EECS126 Course Notes

Anmol Parande

Spring 2021 - Professor Thomas Courtade

Disclaimer: These notes reflect 126 when I took the course (Spring 2021). They may not accurately reflect current course content, so use at your own risk. If you find any typos, errors, etc, please raise an issue on the GitHub repository.

Contents

1 Introduction to Probability

Definition 1 *A probability space is a triple $(\Omega, \mathcal{F}, P)$ where Ω is a set of objects called the sample space, $\mathcal{F}$ is a family of subsets of Ω called events, and the probability measure $P : \mathcal{F} \to [0, 1]$.*

One key assumption we make is that $\mathcal{F}$ is a σ-algebra containing Ω, meaning that countably many complements, unions, and intersections of events in $\mathcal{F}$ are also events in $\mathcal{F}$. The probability measure P must obey **Kolmogorov's Axioms**.

1. $\forall A \in \mathcal{F},\ P(A) \geq 0$
2. $P(\Omega) = 1$
3. If $A_1, A_2, \cdots \in \mathcal{F}$ and $\forall i \neq j,\ A_i \bigcap A_j = \emptyset$, then $P\left(\bigcup_{i\geq 1} A_i\right) = \sum_{i\geq 1} P(A_i)$

We choose Ω and $\mathcal{F}$ to model problems in a way that makes our calculations easy.

Theorem 1

$$P(A^c) = 1 - P(A)$$

Theorem 2 (Inclusion-Exclusion Principle)

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{k=1}^{n} (-1)^{k+1} \left(\sum_{1\leq i_1 < \cdots < i_k \leq n} P(A_{i_1} \cap \cdots \cap A_{i_k}) \right)$$

Theorem 3 (Law of Total Probability) *If $A_1, A_2, \cdots$ partition Ω (i.e A_i are disjoint and $\cup A_i = \Omega$), then for event B,*

$$P(B) = \sum_i P(B \cap A_i)$$

1.1 Conditional Probability

Definition 2 *If B is an event with $P(B) > 0$, then the conditional probability of A given B is*

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Intuitively, conditional probabilty is the probability of event A given that event B has occurred. In terms of probability spaces, it is as if we have taken $(\Omega, \mathcal{F}, P)$ and now have a probabilty measure $P(\cdot|C)$ belonging to the space $(\Omega, \mathcal{F}, P(\cdot|C))$.

Theorem 4 (Bayes Theorem)

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

1.2 Independence

Definition 3 *Events A and B are independent if $P(A \cap B) = P(A)P(B)$*

If $P(B) > 0$, then A, B are independent if and only if $P(A|B) = P(A)$. In other words, knowing B occurred gave no extra information about A.

Definition 4 *If A, B, C with $P(C) > 0$ satisfy $P(A \cap B|C) = P(A|C)P(B|C)$, then A and B are conditionally independent given C.*

Conditional independence is a special case of independence where A and B are not necessarily independent in the original probability space which has the measure P, but are independent in the new probability space conditioned on C with the measure $P(\cdot|C)$.

2 Random Variables and their Distributions

Definition 5 *A random variable is a function $X : \Omega \to \mathbb{R}$ with the property $\forall \alpha \in \mathbb{R}$, $\{\omega \in \Omega : \ X(\omega) \leq \alpha\} \in \mathcal{F}$.*

The condition in definition 5 is necessary to compute $P(X \leq \alpha)$, $\forall \alpha \in \mathbb{R}$. This requirement also let us compute $P(X \in B)$ for most sets by leveraging the fact that $\mathcal{F}$ is closed under complements, unions, and intersections. For example, we can also compute $P(X > \alpha)$ and $P(\alpha < X \leq \beta)$. In this sense, the property binds the probability space to the random variable.

definition 5 also implies that random variables satisfy particular algebraic properties. For example, if X, Y are random variables, then so are $X+Y$, XY, X^p, $\lim_{n\to\infty} X_n$, etc.

Definition 6 *A discrete random variable is a random variable whose codomain is countable.*

Definition 7 *A continuous random variable is a random variable whose codomain is the real numbers.*

Although random variables are defined based on a probability space, it is often most natural to model problems without explicitly specifying the probability space. This works so long as we specify the random variables and their distribution in a "consistent" way. This is formalized by the so-called Kolmogorov Extension Theorem but can largely be ignored.

2.1 Distributions

Roughly speaking, the distribution of a random variable gives an idea of the likelihood that a random variable takes a particular value or set of values.

Definition 8 *The probability mass function (or distribution) of a discrete random variable X is the frequency with which X takes on different values.*

$$p_X : \mathcal{X} \to [0,1] \text{ where } \mathcal{X} = range(X), \qquad p_X(x) = Pr\left\{X = x\right\}.$$

Note that $\sum_{x\in\mathcal{X}} p_X(x) = 1$ since $\bigcap_{x\in\mathcal{X}}\{w : X(w) = x\} = \Omega$.

Continuous random variables are largely similar to discrete random variables. One key difference is that instead of being described by a probability "mass", they are instead described by a probability "density".

Definition 9 *The probability density function (distribution) of a continuous random variable describes the density by which a random variable takes a particular value.*

$$f_X : \mathbb{R} \to [0,\infty) \text{ where } \int_{-\infty}^{\infty} f_X(x)dx = 1 \text{ and } Pr\left\{X \in B\right\} = \int_B f_X(x)dx$$

Observe that if a random variable X is continuous, then the probability that it takes on a particular value is zero.

$$\Pr\left\{X = x\right\} = \lim_{\delta\to 0} \Pr\left\{x \leq X \leq x + \delta\right\} = \lim_{\delta\to 0} \int_x^{x+\delta} f_X(u)du = \int_x^x f_X(u)du = 0$$

Definition 10 *The cumulative distribution function (CDF) gives us the probability of a random variable X being less than or equal to a particular value.*

$$F_X : \mathbb{R} \to [0,1], \quad F_X(x) = Pr\left\{X \leq x\right\}$$

Note that by the Kolomogorov axioms, F_X must satisfy three properties:

1. F_X is non-decreasing.
2. $\lim_{x \to 0} F_X(x) = 0$ and $\lim_{x \to \infty} F_X(x) = 1$.
3. F_X is right continuous.

It turns out that if we have any function F_X that satisfies these three properties, then it is the CDF of some random variable on some probability space. Note that $F_X(x)$ gives us an alternative way to define continuous random variables. If $F_X(x)$ is absolutely continuous, then it can be expressed as

$$F_X(x) = \int_{-\infty}^{x} f_X(x) dx$$

for some non-negative function $f_X(x)$, and this is the PDF of a continuous random variable.

Often, when modeling problems, there are multiple random variables that we want to keep track of.

Definition 11 *If X and Y are random variables on a common probability space $(\Omega, \mathcal{F}, P)$, then the joint distribution (denoted $p_{XY}(x, y)$ or $f_{XY}(x, y)$ describes the frequencies of joint outcomes.*

Note that it is possible for X to be continuous and Y to be discrete (or vice versa).

Definition 12 *The marginal distribution of a joint distribution is the distribution of a single random variable.*

$$p_X(x) = \sum_y p_{XY}(x, Y = y), \qquad f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y) dy$$

Definition 13 *Two random variables X and Y are independent if their joint distribution is the product of the marginal distributions.*

Just like independence, we can extend the notion of conditional probability to random variables.

Definition 14 *The conditional distribution of X given Y captures the frequencies of X given we know the value of Y.*

$$p_{X|Y}(x|y) = \frac{P_{XY}(x, y)}{p_Y(y)}, \qquad f_{X|Y}(x|y) = \frac{f_{XY}(x, y)}{f_Y(y)}$$

Often, we need to combine or transform several random variables. A derived distribution is the obtained by arithmetic of several random variables or applying a function to several (or many) random variables. Since the CDF of a distribution essentially defines that random variable, it can often be easiest to work backwards from the CDF to the PDF or PMF. In the special case where we want to find $Y = g(X)$ for a function g.

$$F_y(y) = \Pr\left\{Y \le y\right\} = \Pr\left\{g(x) \le y\right\} = \Pr\left\{X \in g^{-1}([-\infty, y])\right\}, \quad g^{-1}(y) = \{x : g(x) = y\}.$$

Another special case of a derived distribution is when adding random variables together.

Theorem 5 *The resulting distribution of a sum of two independent random variables is the convolution of the distributions of the two random variables.*

$$p_{X+Y}(z) = \sum_{k=-\infty}^{\infty} p_X(k)p_Y(z-k), \quad f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(x)f_Y(z-x)dx$$

2.2 Properties of Distributions

2.2.1 Expectation

Definition 15 *The expectation of a random variable describes the center of a distribution,*

$$\mathbb{E}\left[X\right] = \sum_{x\in\mathcal{X}} xp_X(x), \quad \mathbb{E}\left[X\right] = \int_{-\infty}^{\infty} xf_X(x)dx$$

provided the sum or integral converges.

Expectation has several useful properties. If we want to compute the expectation of a function of a random variable, then we can use the law of the unconscious statisitician.

Theorem 6 (Law of the Unconscious Statistician)

$$\mathbb{E}\left[g(X)\right] = \sum_{x\in\mathcal{X}} g(x)p_X(x), \quad \mathbb{E}\left[g(X)\right] = \int_{-\infty}^{\infty} g(x)f_X(x)dx$$

Another useful property is its linearity.

$$\mathbb{E}\left[aX + bY\right] = a\mathbb{E}\left[X\right] + b\mathbb{E}\left[Y\right], \ \forall a, b \in \mathbb{R}.$$

Sometimes it can be difficult to compute expectations directly. For disrete distributions, we can use the tail-sum formula.

Theorem 7 (Tail Sum) *For a non-negative integer random variable,*

$$\mathbb{E}\left[X\right] = \sum_{k=1}^{\infty} Pr\left\{X \geq k\right\}.$$

When two random variables are independent, expectation has some additional properties.

Theorem 8 *If X and Y are independent, then*

$$\mathbb{E}\left[XY\right] = \mathbb{E}\left[X\right]\mathbb{E}\left[Y\right].$$

Earlier, we saw that we find a derived distribution by transforming and combining random variables. Sometimes, we don't need to actually compute the distribution, but only some of its properties.

Definition 16 *The nth moment of a random variable is $\mathbb{E}\left[X^n\right]$.*

It turns out that we can encode the moments of a distribution into the coefficients of a special power series.

Definition 17 *The moment generating function of a random variable X is given by $M_X(t) = \mathbb{E}\left[e^{tX}\right]$.*

Notice that if we apply the power series expansion of e^{tX}, we see that

$$M_X(t) = \sum_{n=0}^{\infty} \frac{t!}{n!}\mathbb{E}\left[X^n\right].$$

Thus the nth moment is encoded in the coefficients of the power series and we can retrieve them by taking a derivative:

$$\mathbb{E}\left[X^n\right] = \frac{\mathrm{d}^n}{\mathrm{d}t^n} M_X(t).$$

Another interesting point to notice is that for a continuous random variable

$$M_X(t) = \int_{-\infty}^{\infty} f_X(x) e^{tx} dx$$

is the Laplace transform of the distribution over the real line, and for a discrete random variable,

$$M_X(t) = \sum_{x=-\infty}^{\infty} p_X(x) e^{tx}$$

is the Z-transform of the distribution evaluated along the curve at e^{-t}.

Theorem 9 *If the MGF of a function exists, then it uniquely determines the distribution.*

This provides another way to compute the distribution for a sum of random variables because we can just multiply their MGF.

2.2.2 Variance

Definition 18 *The variance of a discrete random variable X describes its spread around the expectation and is given by*

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

Theorem 10 *When two random variables X and Y are independent, then*

$$Var\left(X + Y\right) = Var\left(X\right) + Var\left(Y\right).$$

Definition 19 *The covariance of two random variables describes how much they depend on each other and is given by*

$$Cov\left(X, Y\right) = \mathbb{E}\left[(X - \mathbb{E}\left[X\right])(Y - \mathbb{E}\left[Y\right])\right] = \mathbb{E}\left[XY\right] - \mathbb{E}\left[X\right]\mathbb{E}\left[Y\right].$$

If $Cov\left(X, Y\right) = 0$ then X and Y are uncorrelated.

Definition 20 *The correlation coefficient gives a single number which describes how random variables are correlated.*

$$\rho(X, Y) = \frac{Cov\left(X, Y\right)}{\sqrt{Var\left(X\right)}\sqrt{Var\left(Y\right)}}.$$

Note that $-1 \leq \rho \leq 1$.

2.3 Common Discrete Distributions

Definition 21 *X is uniformly distributed when each value of X has equal probability.*

$$X \sim \textit{Uniform}(\{1,2,\cdots,n\}) \implies p_X(x) = \begin{cases} \frac{1}{n} & x = 1,2,\cdots,n, \\ 0 & \textit{else.} \end{cases}$$

Definition 22 *X is a Bernoulli random variable if it is either 0 or 1 with $p_X(1) = p$.*

$$X \sim \textit{Bernoulli}(p) \implies p_X(x) = \begin{cases} 1-p & x = 0, \\ p & x = 1, \\ 0 & \textit{else.} \end{cases}$$

$$\mathbb{E}\left[X\right] = p \qquad Var\left(X\right) = (1-p)p$$

Bernoulli random variables are good for modeling things like a coin flip where there is a probability of success. Bernoulli random variables are frequently used as indicator random variables $\mathbb{1}_A$ where

$$\mathbb{1}_A = \begin{cases} 1 & \text{if A occurs,} \\ 0 & \text{else.} \end{cases}$$

When paired with the linearity of expectation, this can be a powerful method of computing the expectation of something.

Definition 23 *X is a Binomial random variable when*

$$X \sim \textit{Binomial}(n,p) \implies p_X(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & x = 0,1,\cdots,n \\ 0 & \textit{else.} \end{cases}$$

$$\mathbb{E}\left[X\right] = np \qquad Var\left(X\right) = np(1-p)$$

A binomial random variable can be thought of as the number of successes in n trials. In other words,

$$X \sim \text{Binomial}(n,p) \implies X = \sum_{i=1}^{n} X_i, \quad X_i \sim \text{Bernoulli}(p).$$

By construction, if $X \sim \text{Binomial}(n,p)$ and $Y \sim \text{Binomial}(m,p)$ are independent, then $X + Y \sim \text{Binomial}(m+n,p)$.

Definition 24 *A Geometric random variable is distributed as*

$$X \sim \mathit{Geom}(p) \implies p_X(x) = \begin{cases} p(1-p)^{x-1} & x = 1, 2, \cdots \\ 0 & \mathit{else}. \end{cases}$$

$$\mathbb{E}\left[X\right] = \frac{1}{p} \qquad Var\left(X\right) = \frac{1-p}{p^2}$$

Geometric random variables are useful for modeling the number of trials required before the first success. In other words,

$$X \sim \text{Geom}(p) \implies X = \min\{k \geq 1 : X_k = 1\} \text{ where } X_i \sim \text{Bernoulli}(p).$$

A useful property of geometric random variables is that they are memoryless:

$$\Pr\left\{X = K + M | X > k\right\} = \Pr\left\{X = M\right\}.$$

Definition 25 *A Poisson random variable is distributed as*

$$X \sim Poisson(\lambda) \implies p_X(x) = \begin{cases} \frac{\lambda^x e^{-\lambda}}{x!} & x = 0, 1, \cdots \\ 0 & \textit{else.} \end{cases}$$

$$\mathbb{E}\left[X\right] = \lambda$$

Poisson random variables are good for modeling the number of arrivals in a given interval. Suppose you take a given time interval and divide it into n chunks where the probability of arrival in chunk i is $X_i \sim$ Bernoulli(p_n). Then the total number of arrivals $X_n = \sum_{i=1}^n X_i$ is distributed as a Binomial random variable with expectation $np_n = \lambda$. As we increase n to infinity but keep λ fixed, we arrive at the poisson distribution.

A useful fact about Poisson random variables is that if $X \sim$ Poisson(λ) and $Y \sim$ Poisson(μ) are independent, then $X + Y \sim$ Poisson$(\lambda + \mu)$.

2.4 Common Continuous Distributions

Definition 26 *A continuous random variable is uniformly distributed when the pdf of X is constant over a range.*

$$X \sim \mathit{Uniform}(a, b) \implies f_X(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b, \\ 0 & \textit{else.} \end{cases}$$

The CDF of a uniform distribution is given by

$$F_X(x) = \begin{cases} 0, & x < a, \\ \frac{x-a}{b-a}, & x \in [a, b) \\ 1, & x \geq b. \end{cases}$$

Definition 27 *A continuous random variable is exponentially distributed when its pdf is given by*

$$X \sim Exp(\lambda) \implies f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0 & \textit{else.} \end{cases}$$

Exponential random variables are the only continuous random variable to have the memoryless property:

$$\Pr\{X > t + s | X > s\} = \Pr\{X > t\}, \quad t \geq 0.$$

The CDF of the exponential distribution is given by

$$F_X(x) = \lambda \int_0^x e^{-\lambda u} du = 1 - e^{-\lambda x}$$

Definition 28 *X is a Gaussian Random Variable with mean μ and variance σ^2 (denoted $X \sim \mathcal{N}(\mu, \sigma^2)$) if it has the PDF*

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

The standard normal is $X \sim \mathcal{N}(0, 1)$, and it has the CDF

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-u^2}{2}} du$$

There is no closed from for $\Phi(x)$. It turns out that every normal random variable can be transformed into the standard normal (i.e $\frac{X-\mu}{\sigma} \sim \mathcal{N}(0, 1)$). Some facts about Gaussian random variables are

1. If $X \sim \mathcal{N}(\mu_x, \sigma_x^2),\ Y \sim \mathcal{N}(\mu_y, \sigma_y^2)$ are independent, then $X + Y \sim \mathcal{N}(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$.
2. If X, Y are independent and $(X + Y), (X - Y)$ are independent, then both X and Y are Gaussian with the same variance.

2.4.1 Jointly Gaussian Random Variables

Jointly Gaussian Random Varables, also known as Gaussian Vectors, can be defined in a variety of ways.

Definition 29 *A Gaussian Random Vector* $\boldsymbol{X} = \begin{bmatrix} X_1 & \cdots & X_n \end{bmatrix}^T$ *with density on* $\mathbb{R}^n$, $Cov\left(\boldsymbol{X}\right) = \Sigma, \mathbb{E}\left[X\right] = \boldsymbol{\mu}$ *is defined by the pdf*

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^n det(\Sigma)}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$$

Definition 30 *A joint gaussian random variable is an affine transformation of independent and identically distributed standard normals.*

$$\boldsymbol{X} = \boldsymbol{\mu} + A\boldsymbol{W}$$

where $A = \Sigma^{1/2}$ *is a full-rank matrix and* $\boldsymbol{W}$ *is a vector of i.i.d standard normals.*

Definition 31 *A random variable is jointly gaussian if all 1D projections are Gaussian*

$$\boldsymbol{a}^T\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{a}^T\boldsymbol{\mu}, \boldsymbol{a}^T\Sigma\boldsymbol{a})$$

In addition to their many definitions, jointly gaussian random variables also have interesting properties.

Theorem 11 *If* $\boldsymbol{X}$ *and* $\boldsymbol{Y}$ *are jointly gaussian random variables, then*

$$X = \mu_{\boldsymbol{X}} + \Sigma_{\boldsymbol{XY}}(\boldsymbol{Y} - \boldsymbol{\mu}_{\boldsymbol{Y}}) + \boldsymbol{V} \text{ where } V \sim \mathcal{N}(0, \Sigma_X - \Sigma_{\boldsymbol{XY}}\Sigma_Y^{-1}\Sigma\boldsymbol{Y}\boldsymbol{X}$$

theorem 11 tells us that each entry in Gaussian Vector can be thought of as a "noisy" version of the others.

2.5 Hilbert Spaces of Random Variables

One way to understand random variables is through linear algebra by thinking of them as vectors in a vector space.

Definition 32 *An real inner product space V is composed of a vector space V over a real scalar field equipped with an inner product $\langle \cdot, \cdot \rangle$ that satisfies $\forall u, v, w \in V$, $a, b \in \mathbb{R}$,*

1. $\langle u, v \rangle = \langle v, u \rangle$
2. $\langle au + bv, w \rangle = a\langle u, w \rangle + b\langle v, w \rangle$
3. $\langle u, u \rangle \geq 0$ *and* $< u, u >= 0 \Leftrightarrow u = 0$

Inner products spaces are equipped with the norm $\|v\| = \sqrt{\langle v, v \rangle}$.

Definition 33 *A Hilbert Space is a real inner product space that is complete with respect to its norm.*

Loosely, completeness means that we can take limits of without exiting the space. It turns out that random variables satisfy the definition of a Hilbert Space.

Theorem 12 *Let $(\Omega, \mathcal{F}, P)$ be a probability space. The collection of random variables X with $\mathbb{E}\left[X^2\right] < \infty$ on this probability space form a Hilbert Space with respect to the inner product $\langle X, Y \rangle = \mathbb{E}\left[XY\right]$.*

Hilbert spaces are important because they provide a notion of geometry that is compatible with our intuition as well as the geometry of $\mathbb{R}^n$ (which is a Hilbert Space). One geometric idea is that of orthogonality. Two vectors are orthogonal if $\langle X, Y \rangle = 0$. Two random variables will be orthogonal if they are zero-mean and uncorrelated. Using orthogonality, we can also define projections.

Theorem 13 (Hilbert Projection Theorem) *Let $\mathcal{H}$ be a Hilbert Space and $\mathcal{U} \subseteq \mathcal{H}$ be a closed subspace. For each vector $v \in \mathcal{H}$, $\operatorname{argmin} \|u - v\|$ has a unique solution (there is a unique closest point $u \in \mathcal{U}$ to v). If u is the closest point to v, then $\forall u \in \mathcal{U}$, $\langle u - v, u' \rangle$.*

theorem 13 is what gives rise to important properties like the Pythogorean Theorem for any Hilbert Space.

$$\|u\|^2 + \|u - v\|^2 = \|v\| \text{ where } u = \operatorname{argmin} \|u - v\|.$$

Suppose we had to random variables X and Y. What happens if we try and project one onto the other?

Definition 34 *The conditional expectation of X given Y is the bounded continuous function of Y such that $X - \mathbb{E}\left[X|Y\right]$ is orthogonal to all other bounded continuous functions $\phi(Y)$.*

$$\forall \phi,\ \mathbb{E}\left[(X - \mathbb{E}\left[X|Y\right])\phi(Y)\right] = 0.$$

Thus, the conditional expectation is the function of Y that is closest to X. It's interpretation is that the expectation of X can change after observing some other random variable Y. To find $\mathbb{E}\left[X|Y\right]$, we can use the conditional distribution of X and Y.

Theorem 14 *The conditional expectation of a conditional distribution is given by*

$$\mathbb{E}\left[X|Y=y\right] = \sum_{x\in\mathcal{X}} x p_{X|Y}(x|y), \quad \mathbb{E}\left[X|Y=y\right] = \int_{-\infty}^{\infty} x f_{X|Y}(x,y)dx$$

Notice that $\mathbb{E}\left[X|Y\right]$ is a function of the random variable Y, meaning we can apply theorem 6.

Theorem 15 (Tower Property) *For all functions f,*

$$\mathbb{E}\left[f(Y)X\right] = \mathbb{E}\left[f(Y)\mathbb{E}\left[X|Y\right]\right]$$

Alternatively, we could apply lineary of expectation to definition 34 to arrive at the same result. If we apply theorem 15 to the function $f(Y) = 1$, then we can see that $\mathbb{E}\left[\mathbb{E}\left[X|Y\right]\right] = \mathbb{E}\left[X\right]$.

Just as expectation can change when we know additional information, so can variance.

Definition 35 *Conditional Variance is the variance of X given the value of Y.*

$$Var\left(X|Y=y\right) = \mathbb{E}\left[(X-\mathbb{E}\left[X|Y=y\right])^2|Y=y\right] = \mathbb{E}\left[X^2|Y=y\right] - \mathbb{E}\left[X|Y=y\right]^2$$

Conditional variance is a random variable just as expectation is.

Theorem 16 (Law of Total Variance)

$$Var\left(X\right) = \mathbb{E}\left[Var\left(X|Y\right)\right] + Var\left(\mathbb{E}\left[X|Y\right]\right)$$

The second term in the law of total variance ($Var\left(\mathbb{E}\left[X|Y\right]\right)$) can be interpreted as on average, how much uncertainty there is in X given we know Y.

3 Concentration

In real life, for the most part, we can't compute probabilities in closed form. Instead, we either bound them, or we want to show that $P(A) \approx 0$ or $P(A) \approx 1$.

3.1 Concentration Inequalities

Theorem 17 (Markov's Inequality) *For a non-negative random variable X,*

$$Pr\left\{X \geq t\right\} \leq \frac{\mathbb{E}\left[X\right]}{t}, \quad t \geq 0.$$

Theorem 18 (Chebyshev's Inequality) *If X is a random variable, then*

$$Pr\left\{|X - \mathbb{E}\left[X\right]| \geq t\right\} \leq \frac{Var\left(X\right)}{t^2}.$$

Intuitively, theorem 18 gives gives a "better" bound than theorem 17 because it incorporates the variance of the random variable. Using this idea, we can define an even better bound that incorporates information from all moments of the random variable.

Definition 36 (Chernoff Bound) *For a random variable X and $a \in \mathbb{R}$,*

$$Pr\left\{X \geq a\right\} \leq \frac{\mathbb{E}\left[e^{tX}\right]}{e^{ta}} = e^{-ta}M_x(t).$$

After computing the Chernoff bound for a general t, we can then optimize over it to compute the best bound possible.

3.2 Convergence

The idea of convergence brings the mathematical language of limits into probability. The fundamental question we want to answer is given random variables $X_1, X_2, \cdots$, what does it mean to compute

$$\lim_{n \to \infty} X_n.$$

This question is not as straightforward as it seems because random variables are functions, and there are many ways to define the convergence of functions.

Definition 37 *A sequence of random variables converges almost surely to X if*

$$P\left(\lim_{n\to\infty} X_n = X\right) = 1$$

One result of almost sure convergence deals with deviations around the mean of many samples.

Theorem 19 (Strong Law of Large Numbers) *If $X_1, X_2, \cdots, X_n$ are independently and identically distributed to X where $\mathbb{E}\left[X\right] < \infty$, then $\frac{1}{n}\sum_i X_i$ converges almost surely to $\mathbb{E}\left[X\right]$.*

The strong law tells us that for any observed realization, there is a point after which there are no deviations from the mean.

Definition 38 *A sequence of random variables converges in probability if*

$$\forall \epsilon > 0, \quad \lim_{n\to\infty} P(|X_n - X| > \epsilon) = 0$$

Convergence in probability can help us formalize the intuition that we have which says probability is the frequency with which an even happens over many trials of an event.

Theorem 20 (Weak Law of Large Numbers) *Let $X_1, X_2, \cdots, X_n$ be independently and identically distributed according to X, and let $M_n = \frac{1}{n}\sum X_i$. Then for $\epsilon > 0$,*

$$\lim_{n\to\infty} Pr\left\{|M_n - \mathbb{E}\left[X\right]| > \epsilon\right\} = 0.$$

It tells us that the probability of a deviation of ϵ from the true mean will go to 0 in the limit, but we can still observe these deviations. Nevertheless, the weak law helps us formalize our intuition about probability. If $X_1, X_2, \cdots, X_n$ are independently and identically distributed according to X, then we can define the empirical frequency

$$F_n = \frac{\sum \mathbb{1}_{X_i \in B}}{n} \implies \mathbb{E}\left[F_n\right] = P(X \in B).$$

By theorem 20,

$$\lim_{n\to\infty} \Pr\left\{|F_n - P(X \in B)| > \epsilon\right\} = 0,$$

meaning over many trials, the empirical frequency is equal to the probility of the event, matching intuition.

Definition 39 *A sequence of random variables converges in distribution if*

$$\lim_{n\to\infty} F_{X_n}(x) = F_x(x).$$

An example of convergence in distribution is the central limit theorem.

Theorem 21 (Central Limit Theorem) *If $X_1, X_2, \cdots$ are independently and identically distributed according to X with $Var\left(X\right) = \sigma^2$ and $\mathbb{E}\left[X\right] = \mu$, then*

$$\lim_{n\to\infty} P\left(\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} \leq x\right) = \Phi(x)$$

In other words, a sequence of random variables converges in distribution to a normal distribution with variance σ^2 and mean μ.

These notions of convergence are not identical, and they do not necessarily imply each other. It is true that almost sure convergence implies convergence in probability, and convergence in probability implies convergence in distribution, but the implication is only one way.

Once we know how a random variable converges, we can then also find how functions of that random variable converge.

Theorem 22 (Continuous Mapping Theorem) *If f is a continuous function, then if X_n converges to X, then $f(X_n)$ converges to $f(X)$. The convergence can be almost surely, in probability, or in distribution.*

4 Information Theory

Information Theory is a field which addresses two questions

1. **Source Coding:** How many bits do I need to losslessly represent an observation.

2. **Channel Coding:** How reliably and quickly can I communicate a message over a noisy channel.

4.1 Quantifying Information

Intuitively, for a PMF of a disrete random variable, the surprise associated with a particular realization is $-\log p_X(x)$ since less probable realizations are more surprising. With this intuition, we can try and quantify the "expected surprise" of a distribution.

Definition 40 *For a Discrete Random Variable $X \sim p_X$, the Entropy of X is given by*

$$H(x) = \mathbb{E}\left[-\log_2 p_X(x)\right] = -\sum_{x \in \mathcal{X}} p_X(x) \log_2 p_X(x).$$

Alternative interpretations of entropy are the average uncertainty and how random X is. Just like probabilites, we can define both joint and conditional entropies.

Definition 41 *For Discrete Random Variables X and Y, the joint entropy is given by*

$$H(X,Y) = \mathbb{E}\left[-\log_2 p_{XY}(x,y)\right] = -\sum_{x,y \in \mathcal{X} \times \mathcal{Y}} p_{XY}(x,y) \log_2 p_{XY}(x,y).$$

Definition 42 *For Discrete Random Variable X and Y, the conditional entropy is given by*

$$H(Y|X) = \mathbb{E}\left[-\log_2 p_{Y|X}(y|x)\right] = \sum_{x \in \mathcal{X}} p_X(x) H(Y|X=x).$$

Conditional entropy has a natural interpretation which is that it tells us how surprised we are to see $Y = y$ given that we know $X = x$. If X and Y are independent, then $H(Y) = H(Y|X)$ because realizing X gives no additional information about Y.

Theorem 23 (Chain Rule of Entropy)

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

In addition to knowing how much our surprise changes for a random variable when we observe a different random variable, we can also quantify how much additional information observing a random variable gives us about another.

Definition 43 *For random variables X and Y, the mutual information is given by*

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X).$$

4.2 Source Coding

Source coding deals with finding the minimal number of bits required to represent data. This is essentially the idea of lossless compression. In this case, our message is the sequence of realizations of independently and identically distributed random variables $(X_i)_{i=1}^n \sim p_X$. The probability of observing a particular sequence is then

$$P(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} p_X(x_i).$$

Theorem 24 (Asymptotic Equipartition Property) *If we have a sequence of independently and identically distributed random variables $(X_i)_{i=1}^n \sim p_X$, then $-\frac{1}{n}\log P(x_1, x_2, \cdots, x_n)$ converges to $H(X)$ in probability.*

theorem 24 tells us that with overwhelming probability, we will observe a sequence that is assigned probability $2^{-nH(X)}$. Using this idea, we can define a subset of possible observed sequences that in the limit, our observed sequence must belong to with overwhelming probability.

Definition 44 *For a fixed $\epsilon > 0$, for each $n \geq 1$, the typical set is given by*

$$A_\epsilon^{(n)} = \left\{(x_1, x_2, \cdots, x_n) : 2^{-n(H(X)+\epsilon)} \leq P(x_1, x_2, \cdots, x_n) \leq 2^{-n(H(X)-\epsilon)}\right\}.$$

Two important properties of the typical set are that

1. $\lim_{n\to\infty} P\left((x_1, x_2, \ldots, x_n) \in A_\epsilon^{(n)}\right) = 1$
2. $|A_\epsilon^{(n)}| \leq 2^{n(H(X)+\epsilon)}$

The typical set gives us an easy way to do source coding. If I have N total objects, then I only need $\log N$ bits to represent each object, so I can define a simple protocol which is

1. If $(x_i)_{i=1}^n \in A_{\frac{\epsilon}{2}}^{(n)}$, then describe them using the $\log|A_{\frac{\epsilon}{2}}^{(n)}| \leq n\left(H(X) + \frac{\epsilon}{2}\right)$ bits
2. If $(x_i)_{i=1}^n \notin A_{\frac{\epsilon}{2}}^{(n)}$, then describe them naiively with $n\log|\mathcal{X}|$ bits.

This makes the average number of bits required to describe a message

$$\begin{aligned}\mathbb{E}\left[\text{\# of Bits}\right] &\leq n\left(H(X) + \frac{\epsilon}{2}\right) P\left((x_i)_{i=1}^n \in A_{\frac{\epsilon}{2}}^{(n)}\right) + n\log|\mathcal{X}| P\left((x_i)_{i=1}^n \in A_{\frac{\epsilon}{2}}^{(n)}\right) \\ &\leq n(H(X) + \frac{\epsilon}{2}) + n\frac{\epsilon}{2} \leq n(H(X) + \epsilon)\end{aligned}$$

This is the first half of a central result of source coding.

Theorem 25 (Source Coding Theorem) *If $(X_i)_{i=1}^n \sim p_X$ are a sequence of independently and identically distributed random varibles, then for any $\epsilon > 0$ and n sufficiently large, we can represent $(X_i)_{i=1}^n$ using fewer than $n(H(X) + \epsilon)$ bits. Conversely, we can not losslessly represent $(X_i)_{i=1}^n$ using fewer than $nH(X)$ bits.*

This lends a new interpretation of the entropy $H(X)$: it is the average number of bits required to represent X.

4.3 Channel Coding

Whereas source coding deals with encoding information, channel coding deals with transmitting it over a noisy channel. In general, we have a message M, and encoder, a channel, and a decoder as in fig. 1.

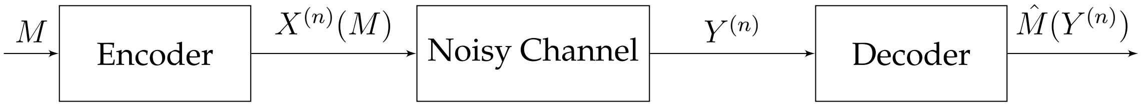

Figure 1: Channel Coding

Each channel can be described by a conditional probability distribution $p_{Y|X}(y|x)$ for each time the channel is used.

Definition 45 *For a channel described by $p_{Y|X}$, the capacity is given by*

$$C = \max_{p_X} I(X;Y).$$

In words, the capacity describes the maximum mutual information between the channel input and output.

Definition 46 *Suppose we use the channel n times to send a message that takes on average $H(m)$ bits to encode, then the rate of the channel is*

$$R = \frac{H(M)}{n}$$

Theorem 26 (Channel Coding Theorem) *For a channel decsribed by $p_{Y|X}$ and $\epsilon > 0$ and $R < C$, for all n sufficiently large, there exists a rate R communication scheme that achieves a probability of error less than ϵ. If $R > C$, then the probability of error converges to 1 for any communication scheme.*

5 Random Processes

Definition 47 *A random/stochastic process is a sequence of random variables $(X_n)_{n\geq 0}$.*

The random variables in a stochastic process do not have to be independently and identically distributed. In fact, if they are not, then we can get additional modeling power.

Definition 48 *A random process $(X_n)_{n\in\mathbb{N}}$ is stationary if for all $k, n > 0$ and all events $A_1, \cdots, A_n$, then*

$$Pr\left\{X_1 \in A_1, \cdots, X_n \in A_n\right\} = Pr\left\{X_{k+1} \in A_1, \cdots, A_{k+n} \in A_n\right\}$$

Stationarity is often a good assumption that can simplify systems which have been running for a long period of time.

5.1 Discrete Time Markov Chains

Definition 49 *$(X_n)_{n\geq 0}$ is a Markov Chain if each random variable X_i takes values in a discrete set S (the state space), and,*

$$\forall n \geq 0,\ i, j \in S,\ Pr\left\{X_{n+1} = j | X_n = i, \cdots, X_0 = x_0\right\} = Pr\left\{X_{n+1} = i | X_n = j\right\}$$

In words, a Markov Chain is a sequence of random variables satisfying the Markov Property where probability of being in a state during the next time step only depends on the current state.

Definition 50 *A temporally homogenous Markov Chain is one where the transition probabilities $Pr\left\{X_{n+1} = j | X_n = i\right\} = p_{ij}$ for all $i, j \in S$ and $n \geq 0$.*

Temporally Homogenous Markov Chains don't change their transition probabilities over time. Since the p_{ij} are conditional probabilities, they must satisfy

1. $\forall i, j \in S,\ p_{ij} \geq 0$
2. $\forall i \in S,\ \sum_{j\in S} p_{ij} = 1$

Definition 51 *The transition matrix of a Markov Chain is a matrix P where the ijth entry $P_{ij} = p_{ij}$ for all $i, j \in S$.*

The transition matrix encodes the one-step transition probabilities of the Markov Chain.

Theorem 27 (Chapman-Kolmogorov Equation) *The n-step transition probabilities (i.e starting in i and ending in j n steps later) of the Markov Chain are given by $p_{ij}^{(n)} = P_{ij}^{n}$.*

One useful thing we can comptue with Markov Chain is when the chain first enters a particular state.

Definition 52 *For a $A \subset S$, the hitting time of A is given by*

$$T_A = \min_n \{n \geq 0 : X_n \in A\}$$

Computing the expected hitting time is an example of a broader type of Markov Chain Analysis called **First Step Analysis**. In First Step Analysis, we set up a system of equations that relies on the Markov property to generate a system of equations that only look at the first transition in the chain. For expected hitting time, these look like

1. For $i \notin A$, $\mathbb{E}\left[T_A | X_0 = i\right] = 1 + \sum_j p_{ij} \mathbb{E}\left[T_A | X_0 = j\right]$
2. For $i \in A$, $\mathbb{E}\left[T_A | X_0 = i\right] = 0$

5.1.1 Properties of Markov Chains

Definition 53 *If $\exists n \geq 1$ such that $p_{ij}^{(n)} \neq 0$, then j is accessible from i, and we write $i \to j$.*

Definition 54 *States i and j communicate with each other when $i \to j$ and $j \to i$. We write this as $i \leftrightarrow j$.*

By convention, we say that $i \leftrightarrow i$. It turns out that $\leftrightarrow$ is an equivalence relation on the state space S. An equivalence relation means that

1. $\forall i \in S,\ i \leftrightarrow i$
2. $\forall i, j \in S,\ i \leftrightarrow j \Leftrightarrow j \leftrightarrow i$
3. $\forall i, j, k \in S, i \leftrightarrow k, k \leftrightarrow j \Rightarrow i \leftrightarrow j$

This means that $\leftrightarrow$ partitions the state-space S into equivalence classes (i.e classes of communicating states).

Definition 55 *A Markov Chain is irreducible if S is the only class.*

Definition 56 *An irreducible Markov Chain is reversible if and only if there exists a probability vector π that satisfies the* ***Detailed Balance Equations****:*

$$\forall i, j \in S,\ \pi_j p_{ij} = \pi_i p_{ji}$$

Markov Chains which satisfy the detailed balance equations are called reversible because if $X_0 \sim \pi$, then the random vectors $(X_0, X_1, \cdots, X_n)$ and $(X_n, X_{n-1}, \cdots, X_0)$ are equal in distribution.

Theorem 28 *If the graph of a Markov Chain (transform the state transition diagram by making edges undirected, removing self-loops, and removing multiple edges) is a tree, then the Markov Chain is reversible.*

5.1.2 Class Properties

A class property is a property where if one element of a class has the property, all elements of the class have the property. Markov Chains have several of these properties which allow us to classify states.

Definition 57 *A state $i \in S$ is recurrent if given that $X_0 = i$, the process revisits state i with probability 1.*

Definition 58 *A state is $i \in S$ is transient if it is not recurrent.*

Recurrence means that we will visit a state infinitely often in the future if we start in that state, while transience means we will only visit the state finitely many times. Recurrence and transience can be easily identified from the transition diagram.

1. Any finite communicating class which has no edges leaving the class is recurrent
2. If a state has an edge leading outside its communicating class, then it is transient
3. If a state is recurrent, then any state it can reach is recurrent

We can further break recurrence down if we modify the definition of hitting time to be $T_i = \min_n\{n \geq 1 : X_n = i\}$ (the first time the chain enters state i).

Definition 59 *State i is positive recurrent if it is recurrent and $\mathbb{E}\left[T_i|X_0 = i\right]$ is finite.*

Definition 60 *State i is null recurrent if it is recurrent and $\mathbb{E}\left[T_i|X_0 = i\right]$ is infinite.*

Positive recurrence means we visit a recurrent state so frequently that we spend a positive fraction of time in that state. Null recurrencce means we visit a recurrent state so infrequently (but still infinitely many times) that we spend virtually no time in that state.

Theorem 29 *Every irreducible finite state Markov Chain is positive recurrent.*

Definition 61 *For a state $i \in S$, we define the period of the state to be*

$$\textit{period}(i) = \textit{GCD}\{n \geq 1 : p_{ii}^{(n)} > 0\}.$$

If we start in state i, then revists to i only occur at integer multiples of the period.

Definition 62 *An irreducible markov chain is aperiodic if any state has period 1.*

All of the above properties are class properties.

5.1.3 Long-Term Behavior of Markov Chains

Since the p_{ij} completely characterize the Markov Chain, we can also describe what happens to the chain in the limit.

Definition 63 *A probability distribution π over the states is a stationary distribution if $\pi = \pi P$*

It is called a stationary distribution because the distribution over states is invariant with time. A Markov Chain is only at stationarity if and only if it has been started from the stationary distribution. The relationship $\pi = \pi P$ can be expanded for the jth element to show that any stationary distribution must satisfy the **Global Balance Equations**:

$$\pi_j = \sum_i p_{ij}\pi_i.$$

Note that if a distribution π satisfies the detailed balance equations from definition 56, then π also satisfies definition 63.

Both the global balance equations and detailed balance equations can be conceptualized as statements of flow. If each π_j indicates how much mass is placed on state j, then the global balance equations tell us the mass leaving the node (going to each neighbor i in proportion to p_{ij}) is equal to the mass entering the node (which must sum to π_j since it is a stationary distribution. Rather than looking at the flow of the whole chain, the detailed balance equations look at the flow between two states. The mass i gives to j is equal to the mass j gives to i.

Theorem 30 *If an irreducible Markov Chain is at stationarity, then the flow-in equals flow-out relationship holds for any cut of the Markov Chain where a cut is a partition of the chain into two disjoint subsets.*

theorem 30 is one useful result can help solve for stationary distributions.

Theorem 31 (Big Theorem for Markov Chains) *Let $(X_n)_{n\geq 0}$ be an irreducible Markov Chain. Then one of the following is true.*

1. Either all states are transient, or all states are null recurrent, and no stationary distribution exists, and $\lim_{n\to\infty} p_{ij}^{(n)} = 0$.

2. All states are positive recurrent and the stationary distribution exists, is unique, and satisfies

$$\pi_j = \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n} P_{ij}^{(k)} = \frac{1}{\mathbb{E}\left[T_j | X_0 = j\right]}.$$

If the Markov Chain is aperiodic, then $\lim_{n\to\infty} p_{ij}^{(n)} = \pi_j$

One consequence of theorem 31 is that it means the stationary distribution π of a reversible Markov Chain is unique. This makes solving the detailed balance equations a good technique of solving for the stationary distribution. If a stationary distribution exists, then we can also say when the chain will converge to the stationary distribution.

Theorem 32 (Convergence Theorem) *If a chain is irreducible, positive, recurrent, and aperiodic with stationary distribution π, then the distribution at time n $\pi_n \to \pi$*

5.2 Continuous Time Markov Chains

Definition 64 *A process $(X_t)_{t\geq 0}$ taking values in a countable state space S is a temporally homogenous continuous time markov chain if it satisfies the Markov Property*

$$Pr\left\{X_{t+\tau} = j | X_t = i, X_s = i_s, 0 \leq s \leq t\right\} = Pr\left\{X_{t+\tau} = j | X_t = i\right\} = Pr\left\{X_\tau = j | X_0 = i\right\}$$

To characterize how a CTMC functions, we need to define some additional quantities.

1. q_i is the transition rate of state i
2. p_{ij} is the transition probability bewteen states i and j

Every time a CTMC enters a state i, it will hold in that state for $\text{Exp}(q_i)$ time before transitioning to the next state j with probability p_{ij}.

Definition 65 *The jump chain is a DTMC which describes the transition probabilities between states in the CTMC*

Note that the jump chain cannot have self-loops ($p_{ii} = 0$) because otherwise the amount of time spent in state i would not be exponentially distributed. An alternative interpretation of a CTMC is

1. Define jump rates $q_{ij} = q_i p_{ij}$
2. On entering state i, jump to $j^\star = \text{argmin}_j T_j$ where $T_j \sim \text{Exp}(q_{ij})$ for all $j \neq i$ and are independent from each other.

Essentially, every time we enter a state, we set an alarm clock for all other states, and then jump to the state whose alarm clock goes off first. This equivalent interpretation allows us to summarize a CTMC using the rate matrix.

$$Q_{ij} = \begin{cases} -q_i & \text{if } i = j \\ q_{ij} & \text{if } i \neq j \end{cases}$$

Following from the first interprentation, all entries of Q are non-negative, and the rows must sum to 0. One useful quantity which we can define is how long it takes to come back to a particular state.

Definition 66 *The time to first re-entry of state j is*

$$T_j = \min\{t \geq 0 : X_t = j \textit{ and } X_s \neq j \textit{ for some } s < t\}$$

Since a CTMC is essentially a DTMC where we hold in each state for an exponential amount of time, we can apply First Step Analysis in essentially the same way that we do for DTMCs. In fact, hitting probabilities will look exactly the same since we can just use the jump chain to comute the transition probabilities. The only differences will arise when we consider the time dependent quantities. For hitting times (how long it takes to enter a state from $A \subseteq S$),

1. If $i \in A, \mathbb{E}\left[T_A | X_0 = i\right] = 0$
2. If $i \notin A, \mathbb{E}\left[T_A | X_0 = i\right] = \frac{1}{q_i} + \sum_{j \in S} p_{ij} \mathbb{E}\left[T_A | X_0 = j\right]$

5.2.1 Class Properties

Just like in DTMCs, we can classify states in the CTMC.

Definition 67 *States i and j communicate with eachc other if i and j communicate in the jump chain.*

Definition 68 *State j is transient if given $X_0 = j$, the process enters j finitely many times with probability 1. Otherwise, it is recurrent.*

Definition 69 *A state j is positive recurrent if its time to first re-entry is finite, and null recurrent otherwise.*

5.2.2 Long Term Behavior of CTMCs

CTMCs also have stationary distributions.

Definition 70 *A probability vector π is a stationary ditribution for a CTMC with rate matrix Q if*

$$\pi Q = 0 \Leftrightarrow \pi_j q_j = \sum_{i \neq j} \pi_i q_{ij}.$$

The stationary distribution of the CTMC is also related to the jump chain, but we need to normalize for the hold times.

Theorem 33 *If π is a stationary distribution for a CTMC, then the stationary distribution of the jump chain is given by*

$$\tilde{\pi}_i = \frac{\pi_i q_i}{\sum_j \pi_j q_j}$$

To describe how a CTMC behaves over time, first define $p_{ij}^{(t)} = \Pr\{X_t = j | X_0 = i\}$ and $m_j = \mathbb{E}[T_j | X_0 = j]$.

Theorem 34 (Big Theorem for CTMCs) *For an irreducible CTMC, exactly one of the following is true.*

1. All states are transient or null recurrent, no stationary distribution exists, and $\lim_{t\to\infty} p_{ij}^{(t)} = 0$

2. All states are positive recurrent, a unique stationary distribution exists, and the stationary distribution satisfies

$$\pi_j = \frac{1}{m_j q_j} = \lim_{t\to\infty} p_{ij}^{(t)}$$

5.2.3 Uniformization

Let $P^{(t)}$ denote the matrix of transition probabiltiies at time $t > 0$. By the Markov property, we know that $P^{(s+t)} = P^{(s)}P^{(t)}$. For $h \approx 0, P^{(h)} \approx I + hQ + o(h)$. This approximation allows us to compute the derivative of $P^{(t)}$.

Theorem 35 (Forward Kolmogorov Equation)

$$\frac{\partial}{\partial t}P^{(t)} = \lim_{h \to 0} \frac{P^{(t+h)} - P^{(t)}}{h} = P^{(t)}Q$$

theorem 35 tells us that the transition probabilties $P^{(t)} = e^{tQ}$ for all $t \geq 0$. This is why Q is sometimes called the generator matrix: it generates the transition probabilities. However, matrix exponentials are difficult to compute. Instead, we can turn to **Uniformization**, which allows us to estimate $P^{(t)}$ by simulating it through a DTMC.

Definition 71 *Given a CTMC where $\exists M$ such that $q_i \leq M$ for all $i, j \in S$. Fix a $\gamma \geq M$, and the uniformized chain will be a DTMC with transition probabilities $p_{ij} = \frac{q_{ij}}{\gamma}$ and $p_{ii} = 1 - \frac{q_i}{\gamma}$.*

$$P_u = I + \frac{1}{\gamma}Q.$$

It turns out that

$$P_u^n = \left(I + \frac{1}{\gamma}Q\right)^n \approx e^{\frac{n}{\gamma}Q}$$

when $\frac{1}{\gamma}$ is small. This means that we can approximate the transition probabilties of the CTMC using the uniformized chain. Observe that uniformization also helps in finding the stationary distribution since the stationary distribution of the uniformized chain is identical to the original chain.

$$\pi P_u = \pi + \frac{1}{\gamma}\pi Q = \pi \Leftrightarrow \pi Q = 0.$$

5.2.4 Poisson Processes

Definition 72 *A counting process $(N_t)_{t \geq 0}$ is a non-decreasing, continuous time, integer valued random process which has right continuous sample paths.*

There are two important metrics which describe counting processes.

Definition 73 *The ith arrival time T_i is given by*

$$T_i = \min_t \{t \geq 0 : \; N_t \geq i\}$$

Definition 74 *The ith inter-arrival time S_i is given by*

$$S_i = T_i - T_{i-1}, i > 0$$

Definition 75 *A rate λ Poisson Process is a counting process with independently and identically distributed inter-arrival times $S_i \sim$ Exp(λ).*

The name Poisson comes from the distribution of each varible in the process.

Theorem 36 *If $(N_t)_{t\geq 0}$ is a rate λ Poisson Process, then for each $t \geq 0$, $N_t \sim$ Poisson(λt)*

A Poisson Process is a special case of a CTMC where the transition rates $q_i = \lambda$ and the transition probabilties p_{ij} are 1 if $j = i + 1$ and 0 otherwise. Since the inter-arrival times are memoryless and i.i.d, Poisson Processes have many useful properties.

Theorem 37 *If $(N_t)_{t\geq 0}$ is a rate λ Poisson Process, then $(N_{t+s} - N_s)_{t\geq 0}$ is also a rate λ Poisson Process for all $s \geq 0$ and is independent of the original process.*

Theorem 38 *For $t_0 < t_1 < \ldots < t_k$, then the increments of a rate λ Poisson Process $(N_{t_1} - N_{t_0}), (N_{t_2} - N_{t_1}), \ldots, (N_{t_k} - N_{t_{k-1}})$ are independent and $N_{t_i} - N_{t_{i-1}} \sim$ Poisson$(\lambda(t_i - t_{i-1}))$*

Poisson Processes are the only counting processes with these particular properties.

It turns out that Poisson Processes can be connected with the Order Statistics of Uniform Random Variables.

Theorem 39 (Conditional Distribution of Arrivals) *Conditioned on $N_t = n$, the random vector $T_1, T_2, \cdots, T_n$ has the same distribution as the order statistics of n random variables $U \sim$ Uniform$(0, t)$.*

What theorem 39 says is that given n arrivals up to time t occur, the distribution of arrival times is equivalent to taking n i.i.d uniform random variables and sorting them.

Two other useful properties of Poisson Processes involve combining and separating them.

Theorem 40 (Poisson Merging) *If $N_{1,t}$ and $N_{2,t}$ are independent Poisson Processes with rates λ_1 and λ_2, then $N_{1,t} + N_{2,t}$ is a Poisson Process with rate $\lambda_1 + \lambda_2$.*

Theorem 41 (Poisson Splitting) *Let $p(x)$ be a probability distribution and N_t be a rate λ Poisson process. If each arrival is marked with the label i independently with probability $p(x = i)$, then $N_{i,t}$, the process counting the number of arrivals labeled i is an independent Poisson Process with rate λp_i.*

6 Random Graphs

A random graph is one which is generated through some amount of randomness.

Definition 76 *An Erdos-Renyi random graph $G(n, p)$ is an undirected graph on $n \geq 1$ vertices where each edge exists independently with probability p.*

With random graphs, we often ask what happens to particular properties as $n \to \infty$ and p scales with some relationship to n. In particular, we want that property to hold with high probability (i.e, as $n \to \infty$, the probabilty that $G(n, p)$ has the property approaches 1).

Theorem 42 *Every monotone graph property (adding more edges doesn't delete the property) has a sharp threshold t_n where if $p \gg t_n$, then $G(n, p)$ has p with high probability and does not have p with high probability if $t_n \ll G(n, p)$.*

One example of a threshold is the connectivity threshold.

Theorem 43 (Erdos-Renyi Connectivity Theorem) *Fix $\lambda > 0$ and let $P_n = \lambda \frac{\log n}{n}$. If $\lambda > 1$, then $P(G(n, p)$ is connected$)$ with probability approaching 1, and if $\lambda < 1$, then $P(G(n, p)$ is disconnected$)$ with probability approaching 1*

7 Statistical Inference

Suppose we have a variable X (may or may not be a random variable) that represents the state of nature. We observe a variable Y which is obtained by some model of the world $P_{Y|X}$.

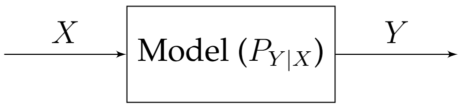

Figure 2: Inference Setup

Suppose we know that $X \sim \pi$ where π is a probability distribution. If we observe $Y = y$, then the *a posteriori* estimate of X is given by Bayes Rule

$$\Pr\{X = x | Y = y\} = \frac{P_{Y|X}(y|x)\pi(x)}{\sum_{\tilde{x}} P_{Y|X}(y|\tilde{x})\pi(\tilde{x})} \propto P_{Y|X}(y|x)\pi(x).$$

Since the estimate is only dependent on the model and the prior, we don't actually need to compute the probabilities to figure out the most likely X.

Definition 77 *The Maximum A Posteriori (MAP) estimate is given by*

$$\hat{X}_{MAP}(y) = \underset{x}{\operatorname{argmax}}\, P_{Y|X}(y|x)\pi(x)$$

If we have no prior information on X, then we can assume π is uniform, reducing definition 77 to only optimize over the model.

Definition 78 *The Maximum Likelihood (ML) estimate is given by*

$$\hat{X}_{ML}(y) = \underset{x}{\operatorname{argmax}}\, P_{Y|X}(y|x)$$

7.1 Binary Hypothesis Testing

Definition 79 *A Binary Hypothesis Test is a type of statistical inference where the unknown variable $X \in \{0, 1\}$.*

Since there are only two possible values of X in a binary test, there are two "hypotheses" that we have, and we want to accept the more likely one.

Definition 80 *The Null Hypothesis H_0 says that $Y \sim P_{Y|X=0}$*

Definition 81 *The Alternate Hypothesis H_1 says that $Y \sim P_{Y|X=1}$*

With two possible hypotheses, there are two kinds of errors we can make.

Definition 82 *A Type I error (false positive) is when we incorrectly reject the null hypothesis. The Type I error probability is then*

$$Pr\left\{\hat{X}(Y) = 1 | X = 0\right\}$$

Definition 83 *A Type II error (false negative) is when we incorrectly accept the null hypothesis. The Type II error probability is then*

$$Pr\left\{\hat{X}(Y) = 0|X = 1\right\}$$

Our goal is to create a decision rule $\hat{X} : \mathcal{Y} \to \{0, 1\}$ that we can use to predict X. Based on what the decision rule is used for, there will be requirements on how large the probability of Type I and Type II errors can be. We can formulate the search for a hypothesis test as an optimization. For some $\beta \in [0, 1]$, we want to find

$$\hat{X}_\beta(Y) = \text{argmin}\,\text{Pr}\left\{\hat{X}(Y) = 0|X = 1\right\} \quad : \quad \text{Pr}\left\{\hat{X}(Y) = 1|X = 0\right\} \leq \beta. \qquad (1)$$

Intuitively, our test should depend on $p_{Y|X}(y|1)$ and $p_{Y|X}(y|0)$ since these quantities give us how likely we are to get our observations if we knew the ground truth. We can define a ratio that formally compares these two quantities.

Definition 84 *The likelihood ratio is given by*

$$L(y) = \frac{p_{Y|X}(y|1)}{p_{Y|X}(y|0)}$$

Notice that we can write MLE as a threshold on the likelihood ratio since if $L(y) \geq 1$, then we say $X = 1$, and vice versa. However, there is no particular reason that 1 must always be the number at which we threshold our likelihood ratio, and so we can generalize this idea to form different forms of tests.

Definition 85 *For some threshold c and randomization probability γ, a threshold test is of the form*

$$\hat{X}(y) = \begin{cases} 1 & \textit{if } L(y) > c \\ 0 & \textit{if } L(y) < c \\ \textit{Bernoulli}(\gamma) & \textit{if } L(y) = c. \end{cases}$$

MAP fits into the framework of a threshold test since we can write

$$\hat{X}_{MAP} = \begin{cases} 1 & \text{if } L(y) \geq \frac{\pi_0}{\pi_1} \\ 0 & \text{if } L(y) < \frac{\pi_0}{\pi_1} \end{cases}$$

It turns out that threshold tests are optimal with respect to solving eq. (1).

Theorem 44 (Neyman Pearson Lemma) *Given $\beta \in [0, 1]$, the optimal decision rule to*

$$\hat{X}_\beta(Y) = \operatorname{argmin} Pr\left\{\hat{X}(Y) = 0 | X = 1\right\} \quad : \quad Pr\left\{\hat{X}(Y) = 1 | X = 0\right\} \leq \beta$$

is a threshold test.

When $L(y)$ is monotonically increasing or decreasing, we can make the decision rule even simpler since it can be turned into a threshold on y. For example, if $L(y)$ is monotonically inreasing, then an optimal decision rule might be

$$\hat{X}(y) = \begin{cases} 1 & \text{if } y > c \\ 0 & \text{if } y < c \\ \text{Bernoulli}(\gamma) & \text{if } y = c. \end{cases}$$

8 Estimation

Whereas hypothesis testing is about discriminating between two or more hypotheses, estimation is about guessing the numerical value of or ground truth of a random variable.

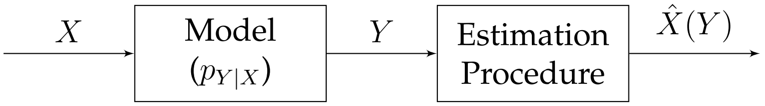

Figure 3: Estimation Setup

In order to measure the quality of our estimation, we need a metric to measure error. One commonly used error is the mean squared error

$$\mathbb{E}\left[(X - \hat{X}(Y))^2\right].$$

Theorem 45 *The minimum mean square estimate (MMSE) of a random variable X is given by the conditional expectation.*

$$\hat{X}(Y) = \mathbb{E}\left[X|Y\right] = \underset{\hat{X}}{\operatorname{argmin}}\, \mathbb{E}\left[(X - \hat{X}(Y))^2\right].$$

This essentially follows from the definition of conditional expectation since it is orthogonal to all other functions of Y, and so by the Hilbert Projection Theorem, it must be the projection of X onto the space of all functions of Y. There are two problems with using MMSE all the time.

1. We often don't know $p_{Y|X}$ explicitly and only have a good model for it.
2. Even if we knew the model $p_{Y|X}$, conditional expectations are difficult to compute.

8.1 Linear Estimation

Since finding the MMSE is difficult, we can restrict ourselves to funtions of a particular type.

Definition 86 *The Linear Least Squares Estimator (LLSE)* $\mathbb{L}\left[\boldsymbol{X}|\boldsymbol{Y}\right]$ *is the projection of a vector of random variables* $\boldsymbol{X}$ *onto the subspace of linear functions of observations* Y_i, $\mathcal{U} = \{\boldsymbol{a} + B\boldsymbol{Y}\}$ *where* $\boldsymbol{Y}$ *is a vector of observations.*

By the orthogonality principle,

1. $\mathbb{E}\left[(\boldsymbol{X} - \mathbb{L}\left[\boldsymbol{X}|\boldsymbol{Y}\right])1\right] = 0 \implies \mathbb{E}\left[\mathbb{L}\left[\boldsymbol{X}|\boldsymbol{Y}\right]\right] = \mathbb{E}\left[\boldsymbol{X}\right]$
2. $\mathbb{E}\left[(\boldsymbol{X} - \mathbb{L}\left[\boldsymbol{X}|\boldsymbol{Y}\right])Y_i\right] = 0$

From here, we can derive a closed form expression for the LLSE. Let $\boldsymbol{\mu_Y} = \mathbb{E}\left[\boldsymbol{Y}\right], \boldsymbol{\mu_X} = \mathbb{E}\left[\boldsymbol{X}\right], \Sigma_{\boldsymbol{Y}} = \mathbb{E}\left[(\boldsymbol{Y} - \boldsymbol{\mu_Y})(\boldsymbol{Y} - \boldsymbol{\mu_Y})^T\right], \Sigma_{\boldsymbol{XY}} = \mathbb{E}\left[(\boldsymbol{X} - \boldsymbol{\mu_X})(\boldsymbol{Y} - \boldsymbol{\mu_Y})^T\right]$. By substituting $\mathbb{L}\left[\boldsymbol{X}|\boldsymbol{Y}\right] = \boldsymbol{a} + B\boldsymbol{Y}$ into the equations we found from the orthogonality principle,

$$\begin{aligned} \boldsymbol{a} + B\boldsymbol{\mu_Y} &= \boldsymbol{\mu_X} \\ a(\boldsymbol{\mu_Y})_i + B\mathbb{E}\left[\boldsymbol{Y}Y_i\right] = \mathbb{E}\left[\boldsymbol{X}Y_i\right] &\implies \boldsymbol{a}(\boldsymbol{\mu_Y})_i + B(\Sigma_{\boldsymbol{Y}})_i + B(\boldsymbol{\mu_Y})_i\boldsymbol{\mu_Y} = (\Sigma_{\boldsymbol{XY}})_i + (\boldsymbol{\mu_Y})_i\boldsymbol{\mu_x} \\ &\implies \boldsymbol{a}\boldsymbol{\mu_Y}^T + B\Sigma_{\boldsymbol{Y}} + B\boldsymbol{\mu_Y}\boldsymbol{\mu_Y}^T = \Sigma_{\boldsymbol{XY}} + \boldsymbol{\mu_X}\boldsymbol{\mu_Y}^T \end{aligned}$$

Solving this system yields

$$B = \Sigma_{\boldsymbol{XY}}\Sigma_{\boldsymbol{Y}}^{-1} \qquad \boldsymbol{a} = \boldsymbol{\mu_X} - \Sigma_{\boldsymbol{XY}}\Sigma_{\boldsymbol{Y}}^{-1}\boldsymbol{\mu_Y}.$$

Theorem 46 *The Linear Least Squares Estimator for vector of random variables* $\boldsymbol{X}$ *given a vector of random variables* $\boldsymbol{Y}$ *is*

$$\mathbb{L}\left[\boldsymbol{X}|\boldsymbol{Y}\right] = \boldsymbol{\mu_X} + \Sigma_{\boldsymbol{XY}}\Sigma_{\boldsymbol{Y}}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu_Y})$$

If X and Y are both a single random variable, this reduces to

$$\mathbb{L}\left[X|Y\right] = \mu_X + \frac{Cov\left(X, Y\right)}{Var\left(Y\right)}(Y - \mu_Y)$$

Since LLSE is essentially projection onto a Linear Subspace, if we have an orthogonal basis for the subspace, then we can do the projection onto the subspace one component at a time. The Gram-Schmidt Process turns vectors $Y_1, \cdots, Y_n$ into an orthonormal set $\tilde{Y}_1, \cdots, \tilde{Y}_n$. If we define $Y^{(n)} = (Y_1, \cdots, Y_n)$,

1. $\tilde{Y}_1 = \frac{Y_1}{\|Y_1\|}$
2. $\tilde{Y}_{i+1} = Y_{i+1} - \sum_{k=1}^{i} \langle Y_{i+1}, \tilde{Y}_k \rangle \tilde{Y}_k = Y_{i+1} - \mathbb{L}\left[Y_{i+1} | Y^{(i)}\right]$

Definition 87 *The linear innovation sequence of random variables $Y_1, \cdots, Y_n$ is the orthogonal set $\tilde{Y}_1, \cdots, \tilde{Y}_n$ produced by Gram Schmidt*

Since $\tilde{Y}_n$ is orthogonal to $\mathbb{L}\left[Y_n | \tilde{Y}^{(n-1)}\right]$, they belong to different parts of the subspace formed by $Y_1, \cdots, Y_n$.

Theorem 47

$$\mathbb{L}\left[X | Y^{(n)}\right] = \mathbb{L}\left[X | \tilde{Y}_n\right] + \mathbb{L}\left[X | \tilde{Y}^{(n-1)}\right]$$

Note that in general, the LLSE is not the same as the MMSE. However, if X and Y are Jointly Gaussian, then the LLSE does, in fact, equal the MMSE.

8.2 Kalman Filtering

Definition 88 *A system evolves according to a state space model if the state $\boldsymbol{X}_n$ at time n and observations $\boldsymbol{Y}_n$ at time n are related by*

$$\forall n \geq 0,\ \boldsymbol{X}_{n+1} = A\boldsymbol{X}_n + \boldsymbol{V}_n \qquad \forall n \geq 1,\ \boldsymbol{Y}_n = C\boldsymbol{X}_n + \boldsymbol{W}_n$$

where V_n and W_n are noise terms.

State space models are flexible and describe a variety of processes. Suppose we want to linearly estimate $\boldsymbol{X}_n$ from the $\boldsymbol{Y}_n$ we have seen so far.

Theorem 48 *The linear estimate $\hat{\boldsymbol{X}}_{n|n} = \mathbb{L}\left[\boldsymbol{X}_n | \boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_n\right]$ can be computed recursively via the Kalman Filter.*

1. $\hat{\boldsymbol{X}}_{0|0} = 0, \Sigma_{0|0} = Cov\left(\boldsymbol{X}_0\right)$.

2. For $n \geq 1$, update

$$\hat{\boldsymbol{X}}_{n|n} = A\hat{\boldsymbol{X}}_{n-1|n-1} + K_n\tilde{\boldsymbol{Y}}_n \qquad \tilde{\boldsymbol{Y}}_n = Y_n - C\hat{\boldsymbol{X}}_{n|n-1} \qquad \Sigma_{n|n-1} = A\Sigma_{n-1|n-1}A^T + \Sigma_{\boldsymbol{V}}$$

$$K_n = \Sigma_{n|n-1}C^T(C\Sigma_{n|n-1}C^T + \Sigma_{\boldsymbol{W}})^{-1} \qquad \Sigma_{n|n} = (I - K_nC)\Sigma_{n|n-1}$$

Kalman filtering is a simple algorithm which lets us do online, optimal estimation. Variants of it can do things such as prediction or smoothing.