EECS 126

Albert Ye

November 30, 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 $A_1, A_2, \ldots, A_n$ 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 $A_1, A_2, \ldots, A_n, \ldots \in \mathcal{F}$, 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 $X_1, \ldots, X_n$ 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)

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

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.

10 Markov Chains

Definition 27 (Markov Chain)
$\{X_n\}_{n \in \mathbb{N}}$ is a discrete-time Markov Chain (DTMC) on state space $\mathcal{X}$ if it satisfies the Markov property: For all positive integers n and feasible sequence of states $x_0, x_1, x_2, \ldots, x_{n+1} \in \mathcal{X}$;

$$\Pr(X_{n+1} = x_{n+1} | X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_0 = x_0) = P(X_{n+1} = x_{n+1}, X_n = x_n).$$

We further denote P as the transition probability matrix, which is done by taking the row statistic of $\mathcal{X}$.

10.1 Distributions

Denote distribution of X_n as Π_n. Then, $\Pi_n = \Pi_0 P^n$. We have a **stationarity distribution** $\Pi = \Pi \cdot P$, and this is also called the balance equation.

10.2 Recurrence and Transience

For $x \in \mathcal{X}$, we define $T_x = \min\{n \in \mathbb{N}, X_n = n\}$ as the hitting time of x, and $T_x^+ = \min\{n \in \mathbb{Z}_+ X_n = n\}$.

T_x determines the first time that a Markov chain reaches a certain state, and T_x^+ calculates the same thing except ignoring trivial (initial) cases.

Now, some notation. Let $\Pr_x(A) = \Pr(A|X_0 = x)$ and $E_x[Z] = E[Z|X_0 = x]$. This is probability and expectation given an initial state in the Markov chain. Furthermore, let $\rho_{x,y} = \Pr_x(T_y^+ < \infty)$, $\rho_x = \rho_{x,x}$.

Definition 28

State x is **recurrent** if $\rho_x = 1$, **transient** otherwise.

A recurrent state essentially means that a state in a Markov chain will certainly be reached again.

Proposition 29

Denote $N_x = \sum_{n \in \mathbb{N}} \mathbb{I}(X_n = x)$. Then,

1. If x is recurrent, then $N_x = \infty$ almost surely.
2. If x is transient then $E_x[N_x] = \frac{\rho(x)}{1-\rho(x)}$.

10.3 Classification of States

Definition 30 (Communicating Class)

We say x communicates with y if $\rho_{x,y} > 0$ and $p_{y,x} > 0$.

A **communicating class** is a maximal set of states which communicate with each other.

Definition 31

Markov Chain is **irreducible** if it consists of only a single communicating class.

The class property is a property that's necessarily shared by all members of class. Anyways, now time to start applying the many definitions we've just made:

Theorem 32

Recurence and transience are class properties.

Are we not going over the proof for this?

Proposition 33

Every finite state irreducible chain is recurrent.

Proof. Basically prove that one of the states must be recurrent using the fact that there are finite states, and then use the above theorem to see that this is a class property. □

10.4 Big Theorem

Theorem 34

Suppose a markov chain is irreducible with a stationary distribution Π. Then,

$$\Pi(x) = \frac{1}{\mathbb{E}_X[T_X^+]}$$

To prove this, we introduce another claim.

Theorem 35

Suppose a Markov chain is irreducible, aperiodic, and has stationary distribution Π. Then, as $n \to \infty$, $P_n(x, y) \to \Pi(y)$ for all x, y.

The **aperiodic** assumption is correct, because if the result is periodic, then it is clear to see that this convergence is not true.

Moreover, $P_n(x, y) = \Pr(X_n = y \mid x_0 = x)$.

Proof. Let $\mathcal{X}^2 = \mathcal{X} \times \mathcal{X}$, and we define a new transition probability $\bar{P}$ on $\mathcal{X}^2$. Then, we define $\bar{P}((x_1, y_1), (x_2, y_2)) = P(x_1, x_2)P(y_1, y_2)$. We claim that $\bar{P}$ is irreducible.

Since P is irreducible, there exist K, L such that $P_K(x_1, x_2) > 0$, $P_L(y_1, y_2) > 0$.

Lemma 10.4.1

For irreducible aperiodic Markov Chain there exists m_0 such that $P_m(x, x) > 0$ for all $m > m_0$, where m_0 depends on x.

Proof.

$$\begin{aligned}
\Pr(X_n = y, T \le n) &= \sum_{m=1}^{n} \sum_x \Pr(T = m, X_m = x, Y_n = y) \\
&= \sum_{m=1}^{n} \sum_x \Pr(T = m, X_m = x) P(Y_n = y | X_m = x, T = m) \\
&= \sum_{m=1}^{n} \sum_x \Pr(T = m, X_m = x) P(Y_n = y | Y_m = x) \\
&= \Pr(Y_n = y, T \le n).
\end{aligned}$$

We can extend the Markov property here by applying it recursively to state that conditioned on some event X_m in Markov chain, a future event X_n is conditionally independent of **all** past events $(X_0, \ldots, X_{m-1})$. $\square$

Using the aperiodicity lemma, we know that for M large enough, $P_{K+M}(x_1, x_2) > 0$, and $P_{L+M}(y_1, y_2) > 0$. It then follows that $\bar{P}_{K+L+M}((x_1, y_1), (x_2, y_2)) > 0$.

I honestly am completely lost for the rest of the proof I'll figure it out later... $\square$

10.5 Reversibility

Asking the following question: Does the Markov Chain still work when played in reverse?

We let $(Y_0, Y_1, \ldots, Y_n) \equiv (X_n, X_{n-1}, \ldots, X_0)$.

Lemma 10.5.1

Y_n is still a Markov Chain with transition matrix $\hat{P}$, where

$$\hat{P}(x,y) = \frac{\Pi(y)P(y,x)}{\Pi(x)}.$$

Definition 36 (Reversibility)

We say that a Markov chain is **reversible** if $\hat{P} = P$.

The detailed balance equation states that $\Pi(x)P(x,y) = \Pi(y)P(y,x)$.

11 Poisson Process

This is based on exponential distributions having the memoryless property.

So if we have $X \sim \text{Exp}(\lambda)$, then $f_X(x) = \lambda e^{-\lambda x}, x \geq 0$, then the CDF is $P(x > t) = e^{-\lambda t}$. We then have the following property: $\Pr(X > t + s \mid X > t) = \Pr(X > s)$.

Definition 37 (Poisson Process)

Fix $\lambda > 0$. Assume that inter-arrival times $s_1, s_2, \ldots$ are i.i.d. $\text{Exp}(\lambda)$. For each $x \geq 1$, define

$$T_n = \sum_{j=1}^{n} S_j, T_0 = 0.$$

Moreover,

$$N(t) = \max(n > 0 : T_n \leq t).$$

We call the continuous time stochastic process $\{N(t)\}_{t>0}$ the **Poisson process** $\text{PP}(\lambda)$.

Next, we define the Big Theorem of Poisson processes with its primary key properties.

Definition 38 (Increment of Poisson Process)

$$N(T_1, T_2) = N(_2) - N(T_1), T_2 \geq T_1.$$

Theorem 39 (Big Theorem)

1. **Stationary Increment.** $N(t, t+s)$ has the same distribution as $N(s)$.
2. **Independent Increment.** For $0 < t_1 < t_2 < \ldots < t_k$ the set of random variables $N(t_1), N(t_1, t_2), \ldots, N(t_{k-1}, t_k)$ are jointly independent.
3. $N(t) \sim \text{pois}(\lambda t)$.

We can generalize Poisson processes into multiple dimensions with the Poisson random field, so $\text{pois}(\int_A \lambda)$ finds the Poisson process over arrivals in a region A.

Definition 40 (Erlang Distribution)

$$f_{T_n}(t) = \frac{\lambda^n t^{n-1} e^{\lambda t}}{(n-1)!}$$

There are two ways to look at a Poisson process: the counts and the inter-arrivals.

Theorem 41

Let $S_1, S_2, \dots$ be some set of almost surely positive inter-arrival times and define $T_n = \sum_{j=1}^{n} S_j$, $N(t) = \max(n \geq 0 : T_n \leq t)$.

If $\{N(t)\}_{t \geq 0}$ has stationary independent increments, and $N(t) \sim \text{pois}(\lambda t)$, then $S_1, S_2, \dots$ are i.i.d. $\text{Exp}(\lambda)$ random variables.

Definition 42 (Splitting)

$N \sim \text{PP}(\lambda)$, and $B_1, B_2, \dots \sim \text{Bern}(p)$.

The splitting process essentially assigns to one of N_0 or N_1.

$$N_0(t) = |\{i : B_i = 0, i \leq N(t)\}|$$

$$N_1(t) = |\{i : B_i = 1, i \leq N(t)\}|$$

Then $N_1 \sim \text{PP}(\lambda p), N_0 \sim \text{PP}(\lambda(1-p))$, so without being given any knowledge of N, we have that N_0, N_1 are independent.

Essentially what is happening is that when something arrives, we flip a (weighted) coin and flip a switch to determine which process it actually reaches.

11.1 Random Incidence Property

Let $N \sim \text{PP}(\lambda)$.

Then,

1. The expected interarrival time is $\frac{1}{\lambda}$.
2. Fix time t in the process, what is the expected length of the interarrival interval which t falls into?

Then, we want to find

$$\begin{aligned}
\mathbb{E}[T_{i+1} - T_i] &= \mathbb{E}[(t - T_i) + (T_{i+1} - t)] \\
&= \mathbb{E}[(t - T_i)] + \mathbb{E}[T_{i+1} - t] \\
&= \mathbb{E}[(t - T_i)] + \frac{1}{\lambda}.
\end{aligned}$$

Now, fix a value τ. Then, $\Pr(t - T_i > T) = \Pr(N(t-\tau, t) = 0) = \Pr(N(\tau) = 0) = \Pr(\text{pois}(\lambda\tau) = 0) = e^{-\lambda\tau}$. As a result, $\Pr(t - T_i > \tau) = \begin{cases} e^{-\lambda\tau} & 0 \leq \tau \leq t \\ 0 & \tau > t \end{cases}$.

Taking the integrals, we see that

$$\mathbb{E}[t - T_i] = \int_0^\infty \Pr(t - T_i > \tau) d\tau = \int_0^t e^{-\lambda\tau} d\tau = \frac{1 - e^{-\lambda t}}{\lambda}.$$

As t goes to λ, then this converges to $\frac{1}{\lambda}$.

Thus, $\mathbb{E}[T_{i+1} - T_i] = \boxed{\frac{2}{\lambda}}$.

12 Estimators

12.1 Hilbert Space

12.2 Gram-Schmidt

12.3 LLSE

12.4 MMSE

13 Kalman Filter

Start with discrete time linear system models.

Example 43

We have a particle moving along a line at fixed velocity. Observed every Δ time units.

So, we have $\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = \begin{bmatrix} x_k + y_k\Delta \\ y_k \end{bmatrix} = \begin{bmatrix} 1 & \Delta \\ 0 & 1 \end{bmatrix}$.

Now we have $x_{k+1} = Ax_k$, for some matrix A and $x_k \in \mathbb{R}^2$.

Today we'll cover this example but with noise, i.e. $x_{k+1} = Ax_k + V_k$.

Example 44

Particle moving under fixed acceleration, once again with discrete time and time step Δ. Then,

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \\ z_{k+1} \end{bmatrix} = \begin{bmatrix} x_k + y_k\Delta \\ y_k + z_k\Delta \\ z_k \end{bmatrix}.$$

We can then get a similar matrix:

$$\begin{bmatrix} 1 & \Delta & 0 \\ 0 & 1 & \Delta \\ 0 & 0 & 1 \end{bmatrix}.$$

More generally, in continuous time we have a dynamical system $\left[\frac{d}{dt}x\right] = Ax(t)$ for some matrix A starting at $X(0)$. The solution is $x(t) = e^{At}x(0)$, using the taylor series $e^{At}\sum_{n=1}^{\infty}\frac{(At)^n}{n!}$.

In discrete time, $x_{k+1} = Ax_k$ where $A = e^{A_c\Delta}$.

13.1 Noisy dynamics model

Let's return to our original value, $x_{k+1} = Ax_k + V_k$, where V_k is the noise. We now assume that the state cannot be measured. We have the model for observation: $Y_k = CX_k + W_k$, where C is a fixed scalar and W_k is the observation noise.

We're interested in computing $L[X_n \mid Y_1, \ldots, Y_{n-1}], n \geq 0$. At $n = 0$, this means $\mathbb{E}[X_0] = 0$. We'll see that we need $L[X_n \mid Y_1, \ldots, Y_{n-1}$ as an intermediary. This can be called a **1-step predictor**.

Recall: If (U, V) are jointly defined, $U \in \mathbb{R}^m$ and $V \in \mathbb{R}^n$, then $L[U \mid V]$ denotes the LLSE estimate of U given V, and has the form KV for $K \in \mathbb{R}^{m\times n}$, where $\mathbb{E}[(U - KV)V^T] = 0$, i.e. $\Sigma_{UV} - K\Sigma_V = 0$.

Example 45

Suppose $V = \begin{bmatrix} V_1 \\ V_2 \\ V_3 \end{bmatrix}$, where $V_3 = V_1 + V_2$, for $n = 3, m = 1$. Say $V = V_1 + V_2$. Then, $K = \begin{bmatrix} 1 & 1 & 0 \end{bmatrix}$ gives $KV = V_1 + V_2$. But $K = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}$ also gives the same thing.

We have that $L[U \mid V] = V_1 + V_2$.

The most important property we'll use:

Lemma 13.1.1 (Orthogonal Updates)

$$L[X \mid Y, Z] = L[X \mid Y]L[X \mid Z - L[Z \mid Y]].$$

Essentially, we're finding an orthogonal basis of the span of Y and Z, and then add the individual projections.

It suffices to prove that

$$\mathbb{E}[(X - L[X \mid Y] - L[X \mid W])Y^T] = 0,$$
$$\mathbb{E}[(X - L[X \mid Y] - L[X \mid W])Z^T] = 0.$$

13.1.1 Scalar Case

First, some notation:

Definition 46

Let $\hat{x}$ be the observation and $\tilde{x}$ be the error.

We see that $x_n = ax_{n-1} + v_n$, for $n \geq 1$. Then, $y_n = x_n + w_n$. All the w_n are i.i.d. with $\mathbb{E}[w_n] = 0$, and $\mathbb{E}[w_n^2] = \sigma_w^2 < \infty$. Notice we've ignored the constant C here, this is because if C is nonzero, it doesn't matter, and if $C = 0$, then the data is just noise and completely meaningless.

From the key lemma: $L[X_n \mid y_1, \ldots, y_n] = L[x_n \mid y_1, \ldots, y_{n-1}] + L[x_n \mid \hat{y}_n]$.

$\hat{x}_{n|n} = \hat{x}_{n|n-1} + k_n y_n$. We need to write $\hat{x}_{n|n-1}$ in terms of $x_{n-\hat{1}|n-1}$. We have

$$\hat{x}_{n|n} = L[x_n \mid y_1, \ldots, y_{n-1}] = L[ax_{n-1} \mid y_1, \ldots, y_{n-1}] = a\hat{x}_{n-1|n-1}.$$

13.2 Innovation and Gain

We call the projection of x_n onto $\tilde{y_n}$ $k_n\tilde{y_n}$, as it's a linear function of $\tilde{y_n}$.

$$\hat{x}_{n|n} = L[x_n \mid x_1, \ldots, x_n] = L[x_n \mid x_1, \ldots, x_{n-1}] + L[x_1 \mid \tilde{[}x_1 \mid \tilde{y_n}].$$

Definition 47 (Kalman Gain)

$$k_n = \frac{\sigma_{n|n}^2}{\sigma_{n|n-1}^2 + \sigma_w^2}$$

Proof. We first find $\sigma_{n|n}$.

$$\begin{aligned}
\sigma_{n|n}^2 &= \mathbb{E}[(x_n - \hat{x}_{n|n})^2] \\
&= \mathbb{E}[(x_n - (\hat{x}_{n|n-1} + k_n \hat{y_n})] \\
&= \mathbb{E}[((1-k_n)\hat{x}_n - k_n w_n)^2] \\
&= (1-k_n)^2 \sigma_{n|n-1}^2 + k_n^2 \sigma_w^2.
\end{aligned}$$

□