Lecture-01: Random Variables and Entropy

1 Random Variables

Our main focus will be on the behavior of large sets of discrete random variables.

Definition 1.1. A **discrete random variable**, X, is defined by following information: (i) $\mathcal{X}$: the finite set of values that it may take, (ii) $p_X : \mathcal{X} \to [0,1]$: the probability it takes each value $x \in X$. Of course, the probability distribution p_X must satisfy the normalization condition $\sum_{x\in X} p_X(x) = 1$. If there is no ambiguity, we may use $p(x)$ to denote $p_X(x)$.

Example 1.2. Let the random variable X denote the sum of two fair 6-sided dice. Then, $\mathcal{X} = \{2,3,\ldots,12\}$ and

$$p_X(x) = \frac{6 - |7 - x|}{36}.$$

Definition 1.3. An **event** $A \subseteq \mathcal{X}$ is a subset of values. The probability of an event is denoted

$$\mathbb{P}(X \in A) = \mathbb{P}(A) = \sum_{x\in A} p_X(x) = \sum_{x\in A} \mathbb{P}(X = x).$$

Also, an event is sometimes defined in words, $A =$ "X is even".

Example 1.4. If X is the sum of two fair 6-sided dice and $A =$ "X is even". Then,

$$\mathbb{P}(X \text{ is even}) = \mathbb{P}(A) = \sum_{x\in A} p_X(x) = \frac{1+3+5+5+3+1}{36} = \frac{1}{2}.$$

Definition 1.5. For a discrete random variable, the expected value (or average) of $f : \mathcal{X} \to \mathbb{R}$ is denoted

$$\mathbb{E}[f] = \mathbb{E}[f(X)] = \sum_{x\in\mathcal{X}} p_X(x) f(x).$$

Mathematically, $\mathbb{E}[]$ can be seen as a linear operator from the space of real functions on $\mathcal{X}$ to the set of real numbers. Thus,

$$\mathbb{E}[af(X) + bg(X)] = a\mathbb{E}[f(X)] + b\mathbb{E}[g(X)].$$

Example 1.6. If X is the sum of two fair 6-sided dice and $f(x) = (x-7)^2$, then

$$\mathbb{E}\left[(X-7)^2\right] = \sum_{x\in A} p_X(x)(x-7)^2 = \frac{2(1\cdot 5^2 + 2\cdot 4^2 + 3\cdot 3^2 + 4\cdot 2^2 + 5\cdot 1^2)}{36} = \frac{105}{18}.$$

Since the mean is $\mathbb{E}[X] = 7$, this actually equals the variance of X.

Definition 1.7. A continuous random variable, X, taking values on the set $\mathcal{X} = \mathbb{R}^d$ or in some smooth finite-dimensional manifold is defined by its cumulative distribution function $\mathbb{P}(X \leqslant x)$, where $X \leqslant x$ is used to denote $X_i \leqslant x_i$ for $i = 1, \ldots, d$. For such a r.v., the probability measure with respect to the infinitesimal element dx is denoted by $dp_X(x)$. For a measurable event $\mathcal{A} \subseteq X$, this gives

$$\mathbb{P}(X \in \mathcal{A}) = \int_{\mathcal{A}} dp_X(x) = \int \mathbb{1}_{\{x \in \mathcal{A}\}} dp_X(x),$$

where the indicator function $\mathbb{1}_{\{s\}}$ is 1 if the logical statement s is true and 0 otherwise. If p_X admits a density, with respect to Lebesgue measure, then it will be denoted by $p_X(x)$. In this case, we can write

$$\mathbb{P}(X \in \mathcal{A}) = \int_{\mathcal{A}} p_X(x) dx = \int \mathbb{1}_{\{x \in \mathcal{A}\}} p_X(x) dx.$$

Example 1.8. If X is a continuous random variable defined, for $a, b \in \mathbb{R}$ with $a < b$, by

$$\mathbb{P}(X \leqslant x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \leqslant x \leqslant b \\ 1 & x > b, \end{cases}$$

then it is uniform on $[a, b]$ and its density is given by $p_X(x) = \frac{1}{b-a} \mathbb{1}_{\{x \in [a,b]\}}$.

Definition 1.9. The expected value and variance of a function $f : \mathbb{R}^d \to \mathbb{R}$ of a continuous random variable $X \in \mathbb{R}^d$ are given by

$$\begin{aligned} \mathbb{E}[f] &= \mathbb{E}[f(X)] = \int f(x) dp_X(x), \\ \mathrm{Var}[f] &= \mathrm{Var}[f(X)] = \mathbb{E}\left[(f(X) - \mathbb{E}[f(X)])^2\right] = \mathbb{E}\left[f(X)^2\right] - \mathbb{E}[f(X)]^2. \end{aligned}$$

Example 1.10. If X is a continuous random variable that is uniform on $[a, b]$, then its mean and variance are given by

$$\begin{aligned} \mathbb{E}[X] &= \int_a^b \frac{x}{b-a} dx = \frac{b^2 - a^2}{2(b-a)} = \frac{b+a}{2}, \\ \mathrm{Var}[X] &= \int_a^b \frac{x^2}{b-a} dx - \left(\frac{b+a}{2}\right)^2 = \frac{b^3 - a^3}{3(b-a)} - \left(\frac{b+a}{2}\right)^2 = \frac{(b-a)^2}{12}. \end{aligned}$$

2 Entropy

In statistical mechanics, the entropy is proportional to the logarithm of the number of resolvable microstates associated with a macrostate. In classical mechanics, this quantity contains an arbitrary additive constant associated with the size of a microstate that is considered resolvable. In quantum mechanics, there is a natural limit to resolvability and this constant is related to the Planck constant. For random variables, Shannon chose the following definition which is similar in spirit.

Definition 2.1. The **entropy** (in bits) of a discrete random variable X with probability distribution $p(x)$ is denoted

$$H(X) \triangleq -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x) = \mathbb{E}\left[\frac{1}{\log_2 p(X)}\right],$$

where $0 \log_2 0 = 0$ by continuity. The notation $H(p)$ is used to denote $H(X)$ when $X \sim p(x)$. When there is no ambiguity, H will be used instead of $H(X)$. The unit of entropy is determined by the base of the logarithm with base-2 resulting in "bits" and the natural log (i.e., base-e) resulting in "nats".

Remark 2.2. Roughly speaking, the entropy $H(X)$ measures the uncertainty in the random variable X.

Example 2.3. If X is uniform, then $p(x) = \frac{1}{|\mathcal{X}|}$ and

$$H(X) = \mathbb{E}\left[\log_2 \frac{1}{|\mathcal{X}|}\right] = \log_2 |\mathcal{X}|.$$

Choosing $|\mathcal{X}| = 2$, we see that a uniform random bit has exactly $\log_2 2 = 1$ bit of entropy.

Example 2.4. Let X be a binary r.v. defined by $p(0) = 1 - q$ and $p(1) = q$. In this case, we have

$$H(X) = \mathcal{H}(q) = q \log_2 \frac{1}{q} + (1-q) \log_2 \frac{1}{1-q},$$

where $\mathcal{H}(q)$ is called the binary entropy function. This function is concave and symmetric about $q = \frac{1}{2}$. It also satisfies $\mathcal{H}(0) = \mathcal{H}(1) = 0$ and $\mathcal{H}(1/2) = 1$.

Example 2.5. The number of length-n binary sequences with exactly qn ones is given by $\binom{n}{qn}$. Using Stirling's formula, $n! = \sqrt{2\pi n}(\frac{n}{e})^n(1 + O(\frac{1}{n}))$, we see that

$$\begin{aligned}
\binom{n}{qn} &= \frac{n!}{(n-qn)!(qn)!} \\
&= \frac{\sqrt{2\pi n}(\frac{n}{e})^n(1+O(\frac{1}{n})}{\sqrt{2\pi n(1-q)}(\frac{n(1-q)}{e})^n(1+O(\frac{1}{n(1-q)})\sqrt{2\pi q n}(\frac{qn}{e})^n(1+O(\frac{1}{qn})} \\
&= \frac{1}{\sqrt{2\pi n q(1-q)}} 2^{n\mathcal{H}(q)} \left(1 + O\left(\frac{1}{nq(1-q)}\right)\right).
\end{aligned}$$

Remark 2.6. This shows that the binary entropy determines the exponential growth rate of the number of binary sequences with a fixed fraction of ones. In fact, this is a fundamental property of the entropy. More generally, we will see that the entropy $H(X)$ is the exponential growth rate of the number of length-n sequences (i.e., there are roughly $2^{nH(X)}$ such sequences) where the fraction of x's converges to $np(x)$. This also implies that $nH(X)$ is essentially equal to the minimum number of binary digits required to index all length-n sequences of this type.

Lemma 2.7. *Basic properties of entropy:*

1. *(non-negativity)* $H(X) \geqslant 0$ *with equality iff* X *is constant.*

   Proof. If X is not constant, there is an $x_0 \in \mathcal{X}$ with $p(x_0) \in (0,1)$. Thus,

   $$H(X) \geqslant p(x_0) \log_2(1/p(x_0)) \geqslant 0.$$

   □

2. *(decomposition rule) For any partition* $A = (A_1, A_2, \ldots, A_m)$ *of* $\mathcal{X}$*, we have*

$$H(p) = H(p_A) + \sum_{i=1}^{m} p(A_i) H(p_i),$$

where we define $p_A(i) = p(A_i) = \sum_{x \in A_i} p(x)$ *for* $i \in [m]$ *and* $p_i(x) = \frac{p(x)}{p(A_i)}$ *for* $x \in A_i$*.*

Proof. Observe that

$$H(X) = \sum_{i=1}^{m} \sum_{x \in A_i} p(x) \log_2 \frac{1}{p(x)} = H(p_A) + \sum_{i=1}^{m} p(A_i) \sum_{x \in A_i} \frac{p(x)}{p(A_i)} \log_2 \frac{p(A_i)}{p(x)}$$

□

Example 2.8. Compute the entropy of the distribution $p(x) = \begin{bmatrix} 0.125 & 0.375 & 0.25 & 0.25 \end{bmatrix}$. Using decomposition with $A_1 = \{1,2\}$ and $A_2 = \{3,4\}$, we get

$$H(p) = 1 + 0.5\mathcal{H}(1/4) + 0.5 \approx 1.9056.$$