Stochastic Processes

Jonathan Ma
johnma@udel.edu

March 21, 2025

Contents

Preface

1 Review of Probability

1.1 Why Measure Theory

Consider a probability space $(\Omega, \mathcal{F}, P)$, where Ω is any set, $\mathcal{F} \subset 2^\Omega$ is a σ-algebra over Ω, and $P : \mathcal{F} \to [0,1]$ is a probability measure. We will call sets in $\mathcal{F}$ events. Recall that $P(\Omega) = 1$, and P is countably additive:

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n),$$

when each $A_i \cap A_j = \varnothing$. Then P is a measure with total mass 1. (Some texts call $\mathcal{F}$ a σ-field, which means the same as σ-algebra.)

Example 1.1. If Ω is a nonempty set, and $x \in \Omega$, then the point mass at x is

$$\delta_x(A) = \begin{cases} 1 & x \in A \\ 0 & x \notin A, \end{cases}$$

for all $A \in \mathcal{F}$. If we take $\mathcal{F} = 2^\Omega$, then δ_x is a probability measure. (Homework.)

Example 1.2. Let $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ be finite, and let $\mathcal{F} = 2^\Omega$. Suppose $P_1, \ldots, P_n \in [0,1]$, with $\sum_{i=1}^n P_i = 1$. Define

$$P(A) = \sum_{i:\omega_i \in A} P_i,$$

for $A \in \mathcal{F}$. Then P is a probability measure.

Example 1.3. If $f : \mathbb{R} \to [0, \infty]$ is such that

$$\int_{-\infty}^{\infty} f(x)\, dx = 1.$$

We would like to be able to define P such that if $a < b$, then $P([a,b]) = \int_a^b f(x)\, dx$. To claim such a probability measure exists and identify the proper σ-algebra $\mathcal{F}$ is the task of measure theory in probability theory.

Definition 1.4. The Borel σ-algebra on $\mathbb{R}$, denoted $\mathcal{B}$, is the smallest σ-algebra containing all open subsets of $\mathbb{R}$. Equivalently, $\mathcal{B}$ contains all intervals of the form $[a,b]$.

2 Measure Theory

Proposition 2.1. *Let $(\Omega, \mathcal{F}, P)$ be a probability space, and suppose that $A, B \in \mathcal{F}$. Then*

1. $P(\emptyset) = 0$,
2. $P(A) = 1 - P(A^C)$,
3. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
4. *If $A \subset B$, then $P(A) \leq P(B)$.*

Proposition 2.2. *Suppose $\{A_n\} \subset \mathcal{F}$. Then*

1. $\lim_{n\to\infty} P(\bigcup_{i=1}^{n} A_i) = P(\bigcup_{n=1}^{\infty} A_i)$,
2. $\lim_{n\to\infty} P(\bigcap_{i=1}^{n} A_i) = P(\bigcap_{n=1}^{\infty} A_n)$.
3. *If $A_1 \subset A_2 \subset \ldots$, then* $\lim_{n\to\infty} P(A_n) = P(\bigcup_{n=1}^{\infty} A_n)$.
4. *If $A_1 \supset A_2 \supset \ldots$, then* $\lim_{n\to\infty} P(A_n) = P(\bigcap_{n=1}^{\infty} A_n)$.

Corollary 2.2.1 (Union bound)**.** *We have for the same sets $\{A_i\}$,*

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

2.1 Random Variables

Definition 2.3. A measurable space is a pair $(S, \mathcal{G})$ such that S is a set, and $\mathcal{G}$ is a σ-algebra on S.

Definition 2.4. If $(S_1, \mathcal{G}_1)$ and $(S_2, \mathcal{G}_2)$ are measurable spaces, a function $f : S_1 \to S_2$ is called measureable if $f^{-1}(B) \in \mathcal{G}_1$, for all $B \in \mathcal{G}_2$. This says inverse images of measurable sets are measurable under f.

We will use the notation $(f \in B) = f^{-1}(B) = \{s_1 \in S_1 \mid f(s_1) \in B\}$.

Definition 2.5. An S-valued random variable is a measureable function $X : \Omega \to S$ where Ω is a probability space, and S is a measureable space.

Definition 2.6. If $A \subset \Omega$, the indicator of A is

$$1_A(\omega) = 1(\omega \in A) = \begin{cases} 1 & \omega \in A \\ 0 & \omega \notin A. \end{cases}$$

Theorem 2.7. *If $A \subset \Omega$, then 1_A is a random variable iff $A \in \mathcal{F}$.*

Proof. Homework. □

If $f : S_1 \to S_2$, and $g : S_2 \to S_3$ are measurable, then so is $g \circ f : S_1 \to S_3$. If $f_1, f_2, \ldots, f_n : \Omega \to \mathbb{R}$, the following are equivalent:

1. f_i is a random variable, for all i.
2. The function $f = (f_1, \ldots, f_n)$ is an $\mathbb{R}^n$ valued random variable.

3. If $X, Y : \Omega \to \mathbb{R}^n$ are random variables and $\alpha, \beta \in \mathbb{R}$, then $\alpha X + \beta Y$ is a random variable. (The same occurs when $\mathbb{R}^n, \mathbb{R}$ are replaced by $\mathbb{C}^n, \mathbb{C}$.)

Proposition 2.8. *Consider $\{X_n\}$, where each $X_n : \Omega \to [-\infty, \infty]$ is a random variable. Then*

1. $\inf_n X_n$ *and* $\sup_n X_n$ *are random variables,*
2. $\liminf_{n\to\infty} X_n$ *and* $\limsup_{n\to\infty} X_n$ *are random variables.*
3. *If* $\lim_{n\to\infty} X_n(\omega)$ *exists for all* $\omega \in \Omega$*, then* $X(\omega) = \lim_{n\to\infty} X_n(\omega)$*, for some* $X : \Omega \to [-\infty, \infty]$.

2.2 Distributions of Random Variables

Definition 2.9. Let $X : \Omega \to S$ be a random variable. The distribution of X is the probability measure $\mu_X : \mathcal{G} \to [0, 1]$ defined by

$$\mu_X(A) = P(X \in A) = P(X^{-1}(A)).$$

One can verify that μ_X is indeed a probability measure.

Definition 2.10. If $X : \Omega_1 \to S$, and $Y : \Omega_2 \to S$ are random variables, we say X and Y have the same distribution, and write $X = dY$ if $\mu_X = \mu_Y$.

Definition 2.11. A random variable $X : \Omega \to S$ is called discrete if $\{s\} \subset \mathcal{G}$, for all $s \in S$, and there is a countable set S' such that $P(X \in S') = 1$.

Definition 2.12. If X is discrete, probability mass function of X is $p_X(s) = P(X = s)$. Note that if X is discrete, the probability

$$P(X \in A) = \mu_X(A) = \sum_{s \in A \cap S'} p_X(s).$$

Definition 2.13. A random variable $X : \Omega \to I$, $I \subset \mathbb{R}$ is continuous if there exists a measurable function $f_X : I \to [0, \infty)$ such that

$$P(a \le X \le b) = \mu_X([a, b]) = \int_a^b f_X(u)\, du.$$

This is integration in the Lebesgue sense, when needed.

Functions of continuous random variables may not be continuous.

3 Conditional Probability

Let $(\Omega, \mathcal{F}, P)$ be a probability space.

Theorem 3.1. *Let $X : \Omega \to \Omega$ to be the identity mapping. Then X is a random variable, and $\mu_X = P$. We call X the canonical random variable with distribution P.*

Example 3.2. I want an example of a random variable that is uniformly distributed on [0,2]. We can take $\Omega = [0, 2]$, and $\mathcal{F} = \mathcal{B}$, and $P(\cdot) = \frac{1}{2} m(\cdot)|_{\mathcal{B}}$, where m is the Lebesgue measure. Then our uniform RV is the canonical RV.

3.1 Multiple Random Variables

Consider $X : \Omega \to S_1, Y : \Omega \to S_2$. There is a natural σ-algebra on $S_1 \times S_2$ containing measurable rectangles $A \times B, A \in \mathcal{G}_1, B \in \mathcal{G}_2$, such that $(X, Y) : \Omega \to X_1 \times S_2$ is a random variable. More generally, if $\{X_i\}_{i \in I}$ (possibly uncountable but probably not more than the cardinality of $\mathbb{R}$) is a collection of random variables, with $X_i : \Omega \to S_i$. Then we can consider $(X_i)_{i \in I}$ a single random variable taking its values in $\prod_{i \in I} S_i$. Then we can talk about the distribution of the collection in $\prod_{i \in I} S_i$. Let's revisit the coin flipping example from the beginning of the course. We were given a sequence of random variables $X_1, X_2, \ldots, X_{50}$, where X_i was the location after i flips. By flipping coins, we as a class each generated a sample from the space $\prod_{i=1}^{50} X_i$. This viewpoint is important for us to adopt, because it allows us to ask questions about properties of the graphs of random functions.

3.2 Conditional Probability, Independence

Recall if $P(B) > 0$, $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$. We say A, B are independent if $P(A \cap B) = P(A)P(B)$. A collection $\{A_i\}_{i \in I}$ of events is an independent collection if for all $J \subset I$, $|J| < \infty$,

$$P\left(\bigcap_{j \in J} A_j\right) = \prod_{j \in J} P(A_j).$$

If $\{X_i\}_{i \in I}$ is a collection of RVs with $X_i : \Omega \to S_i$ is called independent if for every finite set $J \subset I$, and $A_j \in \mathcal{G}_j$,

$$P\left(\bigcap_{j \in J} (X_j \in A_j)\right) = \prod_{j \in J} P(X_j \in A_j)$$

Our previous definitions of expected value were as follows. If $X : \Omega \to \mathbb{R}$ is discrete, then $\mathbf{E}[X] = \sum_{x \in \mathbb{R}} xP(X = x)$, when this sum converges. If X is continuous with density f, we said

$$\mathbf{E}[X] = \int_{\mathbb{R}} xf(x)\,dx.$$

Why won't this be enough for our purposes?!! Let $X : \Omega \to [0, 1]$ be uniformly distributed. Let $g(x) = \max\{x, 1/2\}$. What is $\mathbf{E}[g(X)]$? LOTUS tells us $\mathbf{E}[g(X)] = \int_{\mathbb{R}} g(x)f(x)\,dx$, but this is a theorem. What is the definition of $\mathbf{E}[g(X)]$? We only have definitions for $\mathbf{E}[g(X)]$ if $g(X)$ is continuous or discrete. By definition, $g(X)$ is not discrete, since it can take uncountably many values. But note that $P(g(X) = 1/2) = 1/2 = P(X \leq 1/2)$. But if $g(X)$ were continuous, $P(g(X) = a) = 0$, for all a. So $g(X)$ is not continuous. So we are missing a definition for expectation which does not care whether a random variable is discrete or continuous.

4 Integration

4.1 Expected Value

Let $X_1, X_2, \ldots$ be independent and identically distributed (i.i.d.), with $P(X_i = 1) = 1 - P(X_i = 0) = p$. We can show that

$$\frac{1}{n} \sum_{i=1}^{n} X_i \to p$$

almost surely, or in probability. One can show this without using notions of expected value. Therefore, we define $\mathbf{E}[X_i] = p$. If $Y : \Omega \to \mathbb{R}$ takes only finitely many values, that is, there exists a finite $S \subset \mathbb{R}$ such that $P(Y \in S) = 1$, and $Y_1, Y_2, \ldots$ are i.i.d. with the same distribution as Y, then

$$\frac{1}{n}\sum_{i=1}^{n} Y_i \xrightarrow{p} \sum_{s\in S} sP(Y = s).$$

So we define $\mathbf{E}[Y] = \sum_{s\in S} sP(Y = s) = \sum_{s\in S} \mu_Y(s \in S)$. Note that if $X = Y$ in distribution, then $\mathbf{E}[X] = \mathbf{E}[Y]$.

Example 4.1. We know that $\mathbf{E}[X + Y] = \mathbf{E}[X] + \mathbf{E}[Y]$. From the definition we've given, this linearity is readily apparent. But from the definitions typically given in a first course of probability, this isn't so obvious, requiring a calculation involving change of summation order to unravel.

The following remedies this problem.

Theorem 4.2. *If $\Omega = \{\omega_1, \ldots, \omega_n\}$, and $X : \Omega \to \mathbb{R}$ is a RV, then*

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

Note that in this equality, we sum over the domain Ω of $\mathbb{R}$.

Proof. Let $S = X(\Omega)$. Since Ω is finite, so is S, and $P(X \in S) = 1$. Let $S = \{s_1, \ldots, s_n\}$. Note that $(X = s_1), (X = s_2), \ldots, (X = s_m)$ partition Ω. So

$$\begin{aligned}
\mathbf{E}[X] = \sum_{i=1}^{m} s_i P(X = s_i) &= \sum_{i=1}^{m} s_i \sum_{\omega : X(\omega) = s_i} P(\{\omega\}) \\
&= \sum_{i=1}^{m} \sum_{\omega : X(\omega) = s_i} X(\omega)P(\{\omega\}) \\
&= \sum_{\omega\in\Omega} X(\omega)P(\{\omega\}).
\end{aligned}$$

□

Corollary 4.2.1. *If $|\Omega| < \infty$, and $X, Y : \Omega \to \mathbb{R}$ are random variables, then*

$$\mathbf{E}[X + Y] = \mathbf{E}[X] + \mathbf{E}[Y].$$

Also, $|\mathbf{E}[X]| \le \mathbf{E}[|X|]$.

Proof. Consider

$$\begin{aligned}
\mathbf{E}[X + Y] &= \sum_{\omega\in\Omega} (X(\omega) + Y(\omega))P(\{\omega\}) \\
&= \sum_{\omega\in\Omega} X(\omega)P(\{\omega\}) + \sum_{\omega\in\Omega} Y(\omega)P(\{\omega\}).
\end{aligned}$$

Also,

$$|\mathbf{E}[X]| = \left|\sum_{\omega\in\Omega} X(\omega)P(\{\omega\})\right| \le \sum_{\omega\in\Omega} |X(\omega)|P(\{\omega\}) = \mathbf{E}[|X|].$$

□

Theorem 4.3. *Suppose $X, Y : \Omega \to \mathbb{R}$ are RVs taking only finitely many values, and $\alpha, \beta \in \mathbb{R}$. Then*

1. $\mathbf{E}[\alpha X + \beta Y] = \alpha \mathbf{E}[X] + \beta \mathbf{E}[Y]$,
2. $|\mathbf{E}[X]| \leq \mathbf{E}[|X|]$.
3. *If $X \leq Y$, then* $\mathbf{E}[X] \leq \mathbf{E}[Y]$.
4. *If X, Y are independent, then* $\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y]$.

Since this is a stochastic processes course, we will only sketch this proof.

Proof. Since X, Y take only finitely many values, let S be a finite set such that $P(X \in S) = P(Y \in S) = 1$. Let $\Omega' = S \times S$, and let $S' = 2^{\Omega'}$. Define

$$P'(\{(s_1, s_2)\}) = P(X = s_1, Y = s_2).$$

Define $X', Y' : \Omega' \to \mathbb{R}$ by

$$X'(s_1, s_2) = s_1, \quad Y'(s_1, s_2) = s_2.$$

In particular, (X', Y') is the identity on Ω'. Then $(X, Y) = (X', Y')$ in distribution, by construction. Since $(X, Y) = (X', Y')$, $\alpha X' + \beta Y' = \alpha X + \beta Y$ in distribution, and if X, Y are independent, then so are X', Y'. To prove the results, use the fact that expected value only depends on distribution. For example,

$$\begin{aligned}\mathbf{E}[\alpha X + \beta Y] &= \mathbf{E}\left[\alpha X' + \beta Y'\right] \\ &= \alpha \mathbf{E}\left[X'\right] + \beta \mathbf{E}\left[Y'\right] \\ &= \alpha \mathbf{E}[X] + \beta \mathbf{E}[Y].\end{aligned}$$

□

Definition 4.4. A random variable taking only finitely many values is called a simple random variable.

Definition 4.5. If $X : \Omega \to [0, \infty]$. Then,

$$\mathbf{E}[X] = \sup\{\mathbf{E}[Y] \mid Y \text{ is simple}, 0 \leq Y \leq X\}.$$

If $X : \Omega \to [-\infty, \infty]$, let $X^+ = \max\{X, 0\}, X^- = \max\{-X, 0\}$. Then $X = X^+ - X^-$, and $|X| = X^+ + X^-$, and $X^+, X^- \geq 0$.

Definition 4.6. If $X : \Omega \to \mathbb{R}$ is a RV, then we say X has finite expected value if $\mathbf{E}[|X|] < \infty$, and define

$$\mathbf{E}[X] = \mathbf{E}\left[X^+\right] - \mathbf{E}\left[X^-\right].$$

In places where the 630 definition of $\mathbf{E}[X]$ work, these definitions both agree. All properties of for simple random variables extend to general random variables. If we want to emphasize the role of P when writing $\mathbf{E}[X]$, we will use

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

Usually, in 630, to define $\mathbf{E}[X \mid Y = y]$, $P(X \mid Y = y)$ is defined first. In full generality, this definition cannot work.

5 Spaces of Random Variables

5.1 Metric Spaces, Inner Product Spaces of RVs

In 630, we defined for X, Y discrete RVs,

$$\mathbf{E}[X \mid Y = y] = \sum_x xP(X = x \mid Y = y).$$

Our problem is that this conditional distribution does not always exist. If X is discrete, and Y is continuous, what is $\mathbf{E}[X \mid Y = y]$.

Exercise 5.1. What is the 630 style formula for $\mathbf{E}[X \mid Y = y]$?

The basic question of conditioning is: how do you update your beliefs, or your guesses about the future when you learn new information? We know $\mathbf{E}[X \mid Y]$ should be a function of Y, the new information we learn. We want the best guess for X among all such guesses relying on Y. Thus, $h(Y)$ should be closer to X than any other $f(Y)$. Well now we should ask how should we be able to tell distances between random variables? This distance should be one that we already have a lot of good intuition about.

Example 5.1. Let's think of a simple metric space: $(\mathbb{R}^n, \sqrt{\langle \cdot, \cdot \rangle})$. The nice thing about this metric space is that its a separable Hilbert space that we can do some intuitive geometric reasoning about.

Definition 5.2. Let $X, Y : \Omega \to \mathbb{R}$ be RVs such that $\mathbf{E}\left[X^2\right] < \infty, \mathbf{E}\left[Y^2\right] < \infty$. Define

$$\langle X, Y \rangle_2 = \mathbf{E}[XY]$$

.

Example 5.3. If $\Omega = \{1, \ldots, n\}$, and $P(\{i\}) = p_i$, then

$$\langle X, Y \rangle_2 = \mathbf{E}[XY] = \sum_{i=1}^{n} X(i)Y(i)p_i.$$

Proposition 5.4. *Suppose $\alpha, \beta \in \mathbb{R}$, $X, Y, Z : \Omega \to \mathbb{R}$, and $\mathbf{E}\left[X^2\right], \mathbf{E}\left[Y^2\right], \mathbf{E}\left[Z^2\right] < \infty$. Then*

1. $\langle X, Y \rangle_2 = \langle Y, X \rangle_2$,
2. $\langle X, X \rangle_2 \geq 0$,
3. $\langle \alpha X + \beta Y, Z \rangle_2 = \alpha \langle X, Z \rangle_2 + \beta \langle Y, Z \rangle_2$.

We do not get definiteness from this definition: $\langle X, X \rangle_2 = 0$ does not imply $X = 0$.

Proposition 5.5. *If $X : \Omega \to \mathbb{R}$ has $\mathbf{E}\left[X^2\right] < \infty$, then $\langle X, X \rangle_2 = 0$ iff $X = 0$ almost surely, that is $P(X = 0) = 1$.*

Proof. Consider that if $\mathbf{E}\left[X^2\right] = 0$, then $X^2 \geq 0$. If $X^2 > 0$ on a set of positive probability, then $\mathbf{E}\left[X^2\right] > 0$. □

Definition 5.6. If $\mathbf{E}\left[X^2\right] < \infty$, and $\mathbf{E}\left[Y^2\right] < \infty$, the norm of X is $\|X\|_2 = \sqrt{\langle X, X \rangle_2}$. The distance from X to Y is $\|X - Y\|_2$.

Example 5.7. Let $\Omega = \{1, 2, \ldots, n\}$, $0 < p_i \leq 1$, with $P(\{i\}) = p_i$. Let $RV(\Omega, \mathbb{R})$ be the set of real-valued random variables on Ω. Define $\hat{}: RV(\Omega, \mathbb{R}) \to \mathbb{R}^n$ by $\hat{X} = (X(1), X(2), \ldots, X(n))$. Note that $X \mapsto \hat{X}$ is an isomorphism of vector spaces. We are thinking about $\mathbb{R}^n$ with the inner product

$$\langle x, y \rangle_p = \sum_{i=1}^{n} x_i y_i p_i.$$

Example 5.8. Let $n = 2$. Let $p_1 = 1 - p_2$. Plot the unit circle when $p_1 = 1/4, p_1 = 1/2, p_1 = 3/8$.

6 Conditional Expectation

6.1 Conditional Expectation

Observe that RVs if the form $f(Y)$, with $f : S \to \mathbb{R}$ form a subspace of $RV(\Omega, \mathbb{R})$. Once we show it exists, we will set $\mathbf{E}[X \mid Y]$ to be the orthogonal projection of X onto this subspace, i.e. the closest thing in this subspace to X. Note that there are many ways to define a metric on $RV(\Omega, \mathbb{R})$. For example, we could set $d(X, Y) = \mathbf{E}[|X - Y|]$, which would lead us to define what is called the conditional mean, rather than conditional expectation. I suspect the reason why we would prefer our inner product definition has to do with the fact that we get a semblance of a Hilbert space in $RV(\Omega, \mathbb{R})$ under the weighted inner product we have created. (Moreover, if we quotient $RV(\Omega, \mathbb{R})$ by its 0 expectation functions, like how we construct L^p spaces, I believe that we get a Hilbert space.)

Theorem 6.1. *Let $\Omega = \{\omega_1, \ldots, \omega_n\}$. Suppose*

$$0 < P(\{\omega_i\}) = p_i < 1.$$

Let $S = \{s_1, \ldots, s_m\}$ be a finite set. Suppose $X : \Omega \to \mathbb{R}$ and $Y : \Omega \to S$ are random variables. Then there is a unique random variable $Z : \Omega \to \mathbb{R}$ with the following properties.

1. *There is a function $h : S \to \mathbb{R}$ such that $Z = h(Y)$.*
2. *The distance*
$$d_2(X, Z) = \min_{f : s \to \mathbb{R}} d_2(X, f(Y)).$$
Note that Z is unique, and h is not necessarily unique.

We will present a proof using linear algebra, but we can also show this by solving a quadratic optimization problem. As a refresher, let $U \leq \mathbb{R}^n$. Suppose that $x \in \mathbb{R}^n$. How can we show that there exists a unique $z \in U$ such that $d(x, z) = \min_{y \in U} d(x, y)$. Let $u_1, \ldots, u_m$ be an orthogonal basis of U. Then

$$\mathrm{Proj}_U(x) = \sum_{i=1}^{m} \langle x_i, u_i \rangle \frac{u_i}{\|u_i\|^2} = \sum_{i=1}^{m} \left\langle x, \frac{u_i}{\|u_i\|} \right\rangle \frac{u_i}{\|u_i\|}.$$

If $y \in U$, then

$$d(x, y)^2 = d(x, \mathrm{Proj}_U(x))^2 + d(\mathrm{Proj}_U(x), y)^2.$$

To prove the theorem, we would choose a nice basis of the appropriate subspace, and do the same calculation.

7 Orthogonal Projections

Theorem 7.1. *The RVs $1_{\{\omega_1\}}, \ldots, 1_{\{\omega_n\}}$ are an orthogonal basis of $RV(\Omega, \mathbb{R})$.*

Proof. If $i \neq j$, then $1_{\{\omega_i\}} 1_{\{\omega_j\}} = 0$, so $\mathbf{E}\left[1_{\{\omega_i\}} 1_{\{\omega_j\}}\right] = 0$. If $W \in RV(\Omega, \mathbb{R})$, then

$$W = \sum_{i=1}^{n} W(\omega_i) 1_{\{\omega_i\}}.$$

□

Let $U = \{W \in RV(\Omega, \mathbb{R}) \mid W = f(Y), \text{ for some } f : S \to \mathbb{R}\}$.

Lemma 7.2. *The set U is a subspace of $RV(\Omega, \mathbb{R})$.*

Lemma 7.3. *Let $\overline{S} = Y(\Omega)$. Since S is finite, we know $\overline{S}$ is also finite. Let $\overline{S} = \{\overline{s_1}, \ldots, \overline{s_k}\}$. Then*

$$1_{\overline{s}_1} \circ Y, \ldots, 1_{\overline{s}_k} \circ Y$$

are an orthogonal basis for U.

Proof. Note that

$$(1_{\{\overline{s}_i\}} \circ Y)(\omega) = 1_{\{\overline{s}_i\}}(Y(\omega)) = \begin{cases} 1 & Y(\omega) = \overline{s}_i \\ 0 & \text{otherwise.} \end{cases}$$

Let $A_S = Y^{-1}(S)$. Then

$$1_{\overline{s}_i} \circ Y = 1_{A_{\overline{s}_i}}.$$

Then orthogonality follows as above because if $i \neq j$, then $A_{\overline{s}_i} \cap A_{\overline{s}_j} = \emptyset$. If $W = f(Y)$, then if $\overline{s} = Y(\omega)$, then

$$\begin{aligned} W(\omega) = f(Y(\omega)) &= f(\overline{s}) \\ &= f(\overline{s}) 1_{\{\overline{s}\}} Y(\omega) \\ &= f(\overline{s}) 1_{\{\overline{s}\}} Y(\omega) + \sum_{\overline{s}_i \neq \overline{S}} f(\overline{s}_i) 1_{\{\overline{s}_i\}} Y(\omega) \\ &= \sum_{\overline{s}_i \in \overline{S}} f(\overline{s}_i) 1_{\{\overline{s}_i\}} Y(\omega). \end{aligned}$$

□

Proof of Theorem 6.1. Define

$$Z = \sum_{i=1}^{k} \langle X, 1_{\{\overline{s}_i\}} \circ Y \rangle_2 \frac{1_{\{\overline{s}_i\}} \circ Y}{\| 1_{\{\overline{s}_i\}} \circ Y \|^2}.$$

Define $h : S \to \mathbb{R}$ by

$$h = \sum_{i=1}^{k} \langle X, 1_{\{\overline{s}_i\}} \circ Y \rangle \frac{1_{\{s_i\}}}{\| 1_{\{\overline{s}_i\}} \circ Y \|^2}.$$

Then $h(Y) = Z$. To verify $h(Y) = Z$ is the closest thing to X, we show that $X - Z$ is orthogonal to U. Suppose

$$W = \sum_{i=1}^{k} a_i 1_{\{\bar{s}_i\}} \circ Y \in U.$$

Consider that

$$\begin{aligned}
\langle Z, W \rangle &= \left\langle \sum_{i=1}^{k} \frac{\langle x, 1_{\{\bar{s}_i\}} \circ Y \rangle}{\|1_{\{\bar{s}_i\}} \circ Y\|^2} 1_{\{\bar{s}_i\}} \circ Y, \sum_{i=1}^{k} a_i 1_{\{\bar{s}_i\}} \circ Y \right\rangle \\
&\overset{(1)}{=} \sum_{i=1}^{k} a_i \frac{\langle x, 1_{\{\bar{s}_i\}} \rangle \circ Y}{\|1_{\{s_i\}} \circ Y\|^2} \|1_{\{\bar{s}_i\}} \circ Y\|^2 \\
&= \sum_{i=1}^{k} a_i \langle X, 1_{\{\bar{s}_i\}} \circ Y \rangle \\
&= \langle X, W \rangle.
\end{aligned}$$

where in (1), we've applied orthogonality and bilinearity. □

Note that $\mathbf{E}[ZW] = \mathbf{E}[XW]$, for all $W \in U$, and $Z \in U$, which characterize Z as the closest thing to Z in U.

Definition 7.4 (Conditional expectation)**.** Define $\mathbf{E}[X \mid Y] = Z$, where Z is as in Theorem 6.1.

7.1 Hilbert Spaces and Orthogonal Projections

Extending to a more general case, let

$$\mathcal{L}^2(\Omega) = \{X \in RV(\Omega, \mathbb{R}) \mid \mathbf{E}\left[X^2\right] < \infty\}.$$

Define an equivalence relation (same way we would in L^p) by $X \sim Y$ if $X = Y$ almost surely.

Definition 7.5. Define $L^2(\Omega)$ to be that set of equivalence classes under $\sim$.

Just like when talking about L^2 in the real case, we never talk about elements of $L^2(\Omega)$ like they are equivalence classes, rather, they will be treated as functions.

Definition 7.6. For $X, Y \in L^2(\Omega)$, $\langle X, Y \rangle = \mathbf{E}[XY]$, and $d, \|\cdot\|$ are defined similarly.

Theorem 7.7. *The claim is that $L^2(\Omega)$ is a Hilbert space.*

A notable remark we will note here is that in infinite dimensions, we can only project onto closed subspaces. Let's import the theorem.

Theorem 7.8. *Let U be a closed subspace of $L^2(\Omega)$. If $X \in L^2(\Omega)$, then there exists a unique $Z \in U$ such that $d(X, Z) = \inf_{\chi \in U} d(X, \chi)$, and $Z \in U$ has this property iff $\langle X, \chi \rangle = \langle Z, \chi \rangle$, for all $\chi \in U$.*

Define $Z = \text{Proj}_U(X)$.

Definition 7.9. Suppose $X : \Omega \to \mathbb{R}$ has $\mathbf{E}\left[X^2\right] < \infty$, and $Y : \Omega \to S$ is a RV. Let U_Y to be the closure of the subspace spanned by $h(Y)$, where $h : S \to \mathbb{R}$ is measurable and $\mathbf{E}\left[h(Y)^2\right] < \infty$. Then

$$\mathbf{E}[X \mid Y] = \text{Proj}_{U_Y}(X).$$

8 Sub-Sigma Algebras

Theorem 8.1. *Suppose $X : \Omega \to \mathbb{R}$ is a random variable such that $\mathbf{E}\left[X^2\right] < \infty$, and $Y : \Omega \to S$ is a random variable. If $Z \in U_Y$ then $Z = \mathbf{E}[X \mid Y]$ iff*

$$\mathbf{E}\left[X 1_{(Y \in A)}\right] = \mathbf{E}\left[Z 1_{(Y \in A)}\right],$$

for all $A \in \mathcal{G}$.

The claim is that $\operatorname{span}\{1_{(Y \in A)} \mid A \in \mathcal{G}\}$ is dense in Y. Let's replace U_Y because it is too big of a set to deal with right now.

Definition 8.2. Let $Y : \Omega \to S$ be a random variable. The σ-algebra generated by Y, denoted $\sigma(Y)$ is the smallest σ-algebra such that Y is $\sigma(Y)$-measurable.

We can write

$$\sigma(Y) = \{(Y \in A) \mid A \in \mathcal{G}\}.$$

We can think of $\sigma(Y)$ as the minimal amount of information we need to determine Y.

Definition 8.3. Suppose $X : \Omega \to \mathbb{R}$ is a RV such that $\mathbf{E}[X] < \infty$, or $X \geq 0$. Suppose $Y : \Omega \to S$ is a RV. We say Z is the conditional expected value of X given Y if Z is $\sigma(Y)$-measurable and

$$\mathbf{E}\left[X 1_{(Y \in A)}\right] = \mathbf{E}\left[Z 1_{(Y \in A)}\right],$$

for all $A \in \mathcal{G}$.

Definition 8.4. If $X : \Omega \to \mathbb{R}$ is a random variable, and $\mathcal{F}' \subset \mathcal{F}$, is a σ-algebra, the conditional expected value of X given $\mathcal{F}'$, $\mathbf{E}[X \mid \mathcal{F}']$ is a RV Z that is $\mathcal{F}'$ measurable and

$$\mathbf{E}[Z 1_A] = \mathbf{E}[X 1_A],$$

for all $A \in \mathcal{F}'$. We still assume that $\mathbf{E}[X] < \infty$, or $X \geq 0$.

Theorem 8.5. *In the context of the above definition, $\mathbf{E}[X \mid \mathcal{F}']$ is guaranteed to exist and is almost surely unique. If Z, and Z' are both $\mathcal{F}'$-measurable, and both have the property in the definition, then $Z = Z'$ almost surely.*

Theorem 8.6. *Suppose $X : \Omega \to \mathbb{R}$ is a random variable, with $\mathbf{E}[X] < \infty$ or $X \geq 0$ and $\mathcal{F}' \subset \mathcal{F}$ is a σ-algebra. If $W : \Omega \to \mathbb{R}$ is $\mathcal{F}'$-measurable and bounded, then $\mathbf{E}[XW] = \mathbf{E}[\mathbf{E}[X \mid \mathcal{F}']\, W]$.*

Definition 8.7. If $A \in \mathcal{F}$, and $\mathcal{F}' \subset \mathcal{F}$ is a σ-algebra, then the conditional probability of A given $\mathcal{F}'$ is defined by

$$P(A \mid \mathcal{F}') = \mathbf{E}\left[1_A \mid \mathcal{F}'\right].$$

Theorem 8.8 (Properties of conditional expected values)**.** *Suppose that $X : \Omega \to \mathbb{R}$ is such that $\mathbf{E}[X] < \infty$. Suppose that $Y : \Omega \to S$ is discrete. Therefore S is countable. Define $h : S \to \mathbb{R}$ by*

$$h(y) = \begin{cases} \frac{1}{P(Y=y)} \mathbf{E}\left[X 1_{(Y=y)}\right] & \text{if } P(Y = y) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Then $\mathbf{E}[X \mid Y] = h(Y)$.

9 Properties of Conditional Expectation

Proof of Theorem 8.8. We will show that $\mathbf{E}\left[X1_{(Y=y)}\right] = \mathbf{E}\left[h(Y)1_{(Y=y)}\right]$. Consider

$$\begin{aligned}\mathbf{E}\left[h(Y)1_{(Y=y)}\right] &= \mathbf{E}\left[h(y)1_{(Y=y)}\right] \\ &= h(y)\,\mathbf{E}\left[1_{(Y=y)}\right] \\ &= h(y)P(Y=y) \\ &= \frac{1}{P(Y=y)}\,\mathbf{E}\left[X1_{(Y=y)}\right]P(Y=y) \\ &= \mathbf{E}\left[X1_{(Y=y)}\right].\end{aligned}$$

□

Proposition 9.1. *For $\mathcal{F}' \subset \mathcal{F}$,*

$$\mathbf{E}\left[\mathbf{E}\left[X \mid \mathcal{F}'\right]\right] = \mathbf{E}\left[X\right].$$

Proof. In the definition take $A = \Omega$. Then

$$\mathbf{E}\left[X\right] = \mathbf{E}\left[X1_A\right] = \mathbf{E}\left[\mathbf{E}\left[X \mid \mathcal{F}'\right]1_A\right] = \mathbf{E}\left[\mathbf{E}\left[X \mid \mathcal{F}'\right]\right].$$

□

Proposition 9.2 (Linearity)**.** *For $\alpha, \beta \in \mathbb{R}$,*

$$\mathbf{E}\left[\alpha X + \beta Y \mid \mathcal{F}'\right] = \alpha\,\mathbf{E}\left[X \mid \mathcal{F}'\right] + \beta\,\mathbf{E}\left[Y \mid \mathcal{F}'\right].$$

Proof. Recall that for $X : \Omega \to \mathbb{R}$ a RV, and $\mathcal{F}' \subset \mathcal{F}$ a σ-algebra, $\mathbf{E}\left[X \mid \mathcal{F}'\right] := Z$, a RV that is $\mathcal{F}'$ measurable, and $\mathbf{E}\left[Z1_a\right] = \mathbf{E}\left[X1_A\right]$, for all $A \in \mathcal{F}'$. Let $A \in \mathcal{F}'$. Then $Z := \mathbf{E}\left[\alpha X + \beta Y \mid \mathcal{F}'\right]$, with

$$\begin{aligned}\mathbf{E}\left[Z1_A\right] &= \mathbf{E}\left[(\alpha X + \beta Y)1_A\right] \\ &= \mathbf{E}\left[\alpha X1_A + \beta Y1_A\right] \\ &= \alpha\,\mathbf{E}\left[X1_A\right] + \beta\,\mathbf{E}\left[Y1_A\right].\end{aligned}$$

□

Theorem 9.3 (Tower property)**.** *If $\mathcal{F}'' \subset \mathcal{F}' \subset \mathcal{F}$, then*

$$\mathbf{E}\left[\mathbf{E}\left[X \mid \mathcal{F}'\right] \mid \mathcal{F}''\right] = \mathbf{E}\left[X \mid \mathcal{F}''\right].$$

Proof. We WTS if $A \in \mathcal{F}''$, then

$$\mathbf{E}\left[\mathbf{E}\left[X \mid \mathcal{F}'\right]1_A\right] = \mathbf{E}\left[\mathbf{E}\left[X \mid \mathcal{F}''\right]1_A\right].$$

Since $A \in \mathcal{F}''$, $A \in \mathcal{F}'$, so

$$\mathbf{E}\left[\mathbf{E}\left[X \mid \mathcal{F}'\right]1_A\right] = \mathbf{E}\left[X1_A\right],$$

and

$$\mathbf{E}\left[\mathbf{E}\left[X \mid \mathcal{F}''\right]1_A\right] = \mathbf{E}\left[X1_A\right].$$

□

Theorem 9.4. *We have*

$$\mathbf{E}\left[Xh(Y) \mid Y\right] = h(Y)\,\mathbf{E}\left[X \mid Y\right].$$

More generally, if Z is $\mathcal{F}'$ measurable, then

$$\mathbf{E}\left[XZ \mid \mathcal{F}'\right] = Z\,\mathbf{E}\left[X \mid \mathcal{F}'\right].$$

Proof. If $A \in \mathcal{F}'$, $Z1_A$ is $\mathcal{F}'$ measurable, so by Theorem 8.6,

$$\mathbf{E}\left[XZ1_A\right] = \mathbf{E}\left[\mathbf{E}\left[X \mid \mathcal{F}'\right] Z1_A\right].$$

So by definition, $\mathbf{E}\left[XZ \mid \mathcal{F}'\right] = Z\,\mathbf{E}\left[X \mid \mathcal{F}'\right]$. □

Corollary 9.4.1. *If $\mathcal{F}'' \subset \mathcal{F}' \subset \mathcal{F}$, and $\mathbf{E}\left[X \mid \mathcal{F}'\right]$ is $\mathcal{F}''$ measurable, then $\mathbf{E}\left[X \mid \mathcal{F}'\right] = \mathbf{E}\left[X \mid \mathcal{F}''\right]$.*

Proof. Apply the previous theorem, and the tower property. □

There is one... and only one tool... which can help a mathematician compute conditional expectations.

Definition 9.5. Let $X : \Omega \to S$ be a random variable, and $\mathcal{F}' \subset \mathcal{F}$ a σ-algebra. We say X is independent of $\mathcal{F}'$ if X is independent of 1_A for every $A \in \mathcal{F}'$. Equivalently, $P(X \in B, A) = P(X \in B)P(A)$, for all $A \in \mathcal{F}'$.

Theorem 9.6. *If X is independent of $\mathcal{F}'$, then $\mathbf{E}\left[X \mid \mathcal{F}'\right] = \mathbf{E}\left[X\right]$.*

Proof. Homework next week. □

Theorem 9.7. *Suppose $X : \Omega \to S_1$ is a RV, $\mathcal{F}' \subset \mathcal{F}$ is a σ-algebra, and $Y : \Omega \to S_2$ is $\mathcal{F}'$ measurable. Suppose X is independent of $\mathcal{F}'$, and that $f : S_1 \times S_2 \to \mathbb{R}$ is a measurable function such that $\mathbf{E}\left[|f(X,Y)|\right] < \infty$. Define $\varphi : S_2 \to \mathbb{R}$ by*

$$\varphi(y) = \mathbf{E}\left[Xf(X,y)\right].$$

Then φ is measurable, and μ_Y-almost surely finite, and

$$\mathbf{E}\left[f(X,Y) \mid \mathcal{F}'\right] = \varphi(Y).$$

Example 9.8. Suppose that $\mathcal{N}, X_1, X_2, \ldots$ are independent, and X_i's are identically distributed, with $\mathbf{E}\left[X_i\right] < \infty$, and $P(N \in \mathcal{N}) = 1$, with $\mathbf{E}\left[\mathcal{N}\right] < \infty$. Then

$$\mathbf{E}\left[\sum_{i=1}^{\mathcal{N}} X_i \mid \mathcal{N}\right] = \mathcal{N}\,\mathbf{E}\left[X_i\right].$$

10 Stochastic Processes

Continuing Example 9.8,

$$\begin{aligned}
\mathbf{E}\left[\left|\sum_{i=1}^{N} X_i\right|\right] &\leq \mathbf{E}\left[\sum_{i=1}^{N}|X_i|\right] \\
&= \mathbf{E}\left[\sum_{n=1}^{\infty} 1_{(N=n)} \sum_{i=1}^{n}|X_i|\right] \\
&= \sum_{n=1}^{\infty} \mathbf{E}\left[1_{(N=n)} \sum_{i=1}^{n}|X_i|\right] && \text{(MCT)} \\
&= \sum_{n=1}^{\infty} P(N=n)\,\mathbf{E}\left[\sum_{i=1}^{n}|X_i|\right] \\
&= \sum_{n=1}^{\infty} P(N=n) n\,\mathbf{E}\left[|X_1|\right] \\
&= \mathbf{E}\left[|X_1|\right] \sum_{n=1}^{\infty} nP(N=n) \\
&= \mathbf{E}\left[|X_1|\right] \mathbf{E}\left[N\right] < \infty.
\end{aligned}$$

Define $f : N \times \mathbb{R}^N \to \mathbb{R}$ by

$$f(n, x_1, x_2, \ldots) = \sum_{i=1}^{n} x_i.$$

Assume that f is measurable. Note that

$$\sum_{i=1}^{N} X_i = f(N, X_1, X_2, \ldots).$$

Let

$$\varphi(n) = \mathbf{E}\left[f(n, X_1, \ldots,)\right] = \mathbf{E}\left[\sum_{i=1}^{n} X_i\right] = n\,\mathbf{E}\left[X_i\right].$$

Then Theorem 9.7 says that

$$\mathbf{E}\left[\sum_{i=1}^{N} X_i \mid N\right] = \mathbf{E}\left[f(N, X_1, X_2, \ldots) \mid N\right] = \varphi(N) = N\,\mathbf{E}\left[X_i\right].$$

Corollary 10.0.1. *We have*

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

Proof. We know $\mathbf{E}\left[\mathbf{E}\left[X \mid \mathcal{F}\right]\right] = \mathbf{E}\left[X\right]$. □

Example 10.1 (Galton-Watson branching process)**.** Galton was an English eugenecist, concerned with modeling the survival rate of certain surnames. Later, Watson expanded on the works of Galton, and Galton-Watson processes were used to model nuclear fission physics, epidemics, etc. Let ρ be a distribution on $\{0, 1, 2, 3, \ldots\}$ with finite mean. That is

$$\mu = \sum_i ip(\{i\}) < \infty.$$

Let $(\xi_{ij})_{i,j=1}^{\infty}$ be an array of independent random variables all with distribution ρ. Then ξ_{ij} should represent the number of offspring of individual j in the $(i-1)$th generation of the process.

Let $X_0 = \xi_{00}$ be the initial size of the population. Then

$$X_1 = \sum_{i=1}^{X_0} \xi_{1i}.$$

In general,

$$X_n = \sum_{i=1}^{X_{n-1}} \xi_{ni}.$$

The previous calculation shows that $\mathbf{E}[X_{n+1} \mid X_n] = \mathbf{E}[\xi_{n1}] X_n = \mu X_n$. Also,

$$\mathbf{E}[X_{n+1} \mid X_0, X_1, \ldots, X_n] = \mu X_n.$$

Iterating through, we find $\mathbf{E}[X_n] = \mu^n \mathbf{E}[\xi_{00}] = \mu^{n+1}$. If $\mu > 1$, then $\mathbf{E}[X_n] \xrightarrow{n\to\infty} \infty$. If $\mu = 1$, then $\mathbf{E}[X_n] = 1$. If $\mu < 1$, then $\mathbf{E}[X_n] \xrightarrow{n\to\infty} 0$. Later on, we will want to study the actual populations, rather than expected values, and we will use martingales to do this.

10.1 Markov Processes

Definition 10.2. A sequence $(X_n)_{n\geq 0}$ taking values in $(S, \mathcal{G})$ is called a Markov chain, or has the Markov property if for each $n \geq 0$, and $A \in \mathcal{G}$, the following is almost surely true:

$$P(X_{n+1} \in A \mid X_0, \ldots, X_n) = P(X_{n+1} \in A \mid X_n).$$

Theorem 10.3. *Suppose $f : S_1 \times S_2 \to S_1$ is measurable, and let X_0 be an S_1-valued random variable. Let $(U_n)_{n\geq 0}$ be a sequence of S_2-valued independent random variables independent of X_0, all identically distributed. For $n \geq 1$, define $X_n = f(X_{n-1}, U_n)$. Then $(X_n)_{n\geq 0}$ is a Markov chain.*

Example 10.4 (Random logistic map)**.** Consider

$$f(x) = \beta x(1-x), \quad 0 \leq \beta \leq 4.$$

Then f models resource limited growth. If $X_n = f(X_{n-1})$, we get an interesting dynamical system, depending on β. If we let $f(X, \beta) = \beta X(1-X)$, $\beta_1, \ldots, \beta_n$ be independent random variables with the same distribution, then $X_n = f(X_{n-1}, B_n)$ defines a Markov chain chose behavior depends on the distribution of β_i.

11 Examples of Markov Chains

Example 11.1. Let $X_1, X_2, \ldots,$ be independent $\mathbb{R}^n$-valued random variables, all with the same distribution. Let $S_n = X_1 + X_2 + \ldots + X_n$. Take $f(x, y) = x + y$. The theorem shows that this is a Markov chain.

Example 11.2. Let $X_1, X_2, \ldots,$ be i.i.d. RVs in $\mathbb{R}$. Let $P_n = \prod_{i=1}^n X_i$, and taking $f(x, y) = xy$ shows this is Markov.

Example 11.3 (Autoregressive models)**.** Autoregressive models are used by economists to forecast markets. In practice, economists add a moving average that is not Markov. Let $\epsilon_1, \epsilon_2, \ldots$ be i.i.d. normally distributed with parameters 0, σ^2, with $0 < \sigma^2 < \infty$. Fix $\varphi \in \mathbb{R}$, and define $X_n = \varphi X_{n-1} + \epsilon_n$.

Example 11.4 (Stochastic gradient descent)**.** To minimize $h : \mathbb{R}^n \to \mathbb{R}$, given $\gamma > 0$, start at X_0, and set $X_n = X_{n-1}\gamma\nabla h(X_{n-1})$. Under nice conditions, X_n converges to the minimum of h. In applications, gradients are hard to compute. In stochastic gradient descent, we let h_i be the directional derivative of h in the ith coordinate, and let $I_1, I_2, \ldots$ be i.i.d. uniform on $\{1, 2, \ldots, n\}$. We set

$$X_n = X_{n-1} - \gamma h_{I_n}(X_{n-1}).$$

This is a Markov chain.

Example 11.5. The Galton-Watson process is Markov. Consider

$$f(x, (y_1, y_2, \ldots)) = \sum_{i=1}^{x} y_i.$$

Proof of Theorem 10.3. Fix $A \in \mathcal{G}_1$. Then

$$\begin{aligned} P(X_{n+1} \in A \mid X_1, \ldots, X_n) &= \mathbf{E}\left[1(X_{n+1} \in A) \mid X_0, \ldots, X_n\right] \\ &= \mathbf{E}\left[1(f(X_n, U_{n+1}) \in A) \mid X_0, \ldots, X_n\right]. \end{aligned}$$

Let

$$\varphi(s) = \mathbf{E}\left[1(f(s, U_{n+1}) \in A)\right] = \mathbf{E}\left[1(f(s, U_1) \in A)\right] = P(f(s, U_1) \in A).$$

Theorem 9.7 says that

$$P(X_{n+1} \in A \mid X_1, \ldots, X_n) = \mathbf{E}\left[1(f(X_n, U_{n+1}) \in A) \mid X_0, \ldots, X_n\right] = \varphi(X_n).$$

By Corollary 10.0.1, since $P(X_{n+1} \in A \mid X_0, \ldots, X_n)$ is a function of X_n,

$$P(X_{n+1} \in A \mid X_n) = P(X_{n+1} \in A \mid X_0, \ldots, X_n).$$

□

12 Countable State Space Markov Chains

For most of this course, consider S to be a finite or countable. We can then take $\mathcal{G} = 2^S$. In this case, the condition that

$$P(X_{n+1} \in A \mid X_0, \ldots, X_n) = P(X_{n+1} \in A \mid X_n)$$

simplifies to the condition that for every $n \geq 0$, and $y, x_0, x_1, \ldots, x_n \in S$, such that

$$P(X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n) > 0,$$

we have that

$$P(X_{n+1} = y \mid X_0 = x_0, \ldots, X_n = x_n) = P(X_{n+1} = y \mid X_n = x_n).$$

Definition 12.1. A Markov chain $(X_n)_{n\geq 0}$ on S is called homogeneous or time homogeneous if there exists an $S \times S$ matrix $P = (p(i,j))_{i,j\in S}$ such that for every $n \geq 0$, and $y, x_0, \ldots, x_n \in S$ such that

$$P(X_0 = x_0, \ldots, X_n = x_n) > 0,$$

we have that

$$\begin{aligned} P(X_{n+1} = y \mid X_0 = x_0, \ldots, X_n = x_n) &= P(X_{n+1} = y \mid X_n = x_n) = p(X_n, y) \\ &= P(X_1 = y \mid X_0 = x_n), \end{aligned}$$

where the text in red holds if $P(X_0 = x_n) > 0$. A matrix P associated to a Markov chain in this way is called a transition matrix for $(X_n)_{n\geq 0}$. Some states in a Markov chain may be inaccessible, so this association is not unique.

Assume all the Markov chains we deal with in this course are time homogeneous.

Theorem 12.2. *Let $(X_n)_{n\geq 0}$ be a Markov chain with initial distribution μ and transition matrix P. Assume that each row of P sums to 1. Then for all $n \geq 0$, and $s_1, s_2, \ldots, s_n \in S$,*

$$P(X_0 = s_0, X_1 = s_1, \ldots, X_n = s_n) = \mu(\{s_0\}) \prod_{i=0}^{n-1} p(s_i, s_{i+1}).$$

Proof. Let $A_n = (X_0 = s_0, \ldots, X_n = s_n)$. Observe that

$$\begin{aligned} P(X_0 = s_0, \ldots, X_n = s_n) &= P(A_n) \\ &= P(A_{n-1})P(X_n = s_n \mid A_{n-1}) \\ &= P(A_{n-1})P(X_n = s_n \mid X_{n-1} = s_{n-1}) \\ &= P(A_{n-1})P(s_{n-1}, s_n). \end{aligned}$$

By continuing inductively on A_{n-1}, we get our result. □

Example 12.3. Consider a Markov chain on $S = \{1, 2, 3, 4\}$ with transition matrix

$$P = \begin{bmatrix} 0 & 1/2 & 1/4 & 1/4 \\ 1/4 & 1/2 & 0 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1 & 0 & 0 & 0 \end{bmatrix}.$$

Suppose $P(X_0 = 2) = 1$. From the theorem,

$$\begin{aligned} P(X_0 = 2, X_1 = 1, X_2 = 3, X_4 = 2) &= \mu(\{2\})p(2,1)p(1,3)p(3,2) \\ &= 1(1/4)^3. \end{aligned}$$

Now, as an exercise, $P(X_2 = 3)$ can be computed using the Markov property:

$$\begin{aligned} P(X_2 = 3) &= \sum_{i=1}^{4} P(X_2 = 3 \mid X_1 = i) P(X_1 = i) \\ &= \sum_{i=1}^{4} P(X_2 = 3 \mid X_1 = i) \sum_{j=1}^{4} P(X_1 = i \mid X_0 = j) P(X_0 = j). \end{aligned}$$

13 Construction of Markov Chains

Definition 13.1. An $S \times S$ matrix $P = (p_{(i,j)})_{i,j \in S}$ is called stochastic if the rows sum to 1, and the entries are nonnegative.

Theorem 13.2. *Given a stochastic matrix P and an initial distribution on S, there is a Markov chain with initial distribution μ and transition matrix P.*

The infinite sequence case is hard to prove it turns out, but the finite sequence case is easy to prove.

Definition 13.3. A Markov list is a finite sequence $(X_n)_{n=1}^{N}$ with the Markov property.

Theorem 13.4. *Given an initial distribution on S, and transitions matrix P, and $N \in \mathbb{N}$, there exists a Markov list $(X_n)_{n=0}^{N}$ with initial distribution μ, and transition matrix P.*

Proof. Let $\Omega = S^{\{0,1,2,\ldots,N\}}$. For $\omega = (s_0, \ldots, s_n)$, define $X_i(\omega) = s_i$, the projection of ω onto the ith coordinate. Define

$$P_\mu(\{(s_0, \ldots, s_N)\}) = \mu(\{s_0\}) \prod_{i=1}^{N-1} p(s_i, s_{i+1}).$$

Consider that

$$\begin{aligned} P_\mu(\Omega) &= \sum_{\omega \in \Omega} P_\mu(\{\omega\}) \\ &= \sum_{(s_0,\ldots,s_N) \in \Omega} P_\mu(\{s_0, \ldots, s_N\}) \\ &= \sum_{s_0 \in S} \cdots \sum_{s_N \in S} \mu(\{s_0\}) \prod_{i=0}^{N-1} p(s_i, s_{i+1}) \\ &= \sum_{s_0 \in S} \cdots \left[\sum_{s_{N-1} \in S} \left(\mu(\{s_0\}) \prod_{i=0}^{N-2} p(s_i, s_{i+1}) \right) \overbrace{\sum_{s_N \in S} p(s_{N-1}, s_N)}^{=1, \text{ since } P \text{ is stochastic}} \right] \\ &= \sum_{(s_0,\ldots,s_{N-1}) \in S^{\{0,1,\ldots,N-1\}}} \mu(\{s_0\}) \prod_{i=1}^{N-2} p(s_i, s_{i+1}). \end{aligned}$$

Thus $P_\mu(\Omega) = 1$, by induction. Lastly we need to show that $(X_i)_{i=0}^{N}$ satisfies the Markov property. By definition,

$$P_\mu(X_0 = s_0, \ldots, X_N = s_N) = P_\mu(\{(s_0, \ldots, s_N)\}) = \mu(\{s_0\}) \prod_{i=1}^{N-1} p(s_i, s_{i+1})$$

By the same argument as above,

$$P_\mu(X_0 = s_0, \ldots, X_{N-1} = s_{N-1}) = \mu(\{s_0\}) \prod_{i=1}^{N-2} p(s_i, s_{i+1}).$$

By induction, if $0 \leq n \leq N$,

$$P_\mu(X_0 = s_0, \ldots, X_n = s_n) = \mu(\{s_0\}) \prod_{i=1}^{n-1} p(s_i, s_{i+1}).$$

If $P_\mu(X_0 = s_0, \ldots, X_n = s_n) > 0$, then

$$\begin{aligned} P_\mu(X_{n+1} = y \mid X_0 = s_0, \ldots, X_n = s_n) &= \frac{P_\mu(X_{n+1} = y, X_0 = s_0, \ldots, X_n = s_n)}{P_\mu(X_0 = s_0, \ldots, X_n = s_n)} \\ &= \frac{\mu(\{s_0\}) \left(\prod_{i=1}^{n-1} p(s_i, s_{i+1})\right) p(s_n, y)}{\mu(\{s_0\}) \left(\prod_{i=1}^{n-1} p(s_i, s_{i+1})\right)} \\ &= p(s_n, y). \end{aligned}$$

□

14 Chapman-Kolmogorov Equation

We need more theory, we need more theory.

Theorem 14.1. *Let P be a stochastic matrix on S. Let μ be a distribution on S. Define $X_i : S^{\{0,1,2,\ldots\}} = S^{\mathbb{N}}$ to be the projection onto the ith coordinate $X_i(s_0, s_1, \ldots) = s_i$. Then there exists a unique measure P_μ on $S^{\mathbb{N}}$ such that $(X_n)_{n \geq 0}$ is a Markov chain with transition matrix P and initial distribution μ.*

Definition 14.2. Given a measurable space $(S, \mathcal{G})$, the set of all probability measures on S is denoted $P(S)$.

Definition 14.3. If $s \in S$, we define P_s by $P_s = P_{\delta_s}$ where $\delta_s \in P(S)$ gives measure 1 on $\{s\}$.

This P_s is the distribution of the Markov chain started at s.

Theorem 14.4. *If $\mu \in P(S)$, then*

$$P_\mu = \sum_{s \in S} \mu(\{s\}) P_s.$$

Proof. It is enough to check that equality holds on sets of the form $(X = s_0, \ldots, X_n = s_n)$. The theorem follows immediately from our theorem computing P_μ on sets of this form because if $s \neq s_0$, then

$$P_s(X_0 = s_0, \ldots, X_n = s_n) = 0,$$

since $\delta_s(\{s_n\}) = 0$. □

To think about starting the chain with distribution μ, first pick s_0 according to μ, then run the chain started from s_0. Define $p^{(m)}(s,t) = P_s(X_m = t)$ is the m-step transition matrix.

Theorem 14.5. *If $P_\mu(X_n = s) > 0$, then for $t \in S$, $P_\mu(X_{n+m} = t \mid X_n = s) = P_s(X_m = t) = p^{(m)}(s,t)$. More generally,*

$$P_\mu(X_{n+m} = t \mid X_0 = s_0, \ldots, X_n = s) = p^{(m)}(s,t)$$

Definition 14.6. Define $\theta : S^N \to S^N$ by

$$\theta(s_0, s_1, \ldots) = (s_1, s_2, \ldots),$$

the left shift operator on S^N

Theorem 14.7. *If $\mu \in P(S)$, then $P_\mu \circ \theta^{-1} = P_{\mu p}$.*

Here μP is left multiplication of P by μ. We think of μ as a row vector with entries $\mu = (\mu(s_1), \mu(s_2), \ldots)$, where s_1, s_2 is an enumeration of S. That is

$$(\mu P)(\{s^*\}) = \sum_{s \in S} \mu(\{s\}) P(s, s^*).$$

Proof. It suffices to prove equality on sets of the form $A = (X_0 = s_0, X_1 = s_1, \ldots, X_n = s_n)$. Consider

$$\begin{aligned} A &= (X_0 = s_0, X_1 = s_1, \ldots, X_n = s_n) \\ &= \{(\omega_0, \omega_1, \ldots) \in S^N : \omega_i = s_i, 0 \le i \le n\} \\ &= \{s_1\} \times \{s_1\} \times \ldots \times \{s_n\} \times S \times S \times \ldots. \end{aligned}$$

Consider then that

$$\theta^{-1}(A) = S \times \{s_1\} \times \{s_1\} \times \ldots \times \{s_n\} \times S \times S \times \ldots.$$

So

$$\begin{aligned} \theta^{-1}(A) &= \bigcup_{s \in S} \{s\} \times \{s_0\} \times \ldots \times \{s_n\} \times S \times S \times \ldots \\ &= \bigcup_{s \in S} (X_0 = s, X_1 = s_0, X_2 = s_1, \ldots, X_{n+1} = s_n). \end{aligned}$$

Use the computational theorem Theorem 12.2 to compute the probabilities on both sides:

$$\begin{aligned} P_\mu(\theta^{-1}(A)) &= \sum_{s \in S} P_\mu(X_0 = s, X_1 = s_0, \ldots, X_{n+1} = s_n) \\ &= \sum_{s \in S} \mu(\{s\}) p(s, s_0) \prod_{i=1}^{n-1} (s_i, s_{i+1}) \\ &= P_{\mu P}(X_0 = s_0, X_1 = s_1, \ldots, X_n = s_n) \\ &= P_{\mu P}(A). \end{aligned}$$

□

Note. The set $\theta^{-1}(A)$ is the set of all paths whose future is in A.

Corollary 14.7.1. *Inductively,* $P_\mu \circ (\theta^n)^{-1} = p_{\mu P^n}$.

Theorem 14.8. *For* $n \geq 0$,

$$P_\mu((\theta^n)^{-1}(A) \mid X_0 = s_0, \ldots, X_n = s_n) = P_{s_n}(A).$$

The proof is analogous to the proof we've just presented above. The above theorem captures that the entire future after time n is independent of the past given you are at time n.

Theorem 14.9. *If* $(X_n)_{n\geq 0}$ *is a Markov chain with transition matrix* P*, and initial distribution* μ*, then the distribution of* X_n *is* μP^n*, that is*

$$P_\mu(X_n = j) = \sum_{i \in S} \mu(\{i\}) p^{(n)}(i, j).$$

Definition 14.10. Let $\mu_i = \mu(\{i\})$.

Theorem 14.9. Consider that

$$P_\mu(X_1 = j) = \sum_{i \in S \mid \mu_i > 0} P(X_1 = j \mid X_0 = i) P(X_0 = i) = \sum_{i \in S} \mu_i p(i, j).$$

Since $P^{n+1} = P^n P$, we know

$$p^{(n+1)}(i, j) = \sum_{k \in S} p^{(n)}(i, k) p(k, j),$$

and the result follows by induction. □

Theorem 14.11 (Chapman-Kolmogorov equation)**.** *We have the following to consider:*

$$P_\mu(X_{n+m} = i) = \sum_{s \in S} P_s(X_m = i) P_m(X_n = s).$$

Proof. Consider $\mu P^{n+m} = \mu(P^n P^m) = (\mu P^n) P^m$. □

15 Stationary Distributions

Example 15.1. Let $(X_n)_{n\geq 0}$ be a Markov chain with state space $S = \{1, 2\}$ with transition matrix

$$P = \begin{bmatrix} 1/2 & 1/2 \\ 1/4 & 3/4 \end{bmatrix}.$$

Find $P_\mu(X_n = 1)$, and $P_\mu(X_n = 2)$. Compute $\lim_{n\to\infty} P_\mu(X_n = 1)$ and $\lim_{n\to\infty} P_\mu(X_n = 2)$.

Proof. Note that since the rows of P sum to 1, $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ is an eigenvector of P corresponding to eigenvalue 1. A diagonalization of P is is

$$P = QDQ^{-1} = \begin{bmatrix} 1 & -2 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1/4 \end{bmatrix} \begin{bmatrix} 1/3 & 2/3 \\ -1/3 & 1/3 \end{bmatrix}.$$

Hence

$$\begin{aligned}
\begin{bmatrix} P_\mu(X_n = 1) & P_\mu(X_n = 2) \end{bmatrix} &= \begin{bmatrix} \mu_1 & \mu_2 \end{bmatrix} \begin{bmatrix} 1 & -2 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1/4 \end{bmatrix} \begin{bmatrix} 1/3 & 2/3 \\ -1/3 & 1/3 \end{bmatrix} \\
\implies \lim_{n\to\infty} \begin{bmatrix} P_\mu(X_n = 1) & P_\mu(X_n = 2) \end{bmatrix} &= \begin{bmatrix} \mu_1 & \mu_2 \end{bmatrix} \begin{bmatrix} 1 & -2 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1/3 & 2/3 \\ -1/3 & 1/3 \end{bmatrix} \\
&= \begin{bmatrix} \mu_1 & \mu_2 \end{bmatrix} \begin{bmatrix} 1/3 & 2/3 \\ 1/3 & 2/3 \end{bmatrix} \\
&= \begin{bmatrix} \frac{\mu_1+\mu_2}{3} & \frac{2(\mu_1+\mu_2)}{3} \end{bmatrix} \\
&= \begin{bmatrix} 1/3 & 2/3 \end{bmatrix}.
\end{aligned}$$

Note that

$$\begin{bmatrix} 1/3 & 2/3 \end{bmatrix} \begin{bmatrix} 1/2 & 1/2 \\ 1/4 & 3/4 \end{bmatrix} = \begin{bmatrix} 1/3 & 2/3 \end{bmatrix}.$$

Interestingly the limiting distribution of P is the left eigenvector of P with eigenvalue 1, which is normalized to have nonnegative entries that sum to 1. □

Definition 15.2. If $(X_n)_{n\geq 0}$ is a Markov chain, a distribution π is called stationary for $(X_n)_{n\geq 0}$ if $\pi P = \pi$. Note that $P_\pi(X_n = i) = \pi_i$.

16 Random Walks on Graphs

Let $G = (V, E)$ be a connected finite graph. A uniform random walk on G is a Markov chain $(X_n)_n$ such that given $X_n = v$, X_{n+1} is uniformly distributed on the neighbors of v. That is

$$P(v, w) = \begin{cases} \frac{1}{\deg(v)} & v \sim w \\ 0 & \text{otherwise.} \end{cases}$$

17 Stochastic Matrices

These might be useful.

Theorem 17.1 (Perron-Frobenius theorem)**.** *If all entries of a $n \times n$ matrix A are positive, then it has a unique maximal eigenvalue. Its eigenvector has positive entries.*

Theorem 17.2 (Brouwer fixed point theorem)**.** *Let D^n be the closed ball in $\mathbb{R}^n$ and $\phi : D^n \to D^n$ be a continuous map, then ϕ has a fixed point*

Theorem 17.3. *Let S be a finite set, and $(X_n)_n$ be an S-valued Markov chain. Then $(X_n)_n$ has a stationary distribution.*

We will give 3 proofs. Let P be the transition matrix of $(X_n)_n$.

Proof 1. Assume that $S = \{1, 2, \ldots, n\}$. We want some vector $\pi = (\pi_1, \ldots, \pi_n)$, such that $\pi P = \pi$, and $\pi_i \geq 0$, for all i, and $\sum_{i=1}^n \pi_i = 1$. Since P is a stochastic matrix, this follows from the Perron-Frobenius theorem. □

Proof 2. Up to isomorphism, the set of probability measures on S,

$$P(S) = \left\{ (x_1, \ldots, x_n) \in [0,1]^n \mid \sum_{i=1}^{n} x_i = 1 \right\}.$$

Then $P(S)$ is compact and convex. Define $F : P(S) \to P(S)$ by $F(m) = mP$. Then F is continuous. By the Brouwer fixed point theorem, F has a fixed point, $\pi \in S$, that is, $\pi = F(\pi) = \pi P$. □

Proof 3. Let $P(S)$ be above, a compact and convex set. Fix $\mu \in P(S)$, and define for all $n \in \mathbb{Z}^+$,

$$\mu_n = \frac{1}{n} \sum_{i=1}^{n} \mu P^i.$$

Then $\mu_n \in P(S)$. Therefore, there exists a subsequence of $(\mu_n)_n$, $(\mu_{n_k})_k$ such that $\mu_{n_k} \to \pi$, for some $\pi \in P(S)$. Consider

$$\begin{aligned}
\pi P - \pi &= \lim_{k\to\infty} \mu_{n_k} P - \mu_{n_k} \\
&= \lim_{k\to\infty} \frac{1}{n_k} \sum_{i=1}^{m_k} \mu P^{i+1} - \frac{1}{n_k} \sum_{i=1}^{n_k} \mu P^i \\
&= \lim_{n\to\infty} \frac{1}{n_k} \mu (P^{n_k+1} - P) \\
&= 0,
\end{aligned}$$

since the entries μ, P^{n_k+1}, P are bounded. Therefore, $\pi P = \pi$. □

17.1 Strong Markov Property

To develop a stronger intuition behind stationary distributions, let us study the strong Markov property. In Markov chains, the future is independent of the past, given the present.

Definition 17.4 (Random time)**.** A random time is a RV taking values from $\overline{\mathbb{N}} = \mathbb{N} \cup \{\infty\}$.

Suppose $(X_n)_n$ is a stochastic process. Fix $s^* \in S$. If T is a random time, define

$$X_T(\omega) = X_{T(\omega)}(\omega) = \begin{cases} X_n & \text{if } T_n = n \\ s^* & \text{if } T = \infty \end{cases}.$$

Note that if $s \in S$,

$$(X_T = s) = \left(\bigcup_{n\in\mathbb{N}} (X_n = s, T = n) \right) \cup (s = s^*, T = \infty),$$

a countable union of measurable sets, so X_T is a RV. However, for the Markov property to hold at time T, we need T to not look into the future of the process.

Definition 17.5. Let $(X_n)_n$ be a stochastic process. A random time T is called a stopping time if for all $n \in \overline{\mathbb{N}}$, the event $(T = n)$ is in $\sigma(X_0, \ldots, X_n)$.

Equivalently, T is a stopping time if for each $n \in \overline{\mathbb{N}}$, there exists a function $h : S^{\{0,1,\ldots,n\}} \to \{0,1\}$ such that $(T = n) = (h_n(X_1, \ldots, X_n) = 1)$.

We will now present two versions of the strong Markov property.

Theorem 17.6 (Strong Markov property, version 1)**.** *Consider a Markov chain $(X_n)_n$ with transition matrix P. If T is a stopping time, and $S_0, \dots, S_n$ are such that*

$$P(X_0 = s_0, \dots, X_n = s_n, T = n) > 0,$$

then

$$P(X_{n+1} = y \mid X_0 = s_0, \dots, X_n = s_n, T = n) = p(s_n, y).$$

Proof. Since T is a stopping time, there is a function h_n such that

$$(T = n) = (h_n(X_0, \dots, X_n) = 1).$$

So

$$\begin{aligned}(X_0 = s_0, \dots, X_n = s_n, T = n) &= (X_0 = s_0, \dots, X_n = s_n, h_n(X_0, \dots, X_n) = 1) \\ &= (X_0 = s_0, \dots, X_n = s_n, h_n(s_0, \dots, s_n) = 1).\end{aligned}$$

If $h(s_0, \dots, s_n) = 1$, then

$$(X_0 = s_0, \dots, X_n = s_n, T = n) = (X_0 = s_0, \dots, X_n = s_n).$$

If $h(s_0, \dots, s_n) \neq 1$, then

$$(X_0 = s_0, \dots, X_n = s_n, T = n) = \emptyset.$$

Since $P(X_0 = s_0, \dots, X_n = s_n, T = n) > 0$, we are in the first case. By the Markov property,

$$\begin{aligned}P(X_{n+1} = y \mid X_0 = s_0, \dots, X_n = s_n, T = n) &= P(X_{n+1} \mid X_i = s_i, 0 \leq i \leq n) \\ &= p(s_n, y).\end{aligned}$$

□

For reasons we will discuss, this is a disappointing theorem. We would have liked to write

$$P(X_{T+1} = y \mid X_0 = s_0, \dots, X_T = s) = p(s, y).$$

But $(X_0, \dots, X_T)$ is a vector of random length, and $T = \infty$ is a possible value. We would need more tools to make this definition rigorous.

18

19

Theorem 19.1. *Let $(X_n)_{n \geq 0}$ be a Markov chain with transition matrix P. Let T be a stopping time. Then*

$$P(X_{T+1} = y \mid \mathcal{F}_T) = P(X_t, y), \quad (T < \infty).$$

Proof. Recall

$$P(X_{T+1} = y \mid \mathcal{F}_T) = \mathbf{E}\left[1(X_{T+1} = y) \mid \mathcal{F}_T\right].$$

Suppose $\mathcal{A} \in \mathcal{F}_T$. Then by the MCT,

$$\begin{aligned}
\mathbf{E}\left[1(X_{T+1})1_A\right] &= \sum_{n\geq 0} \mathbf{E}\left[1(X_{T+1} = y)1_{A\cap(T=n)}\right] + \mathbf{E}\left[1(s^* = y)1_{A\cap(T=\infty)}\right] \\
&= \sum_{n\geq 0} \mathbf{E}\left[1(X_{n+1} = y)1_{A\cap(T=n)}\right] + \underbrace{\mathbf{E}\left[1(s^* = y)1_{A\cap(T=\infty)}\right]}_{=\alpha} \\
&= \sum_{n\geq 0} \mathbf{E}\left[\mathbf{E}\left[1(X_{n+1} = y)\right] \mid \mathcal{F}_n 1_{A\cap(T=n)}\right] + \alpha \\
&= \sum_{n\geq 0} \mathbf{E}\left[P(X_{n+1} = y \mid \mathcal{F}_n)1_{A\cap(T=n)}\right] + \alpha \\
&= \sum_{n\geq 0} \mathbf{E}\left[p(X_n, y)1_{A\cap(T=n)}\right] + \alpha \\
&= \mathbf{E}\left[p(X_t, y)1_A 1(T < \infty)\right] + \mathbf{E}\left[1(s^* = y)1_A 1(T = \infty)\right] \\
&= \mathbf{E}\left[\underbrace{p(X_t, y)1(T < \infty) + 1(s^* = y)1(T = \infty)}_{\mathcal{F}_T \text{ measurable}} 1_A\right].
\end{aligned}$$

So $P(X_{T+1} \mid \mathcal{F}_T) = p(X_T, y)1(T < \infty) + 1(s^* = y)1(T = \infty)$. In particular, when $T < \infty$,

$$P(X_{T+y} = y \mid \mathcal{F}_T) = p(X_t, y).$$

□

We can give a stronger theorem using the shift map.

Theorem 19.2. *If T is a stopping time and $X = (X_n)_{n\geq 0}$ is a Markov chain, then for all $\mu \in P(S)$, and $A \in \mathcal{F}$,*

$$P_\mu(\Theta^T X \in A, T < \infty \mid \mathcal{F}_T) = P_{X_T}(A)1(T < \infty).$$

Implicitly, $x \mapsto P_X$ is measurable. The proof is the same as before, we would just use the Markov property in terms of shift maps.

Corollary 19.2.1. *If $F : S^N \to \mathbb{R}$ is nonnegative or bounded, and measurable, then*

$$\mathbf{E}_\mu[(F \circ \Theta^T 1(T < \infty) \mid \mathcal{F}_T)] = \mathbf{E}_\mu[(F)1(T < \infty)].$$

19.1 Excursion Theory, Classification of States

Definition 19.3. If $(X_n)_{n\geq 0}$ is a Markov chain, define

$$T_y = \inf\{n \geq 1 \mid X_n = y\}.$$

Theorem 19.4. *T_y is a stopping time.*

Proof. □