# Notes on Probability and Computing

Xu Zhean

January 21, 2022

## Contents

# 1 Events and Probability

A **probability space** is a **measure space** $(\Omega, \mathcal{F}, \mathbb{P})$ consisting of:

- the **sample space** $\Omega$ — a set of outcomes called **sample**;
- the **$\sigma$-algebra** $\mathcal{F}$ — a family of subsets of $\Omega$, called **events**, such that $\Omega \in \mathcal{F}$ and $\mathcal{F}$ is closed under complements (i.e. $\forall A \in \mathcal{F}$, $\Omega \backslash A \in \mathcal{F}$) and countable unions (i.e. $\forall A_i \in \mathcal{F}$, $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$);
- the **probability function** $\mathbb{P} : \mathcal{F} \to [0,1]$ such that $\mathbb{P}(\Omega) = 1$ and $\mathbb{P}$ is **$\sigma$-additive** (i.e. $\mathbb{P}\left(\bigsqcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$).

The motivation behind this complicated definition is that some sets are **non-measurable**, thus mathematicians developed the theory of **measure**. For instance, **Borel set** on real line forms a $\sigma$-algebra which is **generated by** open intervals. **Stieltjes measure** is a **Borel measure** and builds the measure-theoretic foundation of **continuous probability distribution**.

**Lemma 1.1 (Inclusion-exclusion principle)** Let $E_1, \cdots, E_n$ be any n events. Then

$$\mathbb{P}\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{\ell=1}^{n} (-1)^{\ell+1} \sum_{i_1 < i_2 < \cdots < i_\ell} \mathbb{P}\left(\bigcap_{r=1}^{\ell} E_{i_r}\right).$$

Events $E_1, E_2, \cdots, E_n$ are **mutually independent** (simply called **independent** when $k = 2$) if and only if, for any subset $I \subseteq \{1, 2, \cdots, k\}$, $\mathbb{P}\left(\bigcap_{i \in I} E_i\right) = \prod_{i \in I} \mathbb{P}(E_i)$. Note that events $X, Y, Z, \cdots$ are unnecessarily mutually independent when they are pairwise independent.

The **conditional probability** that event $E$ occurs given that event $F$ occurs is $\mathbb{P}(E \mid F) = \mathbb{P}(E \cap F) \,/\, \mathbb{P}(F)$.

**Theorem 1.2 (Law of total probability)** Let events $\bigsqcup_{i=1}^{n} E_i = \Omega$. Then we have $\mathbb{P}(B) = \sum_{i=1}^{n} \mathbb{P}(B \mid E_i) \cdot \mathbb{P}(E_i)$.

**Theorem 1.3 (Bayes's law)** Let events $E_1, E_2, \cdots, E_n$ satisfy $\bigsqcup_{i=1}^{n} E_i = \Omega$. Then we have

$$\mathbb{P}(E_k \mid B) = \frac{\mathbb{P}(E_k \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B \mid E_k) \cdot \mathbb{P}(E_k)}{\sum_{i=1}^{n} \mathbb{P}(B \mid E_i) \cdot \mathbb{P}(E_i)}.$$

In the **Bayesian approach** one starts with a **prior** model, giving some initial value to the model parameters. This model is then modified, by incorporating new observations, to obtain a **posterior** model that captures the new information.

**Exercise 1.6** Using mathematical induction, we have $p_{i,j} = \frac{i-1}{i+j-1} \cdot p_{i-1,j} + \frac{j-1}{i+j-1} \cdot p_{i,j-1} = \frac{i+j-2}{i+j-1} \cdot \frac{1}{i+j-2} = \frac{1}{i+j-1}$.

**Exercise 1.7.b** Let $F_{b_1 b_2 \cdots b_n}$ be the intersection of events $E_i$ ($b_i = 1$) or $\Omega \backslash E_i$ ($b_i = 0$), and $P_k$ be the sum of $\mathbb{P}(F_b)$ where $b$ consists of $k$ one and $n-k$ zero. Then for every $k \geq 1$, we have $\sum_{i=1}^{l} (-1)^{i+1} \binom{k}{i} = 1 + (-1)^{l+1} \binom{k-1}{l} \geq 1$. Multiply both sides by $P_k$ and sum them up. We eventually reach the desired inequality.

**Exercise 1.11.b** $p_3 = p_1 \cdot (1 - p_2) + (1 - p_1) \cdot p_2 \Rightarrow q_3 = 1 - 2p_3 = (1 - 2p_1)(1 - 2p_2) = q_1 q_2$. Is there any underlying motivation?

**Exercise 1.24 (Karger's algorithm)** Let $K$ be the minimum $r$-way cut-set. Considering all $r$-way cut-sets consisting of $r-1$ single vertex, the total size is $m \cdot \binom{n-2}{r-1}$ with an upper bound $(m - |K|) \cdot \binom{n}{r-1}$. It follows that

$$m \cdot \binom{n-2}{r-1} \leq (m - |K|) \cdot \binom{n}{r-1} \quad \Rightarrow \quad 1 - \frac{|K|}{m} \geq \binom{n-2}{r-1} \binom{n}{r-1}^{-1} = \frac{(n-r+1)(n-r)}{n(n-1)}.$$

The probability that $K$ survives all the $n - r$ iterations is at least

$$\prod_{i=0}^{n-r-1} \frac{(n-i+1-r)(n-i-r)}{(n-i)(n-i-1)} = r \cdot \binom{n}{r-1}^{-1} \binom{n-1}{r-1}^{-1}$$

and its reciprocal is the maximum possible number of minimum cardinality of $r$-way cut-sets.

# 2 Discrete Random Variables and Expectation

A **(real-valued) random variable** $X$ on a sample space $\Omega$ is a **measurable function** $X : \Omega \to \mathbb{R}$, and a **discrete random variable** is one which may take on only a countable number of distinct values. "$X = a$" represents the set $\{s \in \Omega \mid X(s) = a\}$, and we denote the probability of that event by $\mathbb{P}(X = a) = \sum_{s \in \Omega : X(s) = a} \mathbb{P}(s)$.

Random variables $X_1, X_2, \cdots, X_n$ are **mutually independent** (simply called **independent** when $k = 2$) if and only

if, for any subset $I \subseteq \{1, 2, \cdots, k\}$ and any values $x_i$ $(i \in I)$, $\mathbb{P}(\bigcap_{i\in I}(X_i = x_i)) = \prod_{i\in I} \mathbb{P}(X_i = x_i)$.

The **expectation** of a discrete random variable $X$, denoted by $\mathbb{E}[X]$, is given by $\mathbb{E}[X] = \sum_i i \cdot \mathbb{P}(X = i)$. Note that the infinite series needs to be **absolutely convergent** (i.e. rearrangements do not change the value of the sum).

**Theorem 2.1 (Linearity of expectation)** For discrete random variables $X_1, X_2, \cdots, X_n$ with finite expectations and any contants $c_1, c_2, \cdots, c_n$, we have $\mathbb{E}[\sum_{i=1}^n c_i X_i] = \sum_{i=1}^n c_i \mathbb{E}[X_i]$.

*Proof.* Observe that we only need to prove the following two cases:

$$\begin{aligned}
\mathbb{E}[X + Y] &= \sum_i \sum_j (i + j) \cdot \mathbb{P}((X = i) \cap (Y = j)) \\
&= \sum_i i \sum_j \mathbb{P}((X = i) \cap (Y = j)) + \sum_j j \sum_i \mathbb{P}((X = i) \cap (Y = j)) = \mathbb{E}[X] + \mathbb{E}[Y], \\
\mathbb{E}[cX] &= \sum_i i \cdot \mathbb{P}(cX = j) = c \cdot \sum_j (j/c) \cdot \mathbb{P}(X = j/c) = c \cdot \sum_k k \cdot \mathbb{P}(X = k) = c \cdot \mathbb{E}[X].
\end{aligned}$$

When there are countably infinite variables, the situation becomes more subtle. We will discuss it later. ◀

**Theorem 2.2 (Jensen's inequality)** If $f$ is a convex function, then $\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$.

*Proof.* Assume that $f$ has a Taylor expansion. Let $\mu = \mathbb{E}[X]$. By Taylor's theorem, there is a value $c$ such that

$$f(x) = f(\mu) + f'(\mu)(x - \mu) + \frac{f''(c)(x - \mu)^2}{2} \geq f(\mu) + f'(\mu)(x - \mu)$$

Taking expectations of both sides

$$\mathbb{E}[f(X)] \geq \mathbb{E}[f(\mu) + f'(\mu)(X - \mu)] = \mathbb{E}[f(\mu)] + f'(\mu)(\mathbb{E}[X] - \mu) = f(\mu) = f(\mathbb{E}[X])$$

An alternative proof will be presented in Exercise 2.10. ◀

Define **conditional expectation** $\mathbb{E}[Y \mid Z = z] = \sum_y y \cdot \mathbb{P}(Y = y \mid Z = z)$ and $\mathbb{E}[Y \mid Z]$ as a random variable $f(Z)$ that takes on the value $\mathbb{E}[Y \mid Z = z]$ when $Z = z$.

**Theorem 2.3 (Law of total expectation)** For any random variables $X$ and $Y$,

$$\mathbb{E}[X] = \sum_y \mathbb{P}(Y = y) \cdot \mathbb{E}[X \mid Y = y] = \mathbb{E}[\mathbb{E}[X \mid Y]].$$

A **Bernoulli** random variable $X$ takes 1 with probability $p$ and 0 with probability $1 - p$. A **binomial** random variable $X$ with parameters $n$ and $p$, denoted by $B(n, p)$, is defined by **probability distribution** $\mathbb{P}(X = k) = \binom{n}{k} \cdot p^k (1 - p)^{n-k}$, $n = 0, 1, \cdots, n$. Its expectation is $np$.

A **geometric** random variable $X$ with parameter $p$ is defined by probability distribution $\mathbb{P}(X = n) = (1-p)^{n-1}p$, $n = 1, 2, \cdots$. Its expectation is $1/p$. Geometric random variables are **memoryless**, that is, one ignores past failures as distribution does not change. Formally, we have the following statement.

**Lemma 2.4 (Memorylessness)** Let $X$ be a geometric random variable with parameter $p$. Then, for $n > 0$,

$$\mathbb{P}(X = n + k \mid X > k) = \mathbb{P}(X = n).$$

**Lemma 2.5** Let $X$ be a discrete random variable that takes on only nonnegative integer values. Then,

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} k \cdot \mathbb{P}(X = k) = \sum_{1 \leq i \leq k} \mathbb{P}(X = k) = \sum_{i=1}^{\infty} \mathbb{P}(X \geq i)$$

**Exercise 2.7** (a) By the memoryless property, we can ignore the case of $X > 1$ and $Y > 1$, thus $\mathbb{P}[X = Y] = \mathbb{P}[(X = 1) \cap (Y = 1)] / (1 - \mathbb{P}[(X > 1) \cap (Y > 1)])$. (b) Consider the first **trial**, and we can get an equation of $\mathbb{E}[\max(X, Y)]$. (c) Construct a **bernoulli trial** that success when there is at least one of two trials success. Its distribution of the first successful time provides the answer. (d) is the same as (a).

**Exercise 2.14 (Negative binomial distribution)** the $k$-th successful time. $\mathbb{P}(X = n) = \binom{n-1}{k-1} p^k (1-p)^{n-k}$, $n \geq k$.

**Exercise 2.16.b** Break the sequence of flips up into disjoint blocks of $\lfloor \log_2 n - 2 \log_2 \log_2 n \rfloor$ consecutive flips. For sufficiently large $n$, the probability is less than

$$\left(1 - 2^{\log_2 n - 2 \log_2 \log_2 n}\right)^{\frac{n}{\log_2 n - 2 \log_2 \log_2 n}} < \left(1 - \frac{n}{\log_2^2 n}\right)^{\frac{n}{\log_2^2 n} \cdot \log_2 n} < e^{-\ln n} = \frac{1}{n}.$$

**Exercise 2.29** If $\{X_n\}$ is a sequence of random variable satisfying $X_n \to X$ **almost surely** (i.e. except possibly on an event of zero probability) then **(monotone convergence)** if $0 \leq X_n \leq X_{n+1}$ for all $n$ almost surely, then

$\mathbb{E}[X_n] \to \mathbb{E}[X]$; **(dominated convergence)** if $|X_n| \le Y$ for all $n$ almost surely and $\mathbb{E}[Y]$ is finite, then $\mathbb{E}[X_n] \to \mathbb{E}[X]$.

Let $Z_n = \sum_{i=0}^{n} X_n$. We have $Z_n \to \sum_{i=0}^{\infty} X_n$ and $|Z_n| \le \sum_{i=0}^{\infty} |X_n|$ whose expectation is finite ($\mathbb{E}[\sum_{i=0}^{\infty} |X_n|] = \sum_{i=0}^{\infty} \mathbb{E}[|X_n|] < \infty$ is a consequence of monotone convergence). By dominated convergence, it follows that

$$\sum_{j=0}^{n} \mathbb{E}[X_j] = \mathbb{E}\left[\sum_{j=0}^{n} X_j\right] = \mathbb{E}[Z_n] \to \mathbb{E}[Z] = \mathbb{E}\left[\sum_{j=0}^{\infty} X_j\right], \qquad n \to \infty.$$

**Exercise 2.32** For $i > m$, $\mathbb{P}(E_i) = \frac{1}{n} \cdot \frac{m}{i-1}$. Putting this all together, we get $\mathbb{P}(E) = \frac{m}{n} \sum_{j=m+1}^{n} \frac{1}{j-1}$. Then,

$$\frac{m}{n} \cdot \ln\left(\frac{n}{m}\right) = \frac{m}{n} \cdot \int_{m+1}^{n+1} \frac{\mathrm{d}x}{x-1} \le \mathbb{P}(E) \le \frac{m}{n} \cdot \int_{m}^{n} \frac{\mathrm{d}x}{x-1} = \frac{m}{n} \cdot \ln\left(\frac{n-1}{m-1}\right)$$

Note that $m(\ln n - \ln m)/n$ is maximized when $m = n/e$ and $\mathbb{P}(E) \ge 1/e$ for this choice of $m$.

# 3 Moments and Deviations

**Theorem 3.1 (Markov's Inequality)** Let $X$ be a random variable with only nonnegative values. Then, for all $a > 0$,

$$\mathbb{P}(X \ge a) \le \frac{\mathbb{E}[X]}{a}$$

*Proof.* For $a > 0$, let $I = 1$ (if $X \ge a$) or 0 (otherwise), and note that $I \le X/a$. Taking expectaions on both sides, thus yields $\mathbb{P}(X \ge a) = \mathbb{E}[I] =\le \mathbb{E}[X/a] = \mathbb{E}[X]/a$. ◀

The **$k$-th moment** of a random variable $X$ is $\mathbb{E}[X^k]$. The **variance** of random variable $X$ is defined as $\mathsf{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$, and the **standard deviation** of a random variable $X$ is $\sigma[X] = \sqrt{\mathsf{Var}[X]}$. The **convariance** of two random variables $X$ and $Y$ is $\mathsf{Cov}(X, y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$, and we have

**Lemma 3.2** For any two random variables $X$ and $Y$, $\mathsf{Var}[X + Y] = \mathsf{Var}[X] + \mathsf{Var}[Y] + 2 \cdot \mathsf{Cov}(X, Y)$.

**Lemma 3.3** For any two independent random variables $X$ and $Y$, $\mathbb{E}[X \cdot Y] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$. (opposite does not hold)

**Corollary 3.4** If $X$ and $Y$ are independent random variables, then $\mathsf{Cov}(X, Y) = 0$.

**Theorem 3.5 (Linearity of variance)** Let $X_1, X_2, \cdots, X_n$ be mutually independent random variables. Then

$$\mathsf{Var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathsf{Var}[X_i]$$

For example, a Bernoulli trial with success probability $p$ has variable $p(1-p)$, therefore the variance of a binomial random variable $X$ with parameters $n$ and $p$ is $np(1-p)$.

**Theorem 3.6 (Chebyshev's inequality)** Let $X$ be a random variable. Then, for any $a > 0$,

$$\mathbb{P}\left(|X - \mathbb{E}[X]| \ge a\right) \le \frac{\mathsf{Var}[X]}{a^2}$$

*Proof.* We can apply Markov's inequality to prove:

$$\mathbb{P}(|X - \mathbb{E}[X]| \ge a) = \mathbb{P}((X - \mathbb{E}[X])^2 \ge a^2) \le \frac{\mathbb{E}[(X - \mathbb{E}[X])^2]}{a^2} = \frac{\mathsf{Var}[X]}{a^2}$$

A useful variant of Chebyshev's inequality is to substitute $a$ with $t \cdot \sigma[X]$ ($t \ge 1$). ◀

The **median** of random variable $X$ is defined to be any value $m$ such that $\mathbb{P}(X \le m) \ge 1/2$ and $\mathbb{P}(X \ge m) \ge 1/2$.

**Theorem 3.7** For any random variable $X$ with finite expectaion $\mathbb{E}[X]$ and finite median $m$,

- the expectaion $\mathbb{E}[X]$ is the value of $c$ that minimizes the expression $\mathbb{E}[(X - c)^2]$.
- the median $m$ is the value of $c$ that minimizes the expression $\mathbb{E}[|X - c|]$.

**Corollary 3.8** $|\mu - m| = |\mathbb{E}[X] - m| = |\mathbb{E}[X - m]| \le \mathbb{E}[|X - m|] \le \mathbb{E}[|X - \mu|] \le \sqrt{\mathbb{E}[(X - \mu)^2]} = \sigma$.

**Exercise 3.10** By the memoryless property, we have $\mathbb{E}[X^k] = (1-p) \cdot \mathbb{E}[(X+1)^k] + p$. A clever way is to use falling factorial, and we will get $\mathbb{E}[X^{\underline{k}}] = k! \cdot (1-p)^{k-1} \cdot p^{-k}$, $\mathbb{E}[X^n] = \sum_{k=0}^{n} \left\{ {n \atop k} \right\} \cdot \mathbb{E}[X^{\underline{k}}]$.

**Exercise 3.15** $\mathsf{Var}[\sum_i X_i] = \sum_i \mathsf{Var}[X_i] + 2 \sum_i \sum_j \mathsf{Cov}(X_i, X_j)$. If $\mathbb{E}[X_i X_j] = \mathbb{E}[X_i]\mathbb{E}[X_j]$, then $\mathsf{Cov}(X_i, X_j) = 0$.

**Exercise 3.18 (Cantelli's inequality)** Let $Y = X - \mathbb{E}[X]$, and it follows that $\mathbb{E}[Y] = 0$ and $\mathsf{Var}[Y] = \mathbb{E}[Y^2] = \sigma^2$.

For any $\lambda, u > 0$ (taking $u = \sigma^2/\lambda$ in last step),

$$\mathbb{P}(Y \geq \lambda) = \mathbb{P}(Y + u \geq \lambda + u) \leq \mathbb{P}\left((Y+u)^2 \geq (\lambda+u)^2\right) \leq \frac{\mathbb{E}[(Y+u)^2]}{(\lambda+u)^2} = \frac{\sigma^2+u^2}{(\lambda+u)^2} = \frac{\sigma^2}{\lambda^2+\sigma^2}$$

**Exercise 3.26 (The weak law of large numbers)** Apply Chebyshev's Inequality, thus for any $\varepsilon > 0$ we have

$$\mathbb{P}\left(\left|\frac{X_1 + X_2 + \cdots + X_n}{n}\right| - \mu\right) \leq \frac{\sigma^2}{\varepsilon^2 \cdot n} \to 0, \qquad \text{as } n \to \infty.$$

# 4 Chernoff and Hoeffding Bounds

The **moment generating function** of a random variable $X$ is $M_X(t) = \mathbb{E}[e^{tX}]$, and we are interested in its existence and properties near zero. It captures all of the moments of $X$,

**Theorem 4.1** Let $X$ be a random variable. Assuming that we can exchange the expectation and differentiation operands, then $M_X^{(n)}(t) = \mathbb{E}[X^n e^{eX}]$. Computed at $t = 0$, we have $M_X^{(n)}(0) = \mathbb{E}[X^n]$.

**Theorem 4.2** Let $X$ and $Y$ be two random variables. If $M_X(t) = M_Y(t)$ for all $t \in (-\delta, \delta)$ for some $\delta > 0$, then $X$ and $Y$ have the same distribution.

**Theorem 4.3** If $X$ and $Y$ are independent random variables, then $M_{X+Y}(t) = M_X(t)M_Y(t)$.

Bounds derived from following approach are called **Chernoff bounds**. Generally, for any $t > 0$,

$$\mathbb{P}(X \geq a) = \mathbb{P}(e^{tX} \geq e^{ta}) \leq \frac{\mathbb{E}[e^t X]}{e^{ta}} \Rightarrow \mathbb{P}(X \geq a) \leq \min_{t>0} \frac{\mathbb{E}[e^{tX}]}{e^{ta}}.$$

We can select an appropriate value of $t$ to obtain the best possible bounds. Similarly, for any $t < 0$,

$$\mathbb{P}(X \leq a) = \mathbb{P}(e^{tX} \geq e^{ta}) \leq \frac{\mathbb{E}[e^t X]}{e^{ta}} \Rightarrow \mathbb{P}(X \leq a) \leq \min_{t<0} \frac{\mathbb{E}[e^{tX}]}{e^{ta}}.$$

Let $X_1, \cdots, X_n$ be a sequence of independent Bernoulli trials with $\mathbb{P}(X_i = 1) = p_i$. The sum $X = \sum_{i=1}^n X_i$ forms a **Poisson binomial distribution**. Let $\mu = \mathbb{E}[X] = \sum_{i=1}^n p_i$, and we have

$$M_X(t) = \prod_{i=1}^n M_{X_i}(t) = \prod_{i=1}^n \left(1 + p_i \cdot (e^t - 1)\right) \leq \prod_{i=1}^n e^{p_i \cdot (e^t - 1)} = e^{(e^t - 1)\cdot \mu}$$

**Theorem 4.4** Let $X$ be a Poisson binomial distribution, and $\mu = \mathbb{E}[X]$. Then the following Chernoff bounds hold:

$$\mathbb{P}(X \geq (1+\delta)\cdot\mu) \leq \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^\mu, \quad \text{for } \delta > 0; \qquad \mathbb{P}(X \leq (1-\delta)\cdot\mu) \leq \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^\mu, \quad \text{for } \delta > 0.$$

**Corollary 4.5** Let $X$ be a Poisson binomial distribution. Then, for $0 < \delta < 1$,

$$\mathbb{P}(X \geq (1+\delta)\cdot\mu) \leq \exp\left(-\mu\delta^2 \cdot (2\ln 2 - 1)\right), \qquad \mathbb{P}(X \geq (1-\delta)\cdot\mu) \leq \exp\left(-\mu\delta^2/2\right).$$

The coefficient $(2\ln 2 - 1)$ and $1/2$ are derived from $\min((1+\delta)\cdot\ln(1+\delta)/\delta^2 - 1/\delta)$ in $\delta \in (0,1)$ and $\delta \in (-1, 0)$.

**Theorem 4.6** Let $X$ be a binomial distribution where $p = 1$. Then,

$$\mathbb{P}(X \geq (1+\delta)\cdot\mu) \leq \exp\left(-\delta^2\mu\right), \quad \text{for } \delta > 0; \qquad \mathbb{P}(X \leq (1-\delta)\cdot\mu) \leq \exp\left(-\delta^2\mu\right), \quad \text{for } 0 < \delta < 1.$$

**Lemma 4.7 (Hoeffding's lemma)** Let $X$ be a random variable such that $\mathbb{P}(X \in [a, b]) = 1$. Then for every $\lambda > 0$,

$$\mathbb{E}[e^{\lambda X}] \leq \exp\left(\lambda\mu + \frac{\lambda^2 \cdot (b-a)^2}{8}\right), \qquad \text{where } \mu = \mathbb{E}[X].$$

*Proof.* Assum $\mathbb{E}[X] = 0$ and $a \leq 0 \leq b$. Since $e^{\lambda x}$ is a convex function, we have

$$\mathbb{E}[e^{\lambda X}] \leq \mathbb{E}\left[\frac{b-X}{b-a}\cdot e^{\lambda a}\right] + \mathbb{E}\left[\frac{X-a}{b-a}\cdot e^{\lambda b}\right] = \frac{b}{b-a}\cdot e^{\lambda a} - \frac{a}{b-a}\cdot e^{\lambda b} = e^{g(u)}, \quad \text{where } u = \lambda\cdot(b-a).$$

Then $g(u) = -c\cdot u + \ln(1 - c + c\cdot e^u)$ with $c = -a\,/\,(b-a)$. We can verify that $g(0) = g'(0) = 0$ and $g''(u) \leq 1/4$. By Taylor's theorem, for any $u > 0$ there is a $u_0 \in [0, u]$ such that $g(u) = g(0) + u\cdot g(0) + u^2\cdot g''(u_0)\,/\,2 \leq u^2\,/\,8$. ◀

**Theorem 4.8 (Hoeffding bound)** Let $X_1, \cdots, X_n$ be independent random variables with $\mathbb{E}[X_i] = \mu_i$ and $\mathbb{P}(a_i \leq X \leq b_i) = 1$ for constants $a_i$ and $b_i$. Then

$$\mathbb{P}\left(\sum_{i=1}^n X_i - \sum_{i=1}^n \mu_i \geq \varepsilon\right) \leq \exp\left(\frac{-2\cdot\varepsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

*Proof.* Let $Z_i = X_i - \mu_i$ and $Z = \sum_{i=1}^n Z_i$. For any $\lambda > 0$, by Chernoff's approach, we have

$$\mathbb{P}(Z \geq \varepsilon) = \mathbb{P}\left(e^{\lambda Z} \geq e^{\lambda\varepsilon}\right) \leq \frac{\mathbb{E}[e^{\lambda Z}]}{e^{\lambda\varepsilon}} = \frac{\prod_{i=1}^n \mathbb{E}[e^{\lambda Z_i}]}{e^{\lambda\varepsilon}} \leq \exp\left(-\lambda\varepsilon + \lambda^2 \cdot \sum_{i=1}^n \frac{(b_i - a_i)^2}{8}\right)$$

Let $\lambda = 4\varepsilon/(\sum_{i=1}^n (b_i - a_i)^2)$, and it follows Hoeffding bound. ◀

**Packet routing in sparse networks**

# 5 Balls, Bins, and Random Graphs

# 6 The Probabilistic Method

# 7 Markov Chains and Random Walks

A **stochastic process** $\mathbf{X} = \{X(t) : t \in T\}$ is a collection of random variables $X(t)$ (interchangeably, $X_t$), the **state** of process at time $t$. Assume stochastic processes below are discrete time and discrete space.

A discrete time stochastic process $X_0, X_1, X_2, \cdots$ is a (time-homogeneous) **Markov chain** if and only if $\mathbb{P}(X_t = a_t \mid X_{t-1} = a_{t-1}, X_{t-2} = a_{t-2}, \cdots, X_0 = a_0) = \mathbb{P}(X_t = a_t \mid X_{t-1} = a_{t-1}) = P_{a_{t-1},a_t}$. The state $X_t$ only depends on the previous state $X_{t-1}$. This is called the **Markov property** or **memoryless property**, and we say that chain is **Markovian**. The transition probabilities form a one-step **transition matrix** $P$, and for all $i$, $\sum_{j\geq 0} P_{i,j} = 1$. Let $p(t) = (p_0(t), p_1(t), p_2(t), \cdots)$ represents the distribution of the state at time $t$, and we have $p(t) = p(t-1) \cdot P$.

In the finite case, it is equivalent to analyzing the connectivity structure of the directed graph (i.e. strongly connected component). It follows several trivial definitions and conclusions. State $j$ is **accessible** from state $i$ if $\exists n \geq 0, P^n_{i,j} > 0$. If two states $i$ and $j$ are accessible from each other, we say that they **communicate**. A Markov chain is **irreducible** if all states belong to one communicating class. Let $r^t_{i,j} = \mathbb{P}(X_t = j \wedge \forall 1 \leq s \leq t-1, X_s \neq j \mid X_0 = i)$. A state is **recurrent** if $\sum_{t\geq 1} r^t_{i,i} = 1$, and **transient** otherwise. Let $h_{i,j} = \sum_{t\geq 1} t \cdot r^t_{i,j}$. A recurrent state $i$ is **positive recurrent** if $h_{i,i} < \infty$. Otherwise, it is **null recurrent** (this occurs only in infinite case).

**Lemma 7.1** In a finite Markov chain, at least one state is recurrent, and all recurrent states are positive recurrent.

A state $j$ in a discrete time Markov chain is **periodic** if there exists and integer $\Delta > 1$ such that $\mathbb{P}(X_{t+s} = j \mid X_t = j) = 0$ unless $\Delta \mid s$. A discrete time Markov chain is periodic if any state in the chain is periodic. A state of chain that is not periodic is **aperiodic**. An aperiodic, positive recurrent state is an **ergodic** state. A Markov chain is ergodic if all its states are ergodic.

A **stationary distribution** $\pi$ of a Markov chain is a probability distribution $\pi$ such that $\pi = \pi \cdot P$.

**Theorem 7.2** Any finite, irreducible, and ergodic Markov chain has the following properties:

- the chain has a unique stationary distribution $\pi = (\pi_0, \pi_1, \cdots, \pi_n)$;
- for all $j$ and $i$, the limit $\lim_{t\to\infty} P^t_{j,i}$ exists and it is independent of $j$;
- $\pi_i = \lim_{t\to\infty} P^t_{j,i} = 1/h_{i,i}$.

**Lemma 7.3** For any irreducible, ergodic Markov chain and for any state $i$, the limit $\lim_{t\to\infty} P^t_{i,i} = 1/h_{i,i}$.

The expected time between visits to $i$ is $h_{i,i}$ and therefore state $i$ is visited $1/h_{i,i}$ of the time. Thus, if $\lim_{t\to\infty} P^t_{i,i}$ exists, it must be $1/h_{i,i}$. In fact, any finite Markov chain has a stationary distribution; but in the case of periodic state $i$, the stationary probability $\pi_i$ is not the limiting probability of being in $i$ (which does not exist) but instead just the long-term frequency of visiting state $i$. We can compute the stationary distribution by solving $\pi \cdot P = \pi$.

Considering the **cut-sets** of Markov chain, for any state $i$ of the chain, $\sum_{j=0}^n \pi_j P_{j,i} = \pi_i = \pi_i \sum_{j=0}^n P_{i,j}$. It follows

**Theorem 7.4** Let $S$ be a set of states of a finite, irreducible, aperiodic Markov chain. In the stationary distribution, the probability that the chain leaves the set $S$ equals the probability that it enters $S$.

**Theorem 7.5** Consider a finite, irreducible, and ergodic Markov chain with transition matrix $P$. If there are nonnegative numbers $\pi = (\pi_0, \cdots, \pi_n)$ such that $\sum_{i=0}^n \pi_i = 1$ and if, for any pair of states $i, j$, $\pi_i P_{i,j} = \pi_j P_{j,i}$, then $\pi$ is the stationary distribution corresponding to $P$.

Chains that satisfy the condition $\pi_i P_{i,j} = \pi_j P_{j,i}$ are called **time reversible**.

**Theorem 7.6** Any irreducible aperiodic Markov chain belongs to one of the following two categories:

- the chain is ergodic – for any pair of states $i$ and $j$, the limit $\lim_{t\to\infty} P^t_{j,i}$ exists and is independent of $j$, and the chain has a unique stationary distribution $\pi_i = \lim_{t\to\infty} P^t_{j,i} > 0$; or
- no state is positive recurrent – for all $i$ and $j$, $\lim_{t\to\infty} P^t_{j,i} = 0$, and the chain has no stationary distribution.

A **random walk** on $G$ is a Markov chain, where $P_{i,j} = 1 \,/\, \deg(i)$.

**Lemma 7.7** A random walk on an undirected graph $G$ is aperiodic if and only if $G$ is not bipartite.

**Theorem 7.8** A random walk on $G$ converges to a stationary distribution $\pi$, where $\pi_v = \deg(v) \,/\, 2|E|$.

Denote **hitting time** $h_{u,v}$ the expected time to reach state $v$ when starting at state $u$. The **cover time** of a graph $G$ is the maximum over all nodes $v \in V$ of the expected time to visit all of the nodes by a random walk starting from $v$.

**Lemma 7.9** If $(u, v) \in E$, the commute time $h_{u,v} + h_{v,u}$ is at most $2|E|$.

*Proof.* We can view the random walk on $G$ as a Markov chain with states of $2|E|$ directed edges. Since it is a **doubly** stochastic (the sum of the entries in each column is 1), it has a uniform stationary distribution. An upper bound for $h_{u,v} + h_{v,u}$ is the interval of visiting time of edge $(u, v)$. ◀

**Lemma 7.10** The cover time of $G = (V, E)$ is bounded above by $2|E|(|V| - 1)$.

**Theorem 7.11 (Matthews' theorem)** The cover time $C_G$ of $G = (V, E)$ with $n$ vertices is bounded by

$$C_G \le H(n-1) \max_{u,v\in V: u\neq v} h_{u,v}.$$

*Proof.* Consider a random permutation $\{Z_1, Z_2, \cdots, Z_n\}$. Assume that we have computed the expected time visiting all of $\{Z_1, \cdots, Z_{j-1}\}$. If $Z_j$ is not the first visiting node in $\{Z_1, \cdots, Z_j\}$, it contributes nothing. Otherwise, it contributes to the answer with the probability of $1/j$. ◀

**Parrondo's paradox** shows that two losing games can be combined to make a winning game.

#### A random algorithm for 3-Satisfiability

**Exercise 7.13** (a) $\mathbb{P}(X_k \mid X_{k+1}, \cdots, X_m) = \mathbb{P}(X_k, \cdots, X_m) \,/\, \mathbb{P}(X_{k+1}, \cdots, X_m) = \mathbb{P}(X_k)\mathbb{P}(X_{k+1} \mid X_k)\mathbb{P}(X_{k+2}, \cdots, X_m \mid X_k, X_{k+1}) \,/\, \mathbb{P}(X_{k+1})\mathbb{P}(X_{k+2}, \cdots, X_m \mid X_{k+1}) = \mathbb{P}(X_k)\mathbb{P}(X_{k+1} \mid X_k) \,/\, \mathbb{P}(X_{k+1})$, thus it is Markovian. (b) Let $\mathbb{P}(X_k = j) = \pi_j$ and $\mathbb{P}(X_{k+1} = j) = \pi_j$. (c) From part (b), we have $\pi_i Q_{i,j} = \pi_j P_{j,i}$. Then $Q_{i,j} = P_{i,j}$.

**Exercise 7.17** Recall that we let $r^t_{0,0}$ be the probability that the first return to 0 from 0 is at time t. Then

$$\sum_{t=0}^{\infty} r^t_{0,0} = \sum_{n=0}^{\infty} C_n p^n (1-p)^{n+1} = (1-p) \cdot \frac{1 - \sqrt{1-4p(1-p)}}{2p(1-p)}, \qquad \text{since } \sum_{n=0}^{\infty} C_n x^n = \frac{1-\sqrt{1-4x}}{2x}.$$

Hence the chain is recurrent if and only if $p \le 1/2$. Let $h^t_{0,0}$ be the expectation. Then

$$\sum_{t=0}^{\infty} h^t_{0,0} = \sum_{n=0}^{\infty} (2n+2) C_n p^n (1-p)^{n+1} = \frac{2(1-p)}{\sqrt{1-4p(1-p)}}, \qquad \text{since } (n+1)C_n = \binom{2n}{n} \text{ and } \sum_{n=0}^{\infty} \binom{2n}{n} = \frac{1}{\sqrt{1-4x}}.$$

Hence $h^t_{0,0}$ is finite when $p < 1/2$ and is infinite when $p = 1/2$.

**Exercise 7.18 (Random walk on $\mathbb{Z}^d$)** Let $P_d(n)$ be the probability that one returns to origin at time $n$. Random walk on $\mathbb{Z}^d$ is recurrent if and only if $\sum_{n\ge 1} P_d(2n)$ is unbound. In case of $d = 2$, we can transform Manhattan distance into Chebyshev distance (i.e. $(x, y) \to (x + y, x - y)$), thus it becomes two independent random walks on $\mathbb{Z}$. Then

$$\sum_{n=1}^{\infty} P_2(2n) = \sum_{n=1}^{\infty} P_1(2n)^2 = \sum_{n=1}^{\infty} \left( \frac{\binom{2n}{n}}{2^{2n}} \right)^2 \simeq \sum_{n=1}^{\infty} \frac{1}{\pi n} = \infty, \qquad \text{since } n! \sim n^n e^{-n} \sqrt{2\pi n} \text{ (Stirling's formula)}$$

In the case of $d = 3$, we have $P_3(n) = \Theta(n^{-3/2})$. The explicitly expectation formula can be derived by Fourier analysis.

**Exercise 7.22** Formulate a new Markov chain with $n^2$ states of the form $(i, j)$. By Lemma 1.9, $h_{u,v} \le 4m^2$, and we can construct a length $O(n)$ path from $(i, j)$ to $(i, i)$, which gives us an upper bound of $O(m^2 n)$.

**Exercise 7.24 (Lollipop graph)** (a) We need to travel from $v$ to $u$ first, and then travel around the clique. Thus $C_G = h_{v,u} + c_u = \Theta(n^2) + \Theta(n \log n) = \Theta(n^2)$. (b) $h_{u,v} \le C_G \le h_{u,v} + c_u$, thus $C_G = \Theta(h_{u,v}) = \Theta(n^3)$.

**Exercise 7.30 (Random walk on hypercube)** Let $f_i$ be the hitting time when there is exactly $i$ bits differ. Then,

$$f_i = \frac{i}{n} \cdot f_{i-1} + \frac{n-i}{n} \cdot f_{i+1} + 1 \Rightarrow i \cdot (f_i - f_{i-1}) = (n-i) \cdot (f_{i+1} - f_i) + n.$$

Denote the difference $f_i - f_{i-1}$ by $g_i$. Then we have $g_n = 1$, and $i \cdot g_i = (n-i) \cdot g_{i+1} + n$. Expand formula, it follows

$$g_1 = \binom{n-1}{1} \cdot g_2 + \binom{n}{1} = \binom{n-1}{2} \cdot g_3 + \binom{n}{2} + \binom{n}{1} = \cdots = \sum_{i=1}^{n} \binom{n}{i} = 2^n - 1.$$

In addition, $g_n \leq g_{n-1} \leq \cdots \leq g_2 \leq \frac{g_1}{n-1}$. Thus, $f_n = \sum_{i=1}^{n} g_i = \Theta(2^n)$ and the cover time is $O(N \log N)$.