[AI2613 Lecture 1] Review of Probability Theory

February 24, 2023

1 *Probability Space*

We start with the notion of probability space. A standard reference for the probability theory is [1].

Definition 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* $A^c \in \mathcal{F}$;
  - *for any finite or countable sequence of sets* $A_1, \dots, A_n, \dots \in \mathcal{F}$, *it holds that* $\bigcup_{i=1}^\infty A_i \in \mathcal{F}$.

$A^c := \Omega \setminus A$.

- *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* $A_1, \dots, A_n, \dots \in \mathcal{F}$, *it holds that* $\mathsf{P}\left(\bigcup_{i=1}^\infty A_i\right) = \sum_{i=1}^\infty \mathsf{P}(A_i)$.

Let $\mathcal{S} \subseteq 2^\Omega$. We use $\sigma(\mathcal{S})$ to denote the minimal σ-algebra containing sets in $\mathcal{S}$. 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 $\mathcal{F}' \subsetneq \mathcal{F}$ such that $\mathcal{S} \subseteq \mathcal{F}'$, $\mathcal{F}'$ is not a σ-algebra.

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

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

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

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*).*

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 $\mathscr{R}$ to denote the collection of Borel sets on $\mathbb{R}$. For any $A \subseteq \mathbb{R}$, we use $\mathscr{R}(A)$ to denote $\mathscr{R} \cap 2^A$.

2 *Random Variables*

Definition 2 (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 3 (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*

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

$X^{-1}(B') \triangleq \{\omega \in \Omega | X(\omega) \in B'\}$ 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 4 (Random Variable)**.** *. Let Ω' and $\mathcal{F}'$ in Definition 3 be $\mathbb{R}$ and the Borel algebra $\mathscr{B}$, then X in Definition 3 is a (real-valued) random variable.*

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

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.*

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.

3 *Distribution*

Let $(\Omega, \mathcal{F}, \mathrm{P})$ 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

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

The function $F(x) := \mathbf{Pr}[X \le x] = \mathrm{P}(X^{-1}(-\infty, x))$ is called the *cumulative distribution function (cdf)* of X.

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

$$\int_a^b f(x)\, \mathrm{d}x = F(b) - F(a),$$

then we call $f(x)$ a *probability density function (pdf)* of X.

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

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

4 Expectation and Variance

Definition 5 (Expectation)**.** *. Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space and $X : \Omega \to \mathbb{R}$ be a random variable.*

- *For a discrete random variable X, its expectation is*

$$\mathbf{E}[X] := \sum_{a \in \mathrm{Ran}(X)} a \cdot \mathbf{Pr}[X = a].$$

  If Ω is at most countable, we can also write

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

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

$$\mathbf{E}[X] := \int_{-\infty}^{\infty} t \cdot f(t)\,\mathrm{d}t.$$

  Sometimes it is more convenient to equivalently write the expectation as

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

  using Lebesgue integration.

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

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

Definition 6 (Variance)**.** *The variance of a random variable X is*

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

Proposition 7. *Let $X_1, \dots, X_n$ be random variables where n is a finite constant. Then*

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

5 Conditional Probability

Definition 8 (Conditional Probability)**.** *Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space. Let $A, B \in \mathcal{F}$ be two events with $\mathrm{P}(B) > 0$. The conditional probability of A given B is*

$$\mathrm{P}(A \mid B) := \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}.$$

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

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

Definition 9 (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*

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

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

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

Proposition 10.

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

Proof. • Since the value of $\mathbf{E}[X \mid Y]$ is determined by $Y(\omega)$, it is clearly $\sigma(Y)$-measurable.

- We compute $\mathbf{E}[f_Y]$ by definition.

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

□

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 $\mathbf{E}[X \mid Y = 0]$ should be identical to $\mathbf{E}[X]$, which is zero. However, we cannot directly adopt the definition before since $\mathbf{Pr}[Y = 0] = 0$.

Definition 11. *Let $(\Omega, \mathcal{F}, \mathrm{P})$ 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*

$$\forall A \in \sigma(Y), \int_A f_Y \, \mathrm{d}P = \int_A X \, \mathrm{d}P.$$

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

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,

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

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

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

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 $\gamma = 0.577\ldots$.

γ is called the Euler constant.

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.

8.1 Markov's Inequality

Theorem 12 (Markov's Inequality)**.** . *For any non-negative random variable X and $a > 0$,*

$$\mathbf{Pr}[X \ge a] \le \frac{\mathbf{E}[X]}{a}.$$

Proof. Since X is non-negative, we have

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

This is equivalent to

$$\mathbf{Pr}[X \ge a] \le \frac{\mathbf{E}[X]}{a}.$$

□

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*

$$\mathbf{Pr}[X \ge c] \le \frac{\mathbf{E}[X]}{c} = \frac{nH_n}{c}.$$

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 $\mathbf{E}[f(X)]$ instead $\mathbf{E}[X]$ for some increasing function $f:\mathbb{R}\to\mathbb{R}$ since

$$\mathbf{Pr}[X\ge a]=\mathbf{Pr}[f(X)\ge f(a)].$$

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

Theorem 13 (Chebyshev's Inequality). *. For any random variable with bounded $\mathbf{E}[X]$ and $a\ge 0$, it holds that*

$$\mathbf{Pr}[|X-\mathbf{E}[X]|\ge a]\le\frac{\mathbf{Var}[X]}{a^2}$$

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

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

□

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

$$\mathbf{Pr}[X\ge nH_n+t]\le\mathbf{Pr}[|X-\mathbf{E}[X]|\ge t]\le\frac{\mathbf{Var}[X]}{t^2}.$$

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

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

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

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

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

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

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

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 $t=\Theta(n\log n)$.

References

[1] Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019. 1