Notes on Probability and Computing

Xu Zhean

January 16, 2022

Contents

1 Events and Probability

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

- the **sample space** Ω — a set of outcomes called **sample**;
- the **σ-algebra** $\mathcal{F}$ — a family of subsets of Ω, 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** $\mathbf{P} : \mathcal{F} \to [0,1]$ such that $\mathbf{P}(\Omega) = 1$ and $\mathbf{P}$ is **σ-additive** (i.e. $\mathbf{P}\left(\bigsqcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbf{P}(A_i)$).

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 σ-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

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

Events $E_1, E_2, \cdots, E_n$ are **mutually independent** (simply called **independent** when $k = 2$) if and only if, for any subset $I \subseteq \{1, 2, \cdots, k\}$, $\mathbf{P}\left(\bigcap_{i \in I} E_i\right) = \prod_{i \in I} \mathbf{P}(E_i)$. 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 $\mathbf{P}(E \mid F) = \mathbf{P}(E \cap F) \,/\, \mathbf{P}(F)$.

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

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

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

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 $\mathbf{P}(F_b)$ 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 Ω 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 $\mathbf{P}(X = a) = \sum_{s \in \Omega : X(s) = a} \mathbf{P}(s)$.

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 $(i \in I)$, $\mathbf{P}(\bigcap_{i \in I}(X_i = x_i)) = \prod_{i \in I} \mathbf{P}(X_i = x_i)$.

The **expectation** of a discrete random variable X, denoted by $\mathbf{E}[X]$, is given by $\mathbf{E}[X] = \sum_i i \cdot \mathbf{P}(X = i)$. 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 $\mathbf{E}[\sum_{i=1}^n c_i X_i] = \sum_{i=1}^n c_i \mathbf{E}[X_i]$.

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

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

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 $\mathbf{E}[f(X)] \geq f(\mathbf{E}[X])$.

Proof. Assume that f has a Taylor expansion. Let $\mu = \mathbf{E}[X]$. 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

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

An alternative proof will be presented in Exercise 2.10. ◀

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

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

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

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** $\mathbf{P}(X = k) = \binom{n}{k} \cdot p^k (1-p)^{n-k}$, $n = 0, 1, \cdots, n$. Its expectation is np.

A **geometric** random variable X with parameter p is defined by probability distribution $\mathbf{P}(X = n) = (1-p)^{n-1}p$, $n = 1, 2, \cdots$. 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 $n > 0$,

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

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

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

Exercise 2.7 (a) By the memoryless property, we can ignore the case of $X > 1$ and $Y > 1$, thus $\mathbf{P}[X = Y] = \mathbf{P}[(X = 1) \cap (Y = 1)] / (1 - \mathbf{P}[(X > 1) \cap (Y > 1)])$. (b) Consider the first **trial**, and we can get an equation of $\mathbf{E}[\max(X, Y)]$. (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. $\mathbf{P}(X = n) = \binom{n-1}{k-1} p^k (1-p)^{n-k}$, $n \geq k$.

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

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

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 ($\mathbf{E}[\sum_{i=0}^{\infty} |X_n|] = \sum_{i=0}^{\infty} \mathbf{E}[|X_n|] < \infty$ is a consequence of monotone convergence). By dominated convergence, it follows that

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

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

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

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

3 Moments and Deviations

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

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

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 $\mathbf{P}(X \ge a) = \mathbf{E}[I] = \le \mathbf{E}[X/a] = \mathbf{E}[X]/a$. ◀

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

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

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

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

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

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

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 $a > 0$,

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

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

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

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 $\mathbf{P}(X \le m) \ge 1/2$ and $\mathbf{P}(X \ge m) \ge 1/2$.

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

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

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