EECS 126

Albert Ye

October 9, 2023

1 Probability Space

1.1 Definition

Essentially from 70. Events happen with some probability in a larger probability space that contains all events that can happen.

1.2 Axioms of Probability

Proposition 1 (Axioms)

1. (Positivity) $P(\omega > 0)$ for any event ω in probability space Ω.
2. (Totality) In any sample space Ω, $P(\Omega) = 1$.
3. (Additivity) If $A_1, A_2, \dots, A_n$ are independent, then

$$\sum_{i=1}^{n} A_i = \bigcup_{i=1}^{n} A_i.$$

From just this, we can get some useful information, such as the union bound.

Theorem 2 (Union Bound)

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i).$$

The proof is left as an exercise to the student, probably in the homework.

1.3 σ-algebra

Definition 3 (σ-algebra)

Given a sample space Ω, a set $\mathcal{F} \subseteq 2^{\Omega}$ is a σ-algebra if:

1. $\Omega \in \mathcal{F}$
2. If any event A is in $\mathcal{F}$, then its complement $\Omega \setminus A$ is also in $\mathcal{F}$.
3. For countably many events $A_1, A_2, \dots, A_n, \dots \in \mathcal{F}$, their union $A = \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

The biggest note is that Ω must be in a σ-algebra in order for any of the axioms of probability to apply.

2 Conditional Probability

2.1 Definition

2.2 Total Probability

2.3 Bayes' Rule

2.4 Continuous Bayes

3 It Depends

3.1 Independence / (Un)correlation

3.2 Conditional Expectation

Notice that $E[X|Y]$ is a random variable, but $E[X|Y = y]$ is a number. We can call $E[X|Y]$ a function $g(Y)$, where then $E[X|Y = y] = g(y)$ is just a value in the function.

3.3 Iterated Expectation

4 Distributions

4.1 Joint Distribution

Definition 4 (Joint Distribution)

A joint distribution $f_{X,Y}(x, y)$

4.2 Marginal Distribution

4.3 Derived Distribution

5 Random Variables

5.1 Discrete

5.1.1 Bernoulli

- PMF: $p_X(k) = \begin{cases} p & k = 1 \\ 1 = p & k = 0 \end{cases}$
- Expected value: p
- Variance: $p(1-p)$.

5.1.2 Binomial

- PMF: $p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}$ over all $k \in 0, 1, \ldots, n$.
- Expected value: np
- Variance: $np(1-p)$.

Run a Bernoulli test n times, find how many are positive.

5.1.3 Geometric

- PMF: $p_X(k) = (1-p)^{k-1} p$, for $k = 1, 2, \ldots$.
- Expected value: $\frac{1}{p}$
- Variance: $\frac{1-p}{p^2}$.

Here, each trial has a p probability of success, and we want to find the # of trials until one success.

5.1.4 Poisson

- PMF: $p_X(k) = \frac{\lambda^k (e^{-\lambda})}{k!}$.
- Expected value: λ
- Variance: λ

Used to simulate arrivals, I guess. More useful later, with Poisson processes.

5.2 Continuous

5.2.1 Uniform

5.2.2 Exponential

5.2.3 Gaussian

5.2.4 Joint Gaussian

The main tips for Joint Gaussian are to approach it as a sort of vectorized Gaussians over a certain number N of dimensions. Most of the addition / whatever operations in a Gaussian can be remodeled as a Joint Gaussian.

6 Moment Generating Functions

Definition 5

The **moment generating function** (also known as a transform) associated with a RV X, is a function $M_X(s)$ of a scalar parameter s defined by $M_X(s) = E(e^{sX})$.

the simpler notation $M(S)$ can be used whenever the underlying random variable X is clear from context. In more detail, when X is a discrete random variable, the corresponding MGF is given by

$$M(s) = \sum_x e^{sx} p_X(x).$$

Analogously, when continuous, we just replace the summation with an integral to get

$$M(s) = \int_{-\infty}^{\infty} e^{sx} f_X(x) dx.$$

Just an example so that I know what the reference is here:

Example 6 (Discrete Example)

Let

$$p_X(x) = \begin{cases} \frac{1}{2} & x = 2 \\ \frac{1}{6} & x = 3 \\ \frac{1}{3} & x = 5. \end{cases}$$

Then the corresponding transform is

$$M(s) = E(e^{sx}) = \frac{1}{2} + \frac{1}{6}e^{3s} + \frac{1}{3}e^{5s}.$$

Example 7 (Continuous Example)

Let X be an exponential RV with parameter λ:

$$f_X(x) = \lambda e^{-\lambda x} \qquad x \geq 0.$$

Then,

$$\begin{aligned} M(s) &= \lambda \int_0^{\infty} e^{sx} e^{-\lambda x} dx \\ &= \lambda \int_0^{\infty} e^{(s-\lambda)x} dx \\ &= \lambda \left(\frac{e^{(s-\lambda)x}}{s-\lambda} \right|_0^{\infty} \\ &= \frac{\lambda}{\lambda - s}. \end{aligned}$$

Notice, in above examples, that MGF is a **function** of parameter s, and not a number. We can also find MGF's for functions of X:

Proposition 8 (MGF of Linear Function of RV)

Let $Y = aX + b$. Then,

$$M_Y(s) = E(e^{s(aX+b)}) = e^{sb} E(e^{saX}) = e^{sb} M_X(sa).$$

From our previous example, we see that $M_X(s) = \frac{1}{1-s}$ where X is the exponential distribution

6.1 Moments

Now that we've established what a moment generating function is, now it's time to understand what is being generated.

Let's do a generic MGF
$$M(s) = \int_{-\infty}^{\infty} e^{sx} f_X(x)dx.$$
Now, we take the derivative of this.
$$\begin{aligned}\frac{d}{ds}M(s) &= \frac{d}{ds}\int_{-\infty}^{\infty} e^{sx} f_X(x)dx \\ &= \int_{-\infty}^{\infty} \frac{d}{ds}e^{sx} f_X(x)dx \\ &= \int_{-\infty}^{\infty} xe^{sx} f_X(x)dx.\end{aligned}$$

When $s = 0$, we have that this evaluates to $\int_{-\infty}^{\infty} xf_X(x)dx = E(X)$. If we differentiate n times, then we will get
$$\left(\frac{d^n}{ds^n}M(s)\right|_{s=0} = \int_{-\infty}^{\infty} x^n f_X(x)dx = E(X^n).$$

6.2 Inversion

Proposition 9 (Inversion Property)

The MGF $M_X(s)$ associated with an RV X uniquely determines the CDF of X, assuming that $M_X(s)$ is finite for all s in some interval $[-a, a]$ for positive a.

6.3 Sum of Independent Random Variables

Proposition 10

Addition of independent random variables corresponds to multiplication of transforms.

Proof. Let $Z = X + Y$. $M_Z(s) = E(e^{sZ}) = E(e^{s(X+Y)}) = E(e^{sX}e^{sY})$. Since X, Y are independent, e^{sX} and e^{sY} are independent random variables for any fixed s. Thus, $E(e^{sX}e^{sY}) = E(e^{sX})E(e^{sY}) = M_X(s)M_Y(s)$. □

We can further extend this; if $X_1, \dots, X_n$ is a collection of independent random variables and $Z = X_1 + \cdots + x_n$, then $M_Z(s) = M_{X_1}(s) \cdots M_{X_n}(s)$.

7 Concentration Inequalities

Theorem 11 (Markov's Inequality)

$P(X > a) = \dfrac{E(X)}{a}$.

Theorem 12 (Chebyshev's Inequality)

$P(|X - E(X)| > a) = \dfrac{\mathrm{Var}(X)}{a^2}$.

Used in lieu of confidence interval tests.

8 Modes of Convergence

8.1 Pointwise

Definition 13 (Pointwise Convergence)

Fix $\omega \in \Omega$, $\{X_n(\omega)\}_{n=1}^{\infty}$ converges **pointwise** if it becomes a real-valued sequence.

Usually, people don't use this because of reasons highlighted in 104.

8.2 Almost Sure

Definition 14 (Almost Sure Convergence)

$\{x_n\}_{n=1}^{\infty}$ converges **almost surely** to X if $P(\{\omega : \omega \in \Omega, \lim_{n\to\infty} X_n(\omega) = X(\omega)\}) = 1$.

This gets rid of ω with probability 0. If you find an ω such that convergence doesn't hold, it's fine as long sa $P(\omega) = 0$.

8.2.1 Checking for Almost Sure Convergence

There are a couple ways to check if some sequence converges almost surely.

8.3 In Probability

This is a weaker bound for convergence than almost sure convergence.

8.4 In distribution

Definition 15 (In Distribution Convergence)

$\{X_n\}_{n=1}^{\infty}$ converges in distribution (i.d.) to X if for every $x \in \mathbb{R}$, $P(X = x) = 0$.

In other words,

$$\lim_{n\to\infty} P(X_n \leq x) = 0.$$

Denote this as $X_n \to^d x$.

There are a couple of notable properties of in distribution convergence:

Theorem 16

In probability convergence implies in distribution convergence.

Proof. Suppose $X_n \to^P x$. □

8.5 Applications

8.5.1 Law of Large Numbers

Theorem 17 (Weak Law of Large Numbers)

Let $\{X_n\}_{n-1}^{\infty}$ be independent and identically distributed (i.i.d) with finite mean $|E[X_1]| < \infty$. Then,

$$\overline{X_n} = \frac{X_1 + X_2 + \cdots + X_n}{n} \to^P E[X_1].$$

Proof. Recall Chebyshev's Inequality, which gives us

$$P(|\bar{X_n} - E[\bar{X_n}]| \geq \epsilon) \leq \frac{E[(\bar{X_n} - E[\bar{X_n}]^2)]}{\epsilon^2}.$$

Now, we calculate the variance:

$$\begin{aligned}
\text{Var}(\bar{X_n}) &= \text{Var}\left(\frac{1}{n}(X_1 + X_2 + \cdots + X_n)\right) \\
&= \frac{1}{n^2}\text{Var}(X_1 + X_2 + \cdots + X_n) \\
&= \frac{1}{n^2}(\text{Var}(X_1) + \text{Var}(X_2) + \text{Var}(X_3) + \cdots + \text{Var}(X_n)) \\
&= \frac{\text{Var}(X_1)}{n},
\end{aligned}$$

because X_i are i.i.d.

Applying Chebyshev gives us

$$\lim_{n\to\infty} P(|\bar{X_n} - E[X_1]| \geq \epsilon) \leq \lim_{n\to\infty} \frac{\text{Var}(X_1)}{n\epsilon^2} = 0.$$

Thus, $\bar{X_n}$ converges in probability to $E[X_1]$. □

The strong law of large numbers has the same claim, except instead of in probability convergence it's almost sure convergence.

8.5.2 Central Limit Theorem

Once again let $\bar{X_n} = \frac{X_1+X_2+\cdots+X_n}{n}, S_n = X)1 + X_2 + \cdots + X_n$. Then, we know

$$\text{Var}(S_n) = n\text{Var}(X_1) \to \infty.$$

We let $Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$.

Theorem 18 (Central Limit Theorem)

We have $\{X_n\}_{n=1}^{\infty}$ is i.i.d, with mean μ and variance σ^2.

Then, $Z_n \to^d \mathcal{N}(0,1)$.

Theorem 19 (Poisson Limit Theorem)

Let $X_n = B(n \cdot \phi_n)$. Assume $\lim_{n\to\infty} n \cdot \phi_n = \lambda > 0$. Then,

$$X_n \to^d \text{pois}(\lambda).$$

Now we see why normal and poisson distribs are so useful.

9 Information Theory

9.1 Entropy

First, we define $\mathcal{X}$ as the range of a random variable X over all events in a probability space.

Definition 20 (Entropy)
Given a discrete random variable X and PMF $P_X(x)$, we have **entropy**

$$H(X) = \sum_{x \in \mathcal{X}} P_X(x) \log \frac{1}{P_X(x)}.$$

Furthermore, the average amount of surprise is defined as $E\left[\log \frac{1}{P_X(x)}\right]$.

Moreover, some properties of entropy:

1. $H(X) \geq 0$
2. $H(X)$ is
3. $H(X) \leq \log |x|$, achieved when X is uniform on x.

Where x is the range of $X(\omega)$ for all $\omega \in \Omega$.

Definition 21 (Joint Entropy)
Joint entropy $(X, Y) \sim P_{X,Y}$:

$$H(X,Y) = \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} P_{X,Y}(x,y) \log \frac{1}{P_{X,Y}(x,y)}.$$

Definition 22 (Conditional Entropy)

$$H(Y|X) = \sum_{x \in \mathcal{X}} H(Y|X = x).$$

Next, we observe some properties of joint and conditional entropy.

Proposition 23 1. (Chain Rule)

$$H(X,Y)H(X) + H(Y|X) = H(Y) + H(X|Y).$$

2. (Conditioning Reduces Entropy)

$$H(Y|X) \leq H(Y).$$

3.

$$H(X,Y) \leq H(X) + H(Y).$$

9.2 Mutual Information

Created by a Bob Fano, who argued more important than entropy.

Definition 24 (Mutual Information)
We define $I(X, Y)$ as the **mutual information** between X and Y, such that

$$\begin{aligned} I(X:Y) &= H(X) - H(X|Y) \geq 0 \\ &= H(X) + H(Y) - H(X,Y) \\ &= H(Y) - H(Y|X). \end{aligned}$$

We can think of $I(X, X) = H(X)$ as well.

Definition 25 (Kullback-Leibler Divergence)
We can also call this **relative entropy**.

$$D(P \parallel Q) = \sum_{x \in \mathcal{X}} P(X) \log \frac{P(x)}{Q(x)} \geq 0.$$

We can see that the mutual information can further be reduced to

$$\begin{aligned} I(X:Y) &= \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} P_{X,Y}(x,y) \log \frac{P_{X,Y}(x,y)}{P_X(x) P_Y(y)} \\ &= D(P_{X,Y} \parallel P_X \otimes P_Y), \end{aligned}$$

where we define $P_X \otimes P_Y$ as the cross product.

9.3 Source Coding

Let $X_1, X_2, \ldots, X_n$ be a string of symbols or binary code or etc. in a file. We want to convert this into some compressed $b(X_1, X_2, \ldots X_n)$.

Theorem 26
We assume $X_1, X_2, \ldots, X_n$ are i.i.d as X.

1. There exists a source code such that

$$\lim_{n \to \infty} E\left[\frac{1}{n} |b(x_1, \cdots, x_n)|\right] \leq H(X) + \epsilon$$

for any $\epsilon > 0$.

2. Conversely, no source code can achieve an average length less than $H(X)$ bits per symbol.