ is a martingale with regard to $\{X_i\}$. We check that T satisfies the third condition of the optional stopping theorem which requires $\mathbf{E}[T] < \infty$ and $\mathbf{E}[|Z_{i+1} - Z_i| \mid \mathcal{F}_i] < \infty$.[34] Note that

$$\begin{aligned}\mathbf{E}[|Z_{i+1} - Z_i| \mid \mathcal{F}_i] &= \mathbf{E}[|X_{i+1} - \mathbf{E}[X_{i+1}]| \mid \mathcal{F}_i] \\ &\leq \mathbf{E}[|X_{i+1}| + |\mathbf{E}[X_{i+1}]| \mid \mathcal{F}_i] \leq 2\mathbf{E}[|X_{i+1}|] < \infty.\end{aligned}$$

Combining the given condition that $\mathbf{E}[T] < \infty$ and applying the optional stopping theorem, we have that $\mathbf{E}[Z_T] = \mathbf{E}[Z_1] = 0$, that is, $\mathbf{E}\left[\sum_{i=1}^T (X_i - \mathbf{E}[X_i])\right] = 0$. Thus we have $\mathbf{E}\left[\sum_{i=1}^T X_i\right] = \mathbf{E}\left[\sum_{i=1}^T \mathbf{E}[X_i]\right] = \mathbf{E}[T] \cdot \mathbf{E}[X_t]$. □

It is easy to verify that $\mathbf{E}[\tau_i] = 1 < \infty$ and $\mathbf{E}[N] < \infty$ in our case. So applying the Wald's equation, we have $\mathbf{E}[N]\,\mathbf{E}[\tau_i] = \mathbf{E}\left[\sum_{i=1}^N \tau_i\right]$ and sequentially

$$\mathbf{E}[N] = \mathbf{E}[T] = \int_0^\infty 1 - \prod_{j=1}^n \left(1 - e^{-p_j t}\right) dt. \tag{17}$$

Then we go back to the coupon collector problem with uniform coupons. Let $x = e^{-\frac{t}{n}}$. If $p_j = \frac{1}{n}$ for any $j \in [n]$, we have

$$\begin{aligned}\mathbf{E}[N] &= \int_0^\infty 1 - \prod_{j=1}^n \left(1 - e^{-p_j t}\right) dt \\ &= n \int_0^\infty 1 - (1-x)^n \, d\log x \\ &= n \int_0^\infty \frac{1}{x} - \frac{(1-x)^n}{x} dx \\ &= n \int_0^\infty \sum_{k=1}^n \frac{(1-x)^{k-1}}{x} - \frac{(1-x)^k}{x} dx \\ &\overset{(\heartsuit)}{=} n \sum_{k=1}^n \int_0^\infty (1-x)^{k-1} dx \\ &= n \sum_{k=1}^n \frac{1}{k} = nH_n,\end{aligned}$$

where the ($\heartsuit$) follows from the *Fubini's theorem*. This verifies the validity of Equation (17) when the types of coupons are uniform.

32 Balls-into-Bins

Recall the balls-into-bins problem where we throw m identical balls into n bins. For $i \in [n]$, let X_i be the number of balls in the i-th bin. Then we have $X_i \sim \mathrm{Binom}(m, \frac{1}{n})$ and $\mathbf{E}[X_i] = \frac{m}{n}$. This model can be used to describe the scheme of the hash table. To avoid frequent collision when mapping the keys into slots, it is natural for us to be concerned about the value of $\max_{i \in [n]} X_i$. However, we are faced with the difficulty that X_i's are not independent when computing the distribution of $\max_{i \in [n]} X_i$. It turns out that one can use independent Poisson variables to approximate the distribution. First we have:

[34] Here $\mathcal{F}_i = \sigma(X_1, X_2, \ldots, X_i)$.

Theorem 32.1 The distribution of $(X_1, X_2, \ldots, X_n)$ is the same as that of $(Y_1, Y_2, \ldots, Y_n)$ on condition that $\sum_{i=1}^n Y_i = m$ where $Y_i \sim Pois(\lambda)$ are independent Poisson random variables with an arbitrary rate λ.

Proof. Given $(a_1, a_2, \ldots, a_n) \in \mathbb{N}^n$ and $\sum_{i=1}^n a_n = m$, we have

$$\mathbf{Pr}\left[(X_1, X_2, \ldots, X_n) = (a_1, a_2, \ldots, a_n)\right] = \frac{1}{n^m} \cdot \frac{m!}{a_1! a_2! \cdots a_n!}. \tag{18}$$

And

$$\begin{aligned}
&\mathbf{Pr}\left[(Y_1, Y_2, \ldots, Y_n) = (a_1, a_2, \ldots, a_n) \,\middle|\, \sum_{i=1}^n Y_i = m\right] \\
&= \frac{\mathbf{Pr}\left[(Y_1, Y_2, \ldots, Y_n) = (a_1, a_2, \ldots, a_n) \wedge \sum_{i=1}^n Y_i = m\right]}{\mathbf{Pr}\left[\sum_{i=1}^n Y_i = m\right]} \\
&= \frac{\prod_{i=1}^n \mathbf{Pr}\left[Y_i = a_i\right]}{\mathbf{Pr}\left[\sum_{i=1}^n Y_i = m\right]} \\
&= \frac{\prod_{i=1}^n e^{-\lambda} \frac{\lambda^{a_i}}{a_i!}}{e^{-\lambda n} \frac{(\lambda n)^m}{m!}} = \frac{1}{n^m} \cdot \frac{m!}{a_1! a_2! \cdots a_n!},
\end{aligned}$$

which equals to the RHS of Equation (18). □

Furthermore, we can deduce the following corollary from Theorem 32.1.

Corollary 32.1 Let $f\colon \mathbb{N}^n \to \mathbb{N}$ be an arbitrary function and $Y_1, Y_2, \ldots, Y_n$ be n independent Poisson random variables with rate $\lambda = \frac{m}{n}$. Then we have

$$\mathbf{E}\left[f(X_1, X_2, \ldots, X_n)\right] \le e\sqrt{m} \cdot \mathbf{E}\left[f(Y_1, Y_2, \ldots, Y_n)\right].$$

Proof. By the law of total probability, we have

$$\begin{aligned}
\mathbf{E}\left[f(Y_1, Y_2, \ldots, Y_n)\right] &= \sum_{k=0}^{\infty} \mathbf{E}\left[f(Y_1, Y_2, \ldots, Y_n) \,\middle|\, \sum_{i=1}^n Y_i = k\right] \mathbf{Pr}\left[\sum_{i=1}^n Y_i = k\right] \\
&\ge \mathbf{E}\left[f(Y_1, Y_2, \ldots, Y_n) \,\middle|\, \sum_{i=1}^n Y_i = m\right] \mathbf{Pr}\left[\sum_{i=1}^n Y_i = m\right] \\
&= \mathbf{E}\left[f(X_1, X_2, \ldots, X_n)\right] \mathbf{Pr}\left[\sum_{i=1}^n Y_i = m\right].
\end{aligned}$$

Note that $\sum_{i=1}^n Y_i \sim \text{Pois}(m)$, then we have

$$\mathbf{Pr}\left[\sum_{i=1}^n Y_i = m\right] = e^{-m} \frac{m^m}{m!} > \frac{1}{e\sqrt{m}},$$

where the inequality comes from the *Stirling's formula.* [35] □

Equipped with Corollary 32.1, we have the following theorem to bound $X = \max_{i \in [n]} X_i$.

Theorem 32.2 (Max Load) When $m = n$, we have $X = \Theta\left(\frac{\log n}{\log \log n}\right)$ w.p. $1 - o(1)$.

Proof. First we prove the upper bound, that is, there exists a constant c_1 such that $\mathbf{Pr}\left[X \ge \frac{c_1 \log n}{\log \log n}\right] = o(1)$. Let $k = \frac{c_1 \log n}{\log \log n}$ for brevity. By union bound, we have

$$\begin{aligned}
\mathbf{Pr}\left[X \ge k\right] &= \mathbf{Pr}\left[\exists i \in [n], X_i \ge k\right] \le \sum_{i=1}^n \mathbf{Pr}\left[X_i \ge k\right] \\
&= n \cdot \mathbf{Pr}\left[X_1 \ge k\right] \le n \cdot \binom{n}{k} \frac{1}{n^k} \le n \cdot \left(\frac{en}{k}\right)^k \frac{1}{n^k} = n \cdot \left(\frac{e}{k}\right)^k.
\end{aligned}$$

Note that

$$\begin{aligned}
k \log k &= \frac{c_1 \log n}{\log \log n} \cdot (\log \log n - \log \log \log n + \log c_1) \\
&> c_1 \log n \left(1 - \frac{\log \log \log n}{\log \log n}\right) > \frac{c_1}{2} \log n.
\end{aligned}$$

[35] We can see from the proof of Corollary 32.1 that the choice of $\lambda = \frac{m}{n}$ is to maximize $\mathbf{Pr}\left[\sum_{i=1}^n Y_i = m\right]$.

Letting $c = 6$, we have that

$$\log n + k - k \log k < -\log n.$$

Thus, $\mathbf{Pr}[X \geq k] \leq n \cdot \left(\frac{e}{k}\right)^k < \frac{1}{n} = o(1)$ for $c_1 = 6$.

Then we prove the lower bound. Again let $g = \frac{c_2 \log n}{\log \log n}$ for a constant c_2. Let $f(X_1, X_2, \ldots, X_n) \triangleq \mathbf{1}[X < g] = \mathbf{1}[\max_{i \in [n]} X_i < g]$. Then by Corollary 32.1,

$$\begin{aligned} \mathbf{Pr}[X < g] &= \mathbf{E}[f(X_1, X_2, \ldots, X_n)] \\ &\leq e\sqrt{n} \cdot \mathbf{E}[f(Y_1, Y_2, \ldots, Y_n)] \\ &= e\sqrt{n} \cdot \mathbf{Pr}\left[\max_{i \in [n]} Y_i < g\right]. \end{aligned} \tag{19}$$

By the definition of Y_i in Corollary 32.1, we have

$$\begin{aligned} \mathbf{Pr}\left[\max_{i \in [n]} Y_i < g\right] &= (\mathbf{Pr}[Y_1 \leq g])^n = (1 - \mathbf{Pr}[Y_1 > g])^n \\ &\leq (1 - \mathbf{Pr}[Y_1 = g + 1])^n = \left(1 - \frac{1}{(g+1)!e}\right)^n \leq e^{-\frac{n}{(g+1)!e}} \end{aligned}$$

Note that

$$\begin{aligned} \log(g+1)! &= \sum_{i=1}^{g+1} \log i < \int_1^{g+2} \log x \, dx \\ &= (g+2)\log(g+2) - g - 1 \leq (g+2)\log g - g + 3 \\ &= \frac{c_2 \log n + 2 \log \log n}{\log \log n} (\log \log n - \log \log \log n + \log c_2) - \frac{c_2 \log n}{\log \log n} + 3 \\ &\leq c_2 \log n - \log \log n - 2. \end{aligned}$$

Letting $c_2 = 1$, we have $\log(g+1)! \leq \log n - \log \log n - 2$ and sequentially

$$e(g+1)! \leq \frac{n}{e \log n}.$$

Thus,

$$\mathbf{Pr}\left[\max_{i \in [n]} Y_i < g\right] \leq e^{-\frac{n}{(g+1)!e}} \leq e^{-e \log n} = n^{-e}.$$

Combining with Equation (19), we have $\mathbf{Pr}\left[X < \frac{\log n}{\log \log n}\right] \leq e\sqrt{n} \cdot n^{-e} = o(1)$. □

33 Brownian Motion

Brownian motion describes the random motion of small particles suspended in a liquid or in a gas. This process was named after the botanist Robert Brown, who observed and studied a jittery motion of pollen grains suspended in water under a microscope. Later, Albert Einstein gave a physical explanation of this phenomenon. In mathematics, Brownian motion is characterized by the *Wiener process*, named after Norbert Wiener, a famous mathematician and the originator of cybernetics.

To motivate the definition of Brownian motion, we start from the 1-D random walk starting from 0. Let Z_t be our position at time t and X_t be the move of the t-th step. The value of X_t is chosen from $\{-1, 1\}$ uniformly at random. Note that $Z_0 = 0$ and $Z_{t+1} = Z_t + X_t$. So $Z_T = \sum_{t=0}^{T-1} X_t$. Then we have

$$\mathbf{E}[Z_T] = 0 \text{ and } \mathbf{Var}[Z_T] = \sum_{t=0}^{T-1} \mathbf{Var}[X_t] = T.$$

Suppose now we move with every Δt seconds and with step length δ. Then our position at time T is $Z(T) = \delta \sum_{t=1}^{\frac{T}{\Delta t}} X_t$. We are interested in the behavior of the prcoess when $\Delta t \to 0$. We have

$$\mathbf{E}[Z(T)] = 0 \text{ and } \mathbf{Var}[Z(T)] = \delta^2 \sum_{t=1}^{\frac{T}{\Delta t}} \mathbf{Var}[X_t] = \delta^2 \cdot \frac{T}{\Delta t}.$$

We can identify the expectation and the variance of this process with the discrete random walk when $\Delta t \to 0$ by choosing $\delta = \sqrt{\Delta t}$. It follows from the central limit theorem that

$$Z(T) = \sqrt{\Delta t} \sum_{t=1}^{\frac{T}{\Delta t}} X_t \xrightarrow{\Delta t \to 0} \sqrt{\Delta t} \mathcal{N}(0, \frac{T}{\Delta t}) = \mathcal{N}(0, T).$$

In other words, the "continuous" version of the 1-D random walk follows $\mathcal{N}(0, T)$ at time T. This is the basis of the Wiener process. Now we introduce its formal definition.

Definition 33.1 (Brownian Motion, Wiener Process) We say a stochastic process $\{W(t)\}_{t\geq 0}$ is a standard Brownian motion or Wiener process if it satisfies

- $W(0) = 0$;
- ***Independent increments***: $\forall 0 \leq t_0 \leq t_1 \leq \cdots \leq t_n$, $W(t_1) - W(t_0)$, $W(t_2) - W(t_1)$, $\ldots$, $W(t_n) - W(t_{n-1})$ are mutually independent;
- ***Stationary increments***: $\forall s, t > 0$, $W(s+t) - W(s) \sim \mathcal{N}(0, t)$;
- $W(t)$ is continuous almost surely.[a]

[a] Let Ω be the sample space. Then W can be viewed as a mapping from $\mathbb{R} \times \Omega$ to $\mathbb{R}$. The meaning of "$W(t)$ is continuous almost surely" is: $\exists \Omega_0 \subseteq \Omega$ with $\mathbf{Pr}[\Omega_0] = 1$ such that $\forall \omega \in \Omega_0$, $W(t, \omega)$ is continuous with regard to t.

Let $\{W(t)\}_{t\geq 0}$ be a a standard Brownian motion. If $\{X(t)\}_{t\geq 0}$ satisfies $X(t) = \mu \cdot t + \sigma W(t)$, we call $\{W(t)\}_{t\geq 0}$ a (μ, σ^2) Brownian motion.

Now we introduce another characterization of Brownian motion. First recall the notion of high dimensional Gaussian distribution. A vector of random variables $(X_1, X_2, \ldots, X_n)$ is said to be Gaussian iff $\forall a_1, a_2, \ldots, a_n$, $\sum_{i=1}^n a_i X_i$ is a one-dimensional Gaussian. Let $\mu = (\mu_1, \mu_2, \ldots, \mu_n)$ where $\mu_i = \mathbf{E}[X_i]$. Let $\Sigma = \left(\mathrm{Cov(X_i, X_j)}\right)_{i,j}$. Then the probability density function f of $(X_1, X_2, \ldots, X_n)$ is

$$\text{for } x = (x_1, x_2, \ldots, x_n),\ f(x) = (2\pi)^{-\frac{n}{2}} \cdot |\mathrm{det}\Sigma|^{-\frac{1}{2}} \cdot e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}.$$

To give another characterization of standard Brownian motion, we first define the *Gaussian process*.

Definition 33.2 (Gaussian Process) A stochastic process $\{W(t)\}_{t\geq 0}$ is called Gaussian process if $\forall 0 \leq t_1 \leq t_2 \leq \cdots \leq t_n$, $(W(t_1), W(t_2), \ldots, W(t_n))$ is a Gaussian.

Definition 33.3 (Brownian Motion, Wiener Process) We say a stochastic process $\{W(t)\}_{t\geq 0}$ is a standard Brownian motion or Wiener process if it satisfies

- $\{W(t)\}_{t\geq 0}$ is an almost surely continuous Gaussian Process;
- $\forall s \geq 0$, $\mathbf{E}[W(s)] = 0$;
- $\forall 0 \leq s \leq t$, $Cov(W(s), W(t)) = s$.

Sometimes it is easier to use Definition 33.3 to show that a stochastic process is a Brownian motion. We now verify the equivalence between Definition 33.1 and Definition 33.3.

Proof. Given Definition 33.1, it is easy to know that $\mathbf{E}[W(s)] = 0$ for all $s \geq 0$ since $W(s) \sim \mathcal{N}(0, s)$. What we need is to verify that $\{W(t)\}_{t\geq 0}$ in Definition 33.1 is a Gaussian process and to compute the covariance of $W(s)$ and $W(t)$ in Definition 33.1.

Note that $\forall 0 \leq s < t$ and $\forall a, b$, we have

$$aW(s) + bW(t) = (a+b)W(s) + b(W(t) - W(s)).$$

Since $W(s)$ and $W(t) - W(s)$ are two independent Gaussian's, $aW(s) + bW(t)$ is still a Gaussian.

By the distributive law of covariance, for any $0 \leq s \leq t$, we have

$$\begin{aligned}\mathrm{Cov}(W(s), W(t)) &= \mathrm{Cov}\left(W(s), W(t) - W(s) + W(s)\right)\\ &= \mathrm{Cov}(W(s), W(t) - W(s)) + \mathrm{Cov}(W(s), W(s))\\ &= \mathbf{Var}[W(s)] = s.\end{aligned}$$

Then we consider the counterpart. Given Definition 33.3, we can deduce the first and fourth property in Definition 33.1 directly. For any $0 \leq t_{i-1} \leq t_i \leq t_{j-1} \leq t_j$, we have

$$\begin{aligned}&\mathrm{Cov}(W(t_i) - W(t_{i-1}), W(t_j) - W(t_{j-1}))\\ &= \mathrm{Cov}(W(t_i), W(t_j)) + \mathrm{Cov}(W(t_{i-1}), W(t_{j-1}))\\ &\quad - \mathrm{Cov}(W(t_i), W(t_{j-1})) - \mathrm{Cov}(W(t_{i-1}), W(t_j))\\ &= t_i + t_{i-1} - t_i - t_{i-1} = 0,\end{aligned}$$

which yields the independence of $W(t_i) - W(t_{i-1})$ and $W(t_j) - W(t_{j-1})$. Thus, the $\{W(t)\}_{t\geq 0}$ in Definition 33.3 satisfies independent increments.

It is easy to verify that $\forall s, t > 0$, $W(s+t) - W(s)$ is a Gaussian with mean 0. Note that

$$\begin{aligned}\mathbf{Var}[W(t+s) - W(s)] &= \mathbf{E}\left[(W(t+s) - W(s))^2\right]\\ &= \mathbf{E}\left[W(t+s)^2\right] + \mathbf{E}\left[W(s)^2\right] - 2\mathbf{E}[W(t+s)W(s)]\\ &= \mathbf{Var}\left[W(t+s)^2\right] + \mathbf{Var}\left[W(s)^2\right] - 2\mathrm{Cov}\left(W(t+s), W(s)\right)\\ &= t + s + s - 2s = t.\end{aligned}$$

Thus, the $\{W(t)\}_{t\geq 0}$ in Definition 33.3 satisfies stationary increments. □

Example 30 Suppose $\{W(t)\}_{t\geq 0}$ is a standard Brownian motion. We claim that $\{X(t)\}_{t\geq 0}$ is also a standard Brownian motion where $X(0) = 0$ and $X(t) = t \cdot W(\frac{1}{t})$ for $t > 0$.

We verify the three requirements in Definition 33.3.

Since $X(t) = t \cdot W(\frac{1}{t})$ which is the compound of two (almost surely) continuous function, $\{X(t)\}_{t\geq 0}$ is also continuous almost surely. For any $a_1, a_2, \ldots, a_n$ and $t_1, t_2, \ldots, t_n \geq 0$, $\sum_{i=1}^n a_i X(t_i) = \sum_{i=1}^n a_i t_i \cdot W(\frac{1}{t_i})$. Since $\{W(t)\}$ is standard Brownian motion, $\sum_{i=1}^n a_i t_i \cdot W(\frac{1}{t_i})$ is Gaussian. Thus, $\{X(t)\}_{t\geq 0}$ is a Gaussian process. For $0 \leq s < t$,

$$\begin{aligned}\mathrm{Cov}(X(s), X(t)) &= \mathrm{Cov}(sW(\frac{1}{s}), tW(\frac{1}{t}))\\ &= st \cdot \mathrm{Cov}(W(\frac{1}{s}), W(\frac{1}{t}))\\ &= st \cdot \frac{1}{t} = s.\end{aligned}$$

Thus, $\{X(t)\}_{t\geq 0}$ is a standard Brownian motion.

Example 31 (Hitting time) We consider the first hitting time of position b in a Brownian motion. Define $\tau_b \triangleq \inf\{t \geq 0 \mid W(t) >$
For any $t > 0$,

$$\begin{aligned}\mathbf{Pr}[\tau_b < t] &= \mathbf{Pr}[\tau_b < t \wedge W(t) > b] + \mathbf{Pr}[\tau_b < t \wedge W(t) < b]\\ &= \mathbf{Pr}[W(t) > b] + \mathbf{Pr}[W(t) < b \mid \tau_b < t] \cdot \mathbf{Pr}[\tau_b < t].\end{aligned}$$

Note that $W(t) \sim \mathcal{N}(0, t)$. Let Φ be the cumulative distribution function of standard Gaussian distribution, that is, $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$. Then

$$\mathbf{Pr}[W(t) > b] = \mathbf{Pr}\left[\frac{W(t)}{\sqrt{t}} > \frac{b}{\sqrt{t}}\right] = 1 - \Phi\left(\frac{b}{\sqrt{t}}\right).$$

Assume we have known the value of τ_b and $\tau_b < t$, we can regard $\{W(t)\}_{t\geq\tau_b}$ as a Brownian motion starting from b. Thus, as Figure 1 shows, $\mathbf{Pr}[W(t) < b \mid \tau_b < t] = \frac{1}{2}$.

By direct calculation, we have $\mathbf{Pr}[\tau_b < t] = 2\left(1 - \Phi\left(\frac{b}{\sqrt{t}}\right)\right)$.

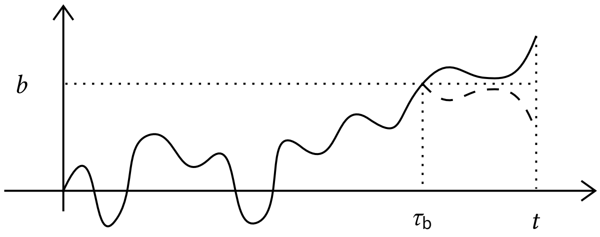


Figure 1: A hitting time and the reflection principle

34 Brownian Bridge

Consider a Brownian motion starting from $W(0) = 0$ and ending at $W(u) = x$. Conditioned on fixed $W(0)$ and $W(u)$, what is the distribution of $W(t)$? By definition, for $t < u$, conditioned on $W(u) = x$, $W(t)$ is a Gaussian. So it is sufficient to compute its mean and variance. As Figure 2 shows, a natural conjecture is that for $0 \leq t \leq u$, $\mathbf{E}[W(t) \mid W(u)] = \frac{t}{u} W(u)$.

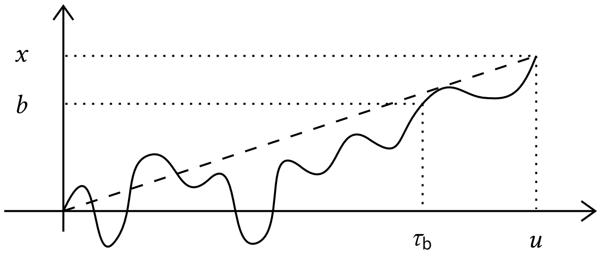


Figure 2: A Brownian bridge

To verify this, we first prove the following proposition.

Proposition 34.1 For any $0 \le t \le u$, $W(t) - \frac{t}{u}W(u)$ is independent of $W(u)$.

Proof.

$$\begin{aligned}\mathrm{Cov}(W(t) - \frac{t}{u}W(u), W(u)) &= \mathrm{Cov}(W(t), W(u)) - \frac{t}{u}\mathbf{Var}\left[W(u)\right]\\ &= t - \frac{t}{u}\cdot u = 0.\end{aligned}$$

The proposition follows from the fact that the two Gaussians are independent iff their covariance is zero. □

Thus, we have

$$\begin{aligned}0 = &\mathbf{E}\left[W(t) - \frac{t}{u}W(u)\right]\\ =&\mathbf{E}\left[W(t) - \frac{t}{u}W(u) \;\middle|\; W(u)\right] = \mathbf{E}\left[W(t) \mid W(u)\right] - \frac{t}{u}W(u).\end{aligned}$$

This confirms the conjecture that for $0 \le t \le u$, $\mathbf{E}\left[W(t) \mid W(u)\right] = \frac{t}{u}W(u)$. Then we consider the variance of $W(t)$ conditioned on $W(u)$.

$$\begin{aligned}\mathbf{Var}\left[W(t) \mid W(u)\right] &= \mathbf{E}\left[(W(t) - \mathbf{E}\left[W(t) \mid W(u)\right])^2 \;\middle|\; W(u)\right]\\ &= \mathbf{E}\left[\left(W(t) - \frac{t}{u}W(u)\right)^2 \;\middle|\; W(u)\right]\\ &= \mathbf{E}\left[\left(W(t) - \frac{t}{u}W(u)\right)^2\right]\\ &= \mathbf{E}\left[W(t)^2\right] + \frac{t^2}{u^2}\mathbf{E}\left[W(u)^2\right] - 2\frac{t}{u}\mathbf{E}\left[W(t)W(u)\right]\\ &= \frac{t(u-t)}{u}.\end{aligned}$$

Finally, to characterize the distribution of $\{W(t)\}$ conditioned on $W(u) = x$, we compute the covariance. Let $p_{W(t)}$ be the probability density function of $W(t)$. For any $s \le t$,

$$\begin{aligned}&\mathrm{Cov}(W(s), W(t)|W(u))\\ =&\mathbf{E}\left[W(s)\cdot W(t) \mid W(u)\right] - \mathbf{E}\left[W(s) \mid W(u)\right]\cdot\mathbf{E}\left[W(t) \mid W(u)\right]\\ =&\int_{\mathbb{R}} y\cdot\mathbf{E}\left[W(s) \mid W(t) = y, W(u)\right]\cdot p_{W(t)}(y|W(u))dy - \frac{st}{u^2}W(u)^2\\ =&\int_{\mathbb{R}} y\cdot\frac{s}{t}y\cdot p_{W(t)}(y|W(u))dy - \frac{st}{u^2}W(u)^2\\ =&\frac{s}{t}\mathbf{E}\left[W(t)^2|W(u)\right] - \frac{st}{u^2}W(u)^2\\ =&\frac{s(u-t)}{u}.\end{aligned}$$

To sum up, conditioned on $W(u)$, $\{W(t)\}$ has the following three properties:

- For $t \in [0,u]$, $\mathbf{E}\left[W(t) \mid W(u)\right] = \frac{t}{u}W(u)$.
- For $t \in [0,u]$, $\mathbf{Var}\left[W(t) \mid W(u)\right] = \frac{t(u-t)}{u}$.
- For any $0 \le s \le t \le u$, $\mathrm{Cov}(W(s), W(t)|W(u)) = \frac{s(u-t)}{u}$.

We call the Brownian motion $\{W(t)\}_{t\ge0}$ which ends at $W(u) = x$ a Brownian bridge. Furthermore, we define the standard Brownian bridge.

Definition 34.1 (Standard Brownian Bridge) A standard Brownian motion ending at $W(1) = 0$ is called a standard Brownian bridge.

We can verify that letting $X(t) = W(t) - tW(1)$ where $\{W(t)\}$ is a standard Brownian motion, then $\{X(t)\}$ is a standard Brownian Bridge.

Example 32 (Hitting Time in a Brownian Bridge) Let $\{W(t)\}_{t\ge0}$ be a standard Brownian motion. Let $\tau_b \triangleq \inf\{t \ge 0 \mid W(t)$ Then we compute $\mathbf{Pr}\left[\tau_b < u \mid W(u) = x\right]$. Note that if $b < x$, $\mathbf{Pr}\left[\tau_b < u \mid W(u) = x\right] = 1$. Let ψ be the probability density function of standard Gaussian distribution, that is, $\psi(x) = \frac{1}{\sqrt{2\pi}}\cdot e^{-\frac{x^2}{2}}$. If $b > x$, letting $\mathbf{d}x = [x, x+h]$ where h is infinitesimal, we have

$$\begin{aligned}\mathbf{Pr}\left[\tau_b < u \mid W(u) = x\right] &= \frac{\mathbf{Pr}\left[\tau_b < u \wedge W(u) \in \mathrm{d}x\right]}{\mathbf{Pr}\left[W(u) \in \mathrm{d}x\right]} \\ &= \frac{\mathbf{Pr}\left[\tau_b < u\right] \cdot \mathbf{Pr}\left[W(u) \in \mathrm{d}x \mid \tau_b < u\right]}{\frac{1}{\sqrt{u}}\psi\left(\frac{x}{\sqrt{u}}\right)\,\mathrm{d}x}.\end{aligned}$$

If we have known the value of τ_b and $\tau_b < u$, we can regard $\{W(u)\}_{t\geq\tau_b}$ as a Brownian motion starting from b. Then we have

$$\begin{aligned}\mathbf{Pr}\left[\tau_b < u\right] \cdot \mathbf{Pr}\left[W(u) \in \mathrm{d}x \mid \tau_b < u\right] &= \mathbf{Pr}\left[\tau_b < u\right] \cdot \mathbf{Pr}\left[W(u) \in 2b - \mathrm{d}x \mid \tau_b < u\right] \\ &= \mathbf{Pr}\left[\tau_b < u \wedge W(u) \in 2b - \mathrm{d}x\right] \\ &= \mathbf{Pr}\left[W(u) \in 2b - \mathrm{d}x\right] \\ &= \frac{1}{\sqrt{u}}\psi\left(\frac{2b-x}{\sqrt{u}}\right)\mathrm{d}x\end{aligned}$$

Thus, when $b > x$, $\mathbf{Pr}\left[\tau_b < u \mid W(u) = x\right] = \frac{\psi\left(\frac{2b-x}{\sqrt{u}}\right)}{\psi\left(\frac{x}{\sqrt{u}}\right)} = e^{-\frac{2b(b-x)}{u}}$.

When $b = x$, we have

$$\mathbf{Pr}\left[\tau_b < u \mid W(u) = b\right] = \frac{\mathbf{Pr}\left[\tau_b < u \wedge W(u) \in \mathrm{d}b\right]}{\mathbf{Pr}\left[W(u) \in \mathrm{d}b\right]}.$$

Note that

$$\begin{aligned}&\mathbf{Pr}\left[\tau_b < u \wedge W(u) \in \mathrm{d}b\right] \\ =&\mathbf{Pr}\left[\tau_b < u\right] - \mathbf{Pr}\left[\tau_b < u \wedge W(u) > b + h\right] - \mathbf{Pr}\left[\tau_b < u \wedge W(u) < b\right].\end{aligned} \tag{20}$$

By Example 31, we have $\mathbf{Pr}\left[\tau_b < u\right] = 2\left(1 - \Phi\left(\frac{b}{\sqrt{u}}\right)\right)$. Note that

$$\begin{aligned}\mathbf{Pr}\left[\tau_b < u \wedge W(u) > b + h\right] &= \mathbf{Pr}\left[W(u) > b + h\right] \\ &= 1 - \Phi\left(\frac{b}{\sqrt{u}}\right) - \mathbf{Pr}\left[W(u) \in \mathrm{d}b\right].\end{aligned}$$

And

$$\begin{aligned}\mathbf{Pr}\left[\tau_b < u \wedge W(u) < b\right] &= \mathbf{Pr}\left[\tau_b < u\right] \cdot \mathbf{Pr}\left[W(u) < b \mid \tau_b < u\right] \\ &= \frac{1}{2}\mathbf{Pr}\left[\tau_b < u\right] = 1 - \Phi\left(\frac{b}{\sqrt{u}}\right).\end{aligned}$$

Thus, Equation (20) equals to $\mathbf{Pr}\left[W(u) \in \mathrm{d}b\right]$ and $\mathbf{Pr}\left[\tau_b < u \mid W(u) = b\right] = 1$.

35 Kolmogorov-Smirnov Test

In this section, we introduce an application of Brownian Bridge, the Kolmogorov-Smirnov test.

Suppose that $U_1, U_2, \ldots, U_n$ are independently sampled from some distribution $[0,1]$ with CDF F. We would like to check if it is a uniform distribution, i.e., if the F satisfies $F(t) = t$ for every $t \in [0,1]$.

Let $\widehat{F}_n$ be the empirical cumulative distribution function, that is, for $t \in [0,1]$, $\widehat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}[U_i \leq t]$. It then follows from the law of large numbers that

$$\widehat{F}_n(t) \overset{n\to\infty}{\longrightarrow} \mathbf{E}\left[\widehat{F}_n(t)\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbf{Pr}\left[U_i \leq t\right] = F(t).$$

The idea of Kolmogorov-Smirnov test is to monitor the variable $\widehat{F}_n(t) - t$ for every $t \in [0,1]$ and reject the uniformity hypothesis if there exists some t that $\left|\widehat{F}_n(t) - t\right|$ is large. Then our goal is to find a suitable rejection threshold b such that if F is indeed a uniform distribution, the failure probability $\lim_{n\to\infty}\mathbf{Pr}\left[\max_{t\in[0,1]}\left|\widehat{F}_n(t) - t\right| \geq b\right]$ is sufficiently small (i.e., $\leq \frac{1}{100}$). If F is a uniform distribution, for a fixed t, we have

$$\begin{aligned}\mathbf{E}\left[\widehat{F}_n(t)\right] &= F(t) = t; \\ \mathbf{Var}\left[\widehat{F}_n(t)\right] &= \frac{1}{n^2}\sum_{i=1}^{n}\mathbf{Var}\left[\mathbf{1}[U_i \leq t]\right] = \frac{1}{n}\cdot t(1-t).\end{aligned}$$

Let $X_n(t) \triangleq \sqrt{n} \cdot (\widehat{F}_n(t) - t)$ for $t \in [0,1]$. By the Central Limit Theorem, we have $X_n(t) \sim \mathcal{N}(0, t(1-t))$ when $n \to \infty$. For any $0 \le s \le t \le 1$,

$$\begin{aligned}
\mathrm{Cov}(X_n(s), X_n(t)) &= n \cdot \mathrm{Cov}\left(\widehat{F}_n(s) - s, \widehat{F}_n(t) - t\right) \\
&= \frac{1}{n}\mathrm{Cov}\left(\sum_{i=1}^{n} \mathbf{1}[U_i \le s], \sum_{i=1}^{n} \mathbf{1}[U_i \le t]\right) \\
&= \mathrm{Cov}\left(\mathbf{1}[U_1 \le s], \mathbf{1}[U_1 \le t]\right) \\
&= \mathbf{Pr}\left[U_1 \le s, U_1 \le t\right] - \mathbf{Pr}\left[U_1 \le s\right]\mathbf{Pr}\left[U_1 \le t\right] \\
&= s(1-t).
\end{aligned}$$

For any $0 \le t_1 \le t_2 \le \cdots \le t_k \le 1$, let $\Sigma = \left(\mathrm{Cov}\left(X_n(t_i), X_n(t_j)\right)\right)_{i,j}$. It follows from the high-dimensional Central Limit Theorem that

$$(X_n(t_1), X_n(t_2), \ldots, X_n(t_k))^T \xrightarrow{D} \mathcal{N}(\mathbf{0}, \Sigma) \sim (X(t_1), X(t_2), \ldots, X(t_k))^T,$$

where $\{X(t)\}$ is a standard Brownian Bridge. Then using the result in Example 3 in Lecture 11, we have

$$\begin{aligned}
\lim_{n\to\infty} \mathbf{Pr}\left[\max_{t\in[0,1]} \widehat{F}_n(t) - t \ge b\right] &= \mathbf{Pr}\left[\max_{t\in[0,1]} X(t) \ge \sqrt{n}b\right] \\
&= \mathbf{Pr}\left[\tau_{\sqrt{n}b} < 1 \,\middle|\, W(1) = 0\right] = \exp\{-2nb^2\}.
\end{aligned}$$

36 Diffusion

36.1 The Definition of Diffusion

A continuous stochastic process with Markov property is called a diffusion.[36] In other words, a diffusion can be viewed as a Markov process in continuous time with continuous sample paths.

Actually, diffusions can be built up from local Brownian motions in the same way as differentiable functions being built up from local linear functions. Imagine that we want to draw the image of a function f with knowing $f'(t) = e^t$ and $f(0) = 1$. How to do this if you are not allowed to integrate $f'(t)$. A natural idea is to approximate f using segmented linear functions:

- Select a step length h;
- Draw a segment on $[0, h]$ which starts from $(0, f(0))$ with slope $f'(0) = 1$;
- Draw a segment on $[h, 2h]$ which starts from $(h, h + f(0))$ with slope $f'(h) = e^h$;
- ....

When $h \to 0$, our drawing is exactly the image of f. This gives an intuition that a differentiable function can be locally approximated as linear functions.

A diffusion $\{X(t)\}_{t\ge0}$ is the stochastic analog of above process. That is, if we are currently at the position $X(t) = X_t$ and consider the small time interval $[t, t+h]$, the process acts as a $\left(\mu(X_t), \sigma^2(X_t)\right)$ Brownian motion where μ and σ^2 are functions of the position X_t. Let $Z \sim \mathcal{N}(0,1)$ be a standard Gaussian. We can break the process into segments and use these normal random variables to simulate the diffusion:

- $X_h = X_0 + \mu(X_0)h + \sigma(X_0)\sqrt{h} \cdot Z_1$;
- $X_{2h} = X_h + \mu(X_h)h + \sigma(X_h)\sqrt{h} \cdot Z_2$;
- ...,

where each Z_i are independent standard Gaussian. Then for any $k \in \mathbb{N}$, $X_{(k+1)h} - X_{kh} \sim \mathcal{N}\left(\mu(X_{kh})h, \sigma^2(X_{kh})h\right)$.

Thus, when $h \to 0$, we can naturally develop a specification of diffusion: A time homogeneous diffusion can be specified by two functions $\mu(x)$ and $\sigma^2(x)$ which satisfies:

- $\forall t$, $\mathbf{E}\left[X(t+h) - X(t) \mid X(t) = x\right] = \mu(x)h + o(h)$;
- $\forall t$, $\mathbf{Var}\left[X(t+h) - X(t) \mid X(t) = x\right] = \sigma^2(x)h + o(h)$;
- $\forall t$, $\mathbf{E}\left[|X(t+h) - X(t)|^p \,\middle|\, X(t) = x\right] = o(h)$ for $p > 2$.

[36] This is an informal definition of diffusions and it is enough for this course.

Note that

$$\begin{aligned}&\mathbf{Var}\left[X(t+h)-X(t)\mid X(t)=x\right]\\&=\mathbf{E}\left[(X(t+h)-X(t))^2\,\middle|\,X(t)=x\right]-\left(\mathbf{E}\left[X(t+h)-X(t)\mid X(t)=x\right]\right)^2\\&=\mathbf{E}\left[(X(t+h)-X(t))^2\,\middle|\,X(t)=x\right]-(\mu(x)h+o(h))^2.\end{aligned}$$

Thus

$$\mathbf{Var}\left[X(t+h)-X(t)\mid X(t)=x\right]=\sigma^2(x)h+o(h)$$

is equivalent to

$$\mathbf{E}\left[(X(t+h)-X(t))^2\,\middle|\,X(t)=x\right]=\sigma^2(x)h+o(h).$$

Recall that in the analog of differentiable functions, we have $\mathrm{d}f(t)=g(t)\,\mathrm{d}t$ where $g(t)$ is the derivative of f. Similarly, for a diffusion $\{X(t)\}_{t\geq 0}$ specified by $\mu(x)$ and $\sigma^2(x)$, we can write it as [37]

$$\mathrm{d}X(t)=\mu(X(t))\,\mathrm{d}t+\sigma(X(t))\,\mathrm{d}W(t),$$

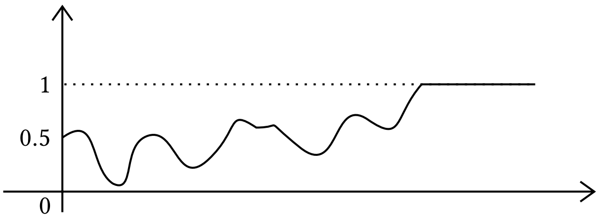
where $\{W(t)\}$ is the standard Brownian motion and $\mathrm{d}W(t)$ can be understood as $\lim_{h\to 0}W(t+h)-W(t)$.

Example 33 (Ornstein-Uhlenbeck Process) Consider a diffusion $\{X(t)\}_{t\geq 0}$ specified by $\sigma(x)=-x$ and $\sigma^2(x)=2$ with $X(0)=0$. This diffusion always has a tendency to 0 since if $X(t)$ is large, $\mu(X(t))$ is also large towards the reverse direction which acts as a spring intuitively. We can write this process as

$$\mathrm{d}X(t)=-X(t)\,\mathrm{d}t+\sqrt{2}\,\mathrm{d}W(t).$$

The process can be used to model the discrete Ehrenfest chain. Suppose we have two boxes with a balls in the first box and b balls in the second box in the initial state. In each round, we choose a ball uniformly at random among the $a+b$ balls and put the chosen ball into the other box. It is more likely to choose the balls in the box with more balls. Thus, this discrete Markov process tends to the equilibrium state where each box has $\frac{a+b}{2}$ balls.

Example 34 (Wright-Fisher Process) Next we consider a stochastic random walk with absorbing boundaries. Let $\mu(x)=0$, $\sigma^2(x)=x(1-x)$ and $X(0)=\frac{1}{2}$. Then this diffusion is jittery around $\frac{1}{2}$ and is more steady around the boundaries.



The process can be used to model the following model of racial reproduction. Assume the total population is N which is invariant over time. At the t-th generation, there is X_t black people and $N-X_t$ white people where X_t is a non-negative random variable. Assume that there is no interracial marriage and the child's race is the same with his or her parents. At the $t+1$-th generation, each person is white w.p. $1-\frac{X_t}{N}$ and is black w.p. $\frac{X_t}{N}$. Assume the race of each individual is independent with other people. If it starts with half white and half black, then we want to ask: Will there be genocide after a long period of time or will the two races tend to keep a balance?

The continuous version of the model is the Wright-Fisher process we just introduced. It is equivalent to ask whether the process tends to keep jittery or be absorbed. Since it seems to be "lazier" when it comes closer to the boundary, the answer of this question is not obvious. In fact, however, after a sufficiently long time, it does reach the boundary.

36.2 Diffusions with a Stopping Time

Let $\{X(t)\}$ be a diffusion specified by $\mu(x)$ and $\sigma^2(x)$, or equivalently,

$$\mathrm{d}X(t)=\mu(X(t))\,\mathrm{d}t+\sigma(X(t))\,\mathrm{d}W(t).$$

The diffusion ends at a random stopping time T. Let $g:\mathbb{R}\to\mathbb{R}$ be a cost function that the diffusion pays at rate $g(x)$ when it arrives at x. We are interested in the average cost of the diffusion, that is, $\mathbf{E}\left[\int_0^T g(X(t))\,\mathrm{d}t\right]$.

Let $w(x)\triangleq\mathbf{E}_x\left[\int_0^T g(X(t))\,\mathrm{d}t\right]=\mathbf{E}\left[\int_0^T g(X(t))\,\mathrm{d}t\,\middle|\,X(0)=x\right]$. We have the following theorem.

[37] This expression is informal as we haven't give a mathematical meaning to the notation $dW(t)$. We will do this in the next lecture, but for now, we can sloppily understand it as a Brownian motion in an infinitisimal time.

Theorem 36.1 The function w satisfies the differential equation:

$$\mu(x)w'(x) + \frac{1}{2}\sigma^2(x)w''(x) = -g(x).$$

Before proving this theorem, we see an application of it. Recall the 1-D random walk with two absorbing barriers $-a$ and b which we discussed in Lecture 7. We used the tool of martingale to show that the average stopping time is ab. If we consider the continuous process, that is, a diffusion specified by $\mu(x) = 0$ and $\sigma^2(x) = 1$ with $X(0) = 0$ and $g(x) \equiv 1$, we have $\mathbf{E}[T] = \mathbf{E}\left[\int_0^T 1 \,\mathrm{d}t\right] = w(0)$ where $T = \min_t \{X(t) = -a \vee X(t) = b\}$. With Theorem 36.1, we have

$$\frac{1}{2}w''(x) = -1.$$

Combining the fact that $w(-a) = w(b) = 0$, we have $w(x) = -(x+a)(x-b)$ and consequentlly $\mathbf{E}[T] = w(0) = ab$.

Then we prove Theorem 36.1.

Proof. [Proof of Theorem 36.1][38] Pick a sufficiently small h such that $\mathbf{Pr}[T > h] = 1 - o(h)$. Then we have

$$\begin{aligned} w(x) &= \mathbf{E}_x\left[\int_0^T g(X(t))\,\mathrm{d}t\right] \\ &= \mathbf{E}_x\left[\int_0^h g(X(t))\,\mathrm{d}t\right] + \mathbf{E}_x\left[\int_h^T g(X(t))\,\mathrm{d}t\right] \\ &= \mathbf{E}_x\left[\int_h^T g(X(t))\,\mathrm{d}t\right] + h\cdot g(x) + o(h). \end{aligned} \tag{21}$$

Note that

$$\begin{aligned} \mathbf{E}_x\left[\int_h^T g(X(t))\,\mathrm{d}t\right] &= \mathbf{E}_x\left[\mathbf{E}_x\left[\int_h^T g(X(t))\,\mathrm{d}t \,\middle|\, X(h)\right]\right] \\ &= \mathbf{E}_x\left[\mathbf{E}_x\left[\int_h^T g(X(t))\,\mathrm{d}t \,\middle|\, X(h), T > h\right]\right] + o(h) \\ &= \mathbf{E}_x[w(X(h))] + o(h). \end{aligned} \tag{22}$$

Using Taylor's expansion, we have

$$\begin{aligned} \mathbf{E}_x[w(X(h))] &= \mathbf{E}_x\left[w(x) + w'(x)(X(h)-x) + \frac{1}{2}w''(x)(X(h)-x)^2\right] + o(h) \\ &= w(x) + h\cdot w'(x)\mu(x) + \frac{h}{2}\cdot w''(x)\sigma^2(x) + o(h). \end{aligned} \tag{23}$$

Combining Equation (21), Equation (22) and Equation (23), we have

$$h\cdot g(x) + h\cdot w'(x)\mu(x) + \frac{h}{2}w''(x)\sigma^2(x) + o(h) = 0.$$

Thus,

$$\begin{aligned} 0 &= \lim_{h\to 0}\frac{h\cdot g(x) + h\cdot w'(x)\mu(x) + \frac{h}{2}w''(x)\sigma^2(x) + o(h)}{h} \\ &= g(x) + \mu(x)w'(x) + \frac{1}{2}\sigma^2(x)w''(x). \end{aligned}$$

□

36.3 Geometric Brownian Motion

Let $\{X(t)\}$ be a (μ, σ^2) Brownian motion, that is, $\mathrm{d}X(t) = \mu\,\mathrm{d}t + \sigma dW(t)$. Define $Y(t) = e^{X(t)}$. Then $\{Y(t)\}$ is called a geometric Brownian motion. Geometric Brownian motion is widely applied to model the stock prices in finance. In fact, we can consider a more generalized situation that $\{Y(t)\}$ is defined by $Y_t = f(X_t)$ where f is strictly monotone and twice differentiable. Then we have the following proposition.

Proposition 36.1 Suppose $\{X(t)\}$ is a diffusion specified by $\mu_X(x)$ and $\sigma_X^2(x)$. Let f be a strictly monotone and twice differentiable function. Define $Y(t) = f(X(t))$. Then $\{Y(t)\}$ is a diffusion specified by $\mu_Y(y)$ and $\sigma_Y^2(y)$ which satisfy

$$\mu_Y(y) = \mu_X(x)f'(x) + \frac{1}{2}\sigma^2(x)f''(x) \quad \textit{and} \quad \sigma_Y^2(y) = (f'(x))^2\sigma_X^2(x)$$

[38] Whether such h exists requires justification and depends on T. Nevertheless, we assume so.

where $x = f^{-1}(y)$.

Proof. For a small h, we have

$$\begin{aligned}
&\mathbf{E}\left[Y(t+h) - Y(t) \mid Y(t) = y\right] \\
=&\mathbf{E}\left[f(X(t+h)) - f(X(t)) \mid X(t) = x\right] \\
=&\mathbf{E}\left[f'(X(t))(X(t+h) - X(t)) + \frac{1}{2}f''(X(t))(X(t+h) - X(t))^2 \,\middle|\, X(t) = x\right] \\
&+ o(h) \\
=&\mu_X(x)f'(x)h + \frac{1}{2}\sigma^2(x)f''(x)h + o(h),
\end{aligned}$$

so that

$$\mu_Y(y) = \lim_{h\to 0} \frac{\mathbf{E}\left[Y(t+h) - Y(t) \mid Y(t) = y\right]}{h} = \mu_X(x)f'(x) + \frac{1}{2}\sigma^2(x)f''(x).$$

Similarly, we have

$$\begin{aligned}
&\mathbf{E}\left[(Y(t+h) - Y(t))^2 \,\middle|\, Y(t) = y\right] \\
=&\mathbf{E}\left[(f(X(t+h)) - f(X(t)))^2 \,\middle|\, X(t) = x\right] \\
=&\mathbf{E}\left[(f'(X(t))(X(t+h)) - X(t))^2 \,\middle|\, X(t) = x\right] + o(h) \\
=&\left(f'(x)\right)^2 \sigma_X^2(x)h + o(h),
\end{aligned}$$

so that

$$\sigma_Y^2(y) = \lim_{h\to 0} \frac{\mathbf{E}\left[(Y(t+h) - Y(t))^2 \,\middle|\, Y(t) = y\right]}{h} = \left(f'(x)\right)^2 \sigma_X^2(x).$$

□

References

[Dur19] Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019. 2

[Mey00] Carl D Meyer. *Matrix analysis and applied linear algebra*, volume 71. SIAM, 2000. 11