Review Sheet

EE 226A

Ahmed Shakil

Contents

1 Tools and Tricks

need to add bayes theorem

graph associated with a markov chain is a tree, then the markov chain is reversible

add tools for min max of exp rv

2 Elements of Probability Theory

2.1 Probability Spaces and Events

Definition 2.1: Kolmogorov's axioms

For any **probability space** $(\Omega, \mathcal{F}, P)$, the function P is called a **probability measure**. It is assumed to satisfy Kolmogorov's axioms:

i.) $P(A) \geqslant 0$ for all $A \in \mathcal{F}$;

ii.) $P(\Omega) = 1$;

iii.) if $A_1, A_2, \cdots \in \mathcal{F}$ are disjoint events, then $P(\cup_{i \geqslant 1} A_i) = \sum_{i \geqslant 1} P(A_i)$.

The probability space we are working in encodes the model of our experiment, with the **measurable space** $(\Omega, \mathcal{F})$ being the most fine-grained representation of outcomes we can hope to observe.

Theorem 2.2

For a probability space $(\Omega, \mathcal{F}, P)$, the probability measure P enjoys the following properties:

i.) Monotonicity: If $A \subset B$ are events, then $P(A) \leqslant P(B)$.

ii.) Subadditivity (Union bound): If $(A_i)_{i \geqslant 1}$ is a sequence of events in $\mathcal{F}$ and $A = \bigcup_{i \geqslant 1} A_i$, then $P(A) \leqslant \sum_{i \geqslant 1} P(A_i)$.

iii.) Continuity from below: If $A_1 \subset A_2 \subset \ldots$ are events in $\mathcal{F}$ and $A = \bigcup_{i \geqslant 1} A_i$, then $P(A_i) \to P(A)$.

iv.) Continuity from above: If $A_1 \supset A_2 \supset \ldots$ are events in $\mathcal{F}$ and $A = \bigcap_{i \geqslant 1} A_i$, then $P(A_i) \to P(A)$.

Theorem 2.3: Law of total probability

If events $A_1, A_2, \ldots$ partition Ω, then

$$P(B) = \sum_{i \geqslant 1} P(A_i \cap B), \quad B \in \mathcal{F}.$$

Definition 2.4: Infinitely often

$$\{A_n \text{ infinitely often}\} = \bigcap_{n \geqslant 1} \bigcup_{i \geqslant n} A_i.$$

We should understand $\{A_n \text{ i.o.}\}$ to be the set of samples $\omega \in \Omega$ such that $\omega \in A_i$ for infinitely many $i \geqslant 1$.

Lemma 2.5: Borel-Cantelli

Let $A_1, A_2, \ldots$ be a sequence of events. If

$$\sum_{i \geqslant 1} P(A_i) < \infty$$

then $P(\{A_i \text{ infinitely often}\}) = 0$.

Definition 2.6: Independent events

A collection of events $A_1, A_2, \ldots$ are **independent** if

$$P\left(\bigcap_{i \in S} A_i\right) = \prod_{i \in S} P(A_i)$$

for every finite subset $S \subset \{1, 2, 3, \ldots\}$. If $A_1, A_2, \ldots$ are independent, then $A_1^C, A_2, \ldots$ are independent. By induction, the complements $A_1^C, A_2^C, \ldots$ are also independent.

Lemma 2.7: Converse to Borel-Cantelli

Let $A_1, A_2, \ldots$ be independent events. If

$$\sum_{i \geqslant 1} P(A_i) = \infty,$$

then $P(\{A_i \text{ infinitely often}\}) = 1$.

Theorem 2.8: Carathéodory's extension theorem

Suppose $\mathcal{G}$ is a family of subsets of Ω that satisfies the following (relatively modest) properties:

i.) $\varnothing, \Omega \in \mathcal{G}$;

ii.) if $A, B \in \mathcal{G}$, then $A \cap B \in \mathcal{G}$;

iii.) if $A, B \in \mathcal{G}$, then there is a *finite* number of *disjoint* sets $C_1, \ldots, C_n \in \mathcal{G}$ such that $A \setminus B = \bigcup_{i=1}^n C_i$.

(Note: (*iii*) is weaker than imposing the assumption $A \in \mathcal{G} \implies A^C \in \mathcal{G}$.)

The extension theorem says that if we assign numbers (i.e., probabilities) $p(A)$ to the sets $A \in \mathcal{G}$ so that

A. $p(A) \geqslant 0$ for $A \in \mathcal{G}$;

B. $p(\Omega) = 1$;

C. if $B \in \mathcal{G}$ and $A_1, A_2, \cdots \in \mathcal{G}$ are disjoint with $B = \cup_{i \geqslant 1} A_i$, then $p(B) = \sum_{i \geqslant 1} p(A_i)$,

then there exists a unique probability measure P on $\sigma(\mathcal{G})$ that satisfies A-C and has the property that $P(A) = p(A)$ for all $A \in \mathcal{G}$.

2.2 Random Variables and Expectation

2.2.1 Random variables and algebraic properties

Definition 2.9: Random Variable

We define a random variable to be a function $X : \Omega \to \overline{\mathbb{R}}$ that satisfies

$$\{\omega \in \Omega : X(\omega) \leqslant \alpha\} \in \mathcal{F} \text{ for each } \alpha \in \overline{\mathbb{R}}.$$

Note that $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$.

A function $X : \Omega \to \overline{\mathbb{R}}$ satisfying the definition above is said to be $\mathcal{F}-$**measurable**. If X does not take values $\pm\infty$ (say, with probability one), then we say it is a **real-valued random variable**.

Proposition 2.10

If X is a random variable, then pX and $|X|^p$ are random variables for $p \in \mathbb{R}$. Moreover, if X, Y are real-valued random variables, then $X + Y$, and XY are also random variables.

Proposition 2.11

If $(X_n)_{n\geqslant 1}$ is a sequence of random variables defined on a common probability space $(\Omega, \mathcal{F}, P)$, then

- $\sup_{n\geqslant 1} X_n$ and $\inf_{n\geqslant 1} X_n$ are random variables; and
- $\limsup_{n\to\infty} X_n$ and $\liminf_{n\to\infty} X_n$ are random variables; and
- if $\lim_{n\to\infty} X_n$ exists point wise, it is also a random variable.

Definition 2.12: Almost sure equivalence of random variables

If X, Y are random variables and $P(\{\omega : X(\omega) \neq Y(\omega)\}) = 0$, then we say $X = Y$ almost surely (abbreviated a.s.), or $X = Y$ with probability one.

2.2.2 Distribution functions and distributions

Definition 2.13: Distribution function

A random variable X on a probability space $(\Omega, \mathcal{F}, P)$ is described in part by its **distribution function** $F_X : \mathbb{R} \to [0, 1]$, defined as

$$F_X(x) \coloneqq P\{X \leqslant x\}, \quad x \in \mathbb{R}.$$

Theorem 2.14: Properties of the distrbution function

A function $F : \mathbb{R} \to [0, 1]$ is the distribution function of a random variable if and only if

i.) F is nondecreasing

ii.) F is right-continuous, that is $\lim_{y \downarrow x} F(y) = F(x)$, for all $x \in \mathbb{R}$.

Moreover, F is the distribution function of a real-valued random variable if and only if it further holds that

$$\lim_{x \to -\infty} F(x) = 0 \text{ and } \lim_{x \to \infty} F(x) = 1.$$

Remark 2.15

Let X be a random variable with distribution function F_X. It follows by continuity from above and below, respectively, that

$$P\{X = -\infty\} = \lim_{x \to -\infty} F_X(x) \text{ and } P\{X = +\infty\} = 1 - \lim_{x \to +\infty} F_X(x).$$

These limits are always well-defined by monotonicity of F_X.

Definition 2.16: Law of a random variable

The function

$$L_X(B) \coloneqq P\{X \in B\}, \quad B \in \mathcal{B}_{\overline{\mathbb{R}}}$$

defines a probability measure on $\overline{\mathbb{R}}$ equipped with the Borel $\sigma-$algebra. This function is called the **law** of X and is synonymous with the distribution of X.

3 Discrete-Time Markov Chains

3.1 Definition of a Markov Chain

Definition 3.1: Markov chain

A Markov chain is a process $(X_n)_{n\geqslant 0}$ satisfying

$$\Pr\{X_{n+1}=j \mid X_n=i, X_{n-1}=i_{n-1},\dots,X_0=i_0\} = \Pr\{X_{n+1}=j \mid X_n=i\}$$

for all $n \geqslant 1$ and $j, i, i_{n-1}, i_0 \in \mathcal{S}$. A Markov chain is said to be **temporally homogeneous** if there are numbers $(p_{ij})_{i,j\in\mathcal{S}}$ such that

$$\Pr\{X_{n+1}=j \mid X_n=i\} = p_{ij}$$

for all $n \geqslant 0$ and all states $i, j \in \mathcal{S}$. The numbers $(p_{ij})_{i,j\in\mathcal{S}}$ are generically referred to as the **transition probabilities** of the Markov chain.

Definition 3.2: Transition matrix

$$P = \begin{bmatrix} p_{00} & p_{01} & \cdots \\ p_{10} & p_{11} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

The matrix P is called the transition matrix. It is a stochastic matrix, which means it is square with non-negative entries, whose rows sum to one. A Markov chain and transition matrix are equivalent representations of each other.

3.1.1 The Chapman-Kolmogorov equations

Proposition 3.3: The Chapman-Kolmogorov Equations

We define **multi-step transition probabilities**

$$P_{ij}^n \coloneqq \Pr\{X_{n+m}=j \mid X_m=i\}, \quad n, m \geqslant 0.$$

The Chapman-Kolmogorov equations give a recursive formula for computing the n-step transition probabilities.

For all $m, n \geqslant 0$ and states i, j,

$$P_{ij}^{m+n} = \sum_k P_{ik}^m P_{kj}^n.$$

In particular, we have

$$P^n = \begin{bmatrix} P_{00}^n & P_{01}^n & \cdots \\ P_{10}^n & P_{11}^n & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

where P^n is the transition matrix P raised to the n^{th} power. Also note $P^{m+n} = P^m P^n$.

Definition 3.4: Irreducible

A **class** of states is a nonempty set of states such that every state in the set can communicate with one another. These classes form an equivalence relation and partition the state space. We say a Markov chain is irreducible if there is only one class.

Definition 3.5: Periodicity

For a state i, define its **period**

$$d(i) \coloneqq \gcd\{n \geqslant 1 : P_{ii}^n > 0\}.$$

States with period 1 are called **aperiodic**. Periodicity is a class property.

Proposition 3.6

Define the **first return time** for state $j \in \mathcal{S}$ as

$$T_j \coloneqq \inf\{n \geqslant 1 : X_n = j\}.$$

Let $(X_n)_{n\geqslant 0}$ be a Markov chain with transition matrix P. Conditioned on the event $\{T_j < \infty\}$, the process $(X_{n+T_j})_{n\geqslant 0}$ is a Markov chain with transition matrix P and starting state j and is independent of $X_0, \ldots, X_{T_j}$.

Proposition 3.7

A state j is **recurrent** if $\Pr\{T_j < \infty \mid X_0 = j\} = 1$, and it is called **transient** if $\Pr\{T_j < \infty \mid X_0 = j\} < 1$. Recurrence and transience are class properties.

Lemma 3.8

State i is recurrent if and only if $\sum_{n=1}^{\infty} P_{ii}^n = \infty$.

Corollary 3.9

If $i \leftrightarrow j$ and j is recurrent, then $\Pr\{T_j < \infty | X_0 = i\} = 1$.

3.2 Markov Limit Theorems

Theorem 3.10: Strong Law of Large Numbers for Markov Chains

Define $N_j(n)$, $n \geqslant 1$, to be the number of transitions into state j, up to and including time n. More precisely,

$$N_j(n) \coloneqq \#\{1 \leqslant k \leqslant n : X_k = j\}.$$

Also define the expected first return time to be

$$\mu_{jj} \coloneqq \mathbb{E}[T_j | X_0 = j].$$

Let $(X_n)_{n\geqslant 0}$ be a Markov chain starting in state $X_0 = i$. If $i \leftrightarrow j$, then

$$\frac{N_j(n)}{n} \to \frac{1}{\mu_{jj}} \quad \text{a.s.}$$

Corollary 3.11

For an irreducible Markov chain $(X_n)_{n\geqslant 0}$, we have

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} P_{ij}^k = \frac{1}{\mu_{jj}}.$$

3.2.1 Stationary distributions and the issue of convergence

Definition 3.12: Stationary Distribution

A probability distribution $(\pi_j)_{j\in\mathcal{S}}$ is a stationary distribution for a Markov chain with transition probability matrix P if $\pi_j = \sum_i \pi_i p_{ij}$ for each $j \in \mathcal{S}$. Equivalently in matrix notation, $\pi = \pi P$, when π is considered as a row vector.

Definition 3.13: Positive and Null recurrence

A recurrent state j is positive recurrent if $\mu_{jj} < \infty$, or null recurrent if $\mu_{jj} = \infty$. Positive and null recurrence are class properties.

It is important to note stationary distributions aren't necessarily unique. It is important to note that stationary distribution does not always exist.

Theorem 3.14

An irreducible Markov chain satisfies exactly one of the following:

1. All states are transient, or all states are null recurrent. In this case, $\frac{1}{n}\sum_{k=1}^{n} P_{ij}^k \to 0$ as $n \to \infty$ for all states i, j, and no stationary distribution exists.
2. All states are positive recurrent. In this case, a unique stationary distribution exists and is given by $\pi_j = \frac{1}{\mu_{jj}} = \lim_{n\to\infty} \frac{1}{n} \sum_{k}^{n} P_{ij}^k$.

Theorem 3.15

Let $(X_n)_{n\geqslant 0}$ be an irreducible, aperiodic, and positive recurrent Markov chain with stationary distribution

π. Then

$$\lim_{n\to\infty} \sum_j \left|P_{ij}^n - \pi_j\right| = 0 \text{ for all } i \in \mathcal{S}.$$

In particular, $P_{ij}^n = \pi_j$ for all i, j.

3.3 Reversibility and Spectral Gap

Definition 3.16: Reversible Markov chain

A Markov chain with transition matrix P and stationary distribution π is reversible if the transition probabilities satisfy

$$\pi_i p_{ij} = \pi_j p_{ji} \text{ for all states } i, j.$$

In this case, we say (P, π) is reversible for convenience.

The equations above are called the detailed balance equations. These equations are also sufficient for reversibility. Hence, if there is a probability distribution π such that the equations are satisfied, then the Markov chain is reversible with stationary distribution π.

3.3.1 Spectral gap and trend to equilibrium

Definition 3.17: Total variation distance

For probability measures μ, ν on a measurable space $(\Omega.\mathcal{F})$, we define their total variation distance

$$\|\mu - \nu\|_{TV} \coloneqq \sup_{A\in\mathcal{F}} |\mu(A) - \nu(A)|.$$

Also note that when Ω is countable, we have

$$\|\mu - \nu\|_{TV} = \frac{1}{2} \sum_\omega |\mu(\omega) - \nu(\omega)|.$$

Definition 3.18: Spectral gap

Define $\mathrm{Var}_\pi(f) = \mathrm{Var}(f(X))$ for $X \sim \pi$ and $f : \mathcal{S} \to \mathbb{R}$. Also define the function $Pf : \mathcal{S} \to \mathbb{R}$ via the matrix-vector multiplication

$$(Pf)(i) = \sum_j p_{ij} f(j), \quad i \in \mathcal{S}.$$

For a reversible Markov Chain (P, π), define the spectral gap $\gamma \coloneqq 1 - \lambda_2$, where λ_2 is the smallest number satisfying

$$\mathrm{Var}_\pi(Pf) \leqslant \lambda_2 \mathrm{Var}_\pi(f) \text{ for all } f : \mathcal{S} \to \mathbb{R} \text{ with } \mathrm{Var}_\pi(f) < \infty.$$

Theorem 3.19

If (P, π) is a reversible Markov chain with spectral gap γ, then

$$\|P^n_{i\bullet} - \pi\|^2_{TV} \leqslant \frac{(1-\gamma)^n}{\pi_i} \text{ for each } n \geqslant 0 \text{ and each state } i.$$

Moreover, if the Markov chain is irreducible, aperiodic and has a finite state space, then $\gamma > 0$.

Definition 3.20

Let μ, π be probability distributions on state space $\mathcal{S}$. We say that μ has density h with respect to π (written $d\mu = h d\pi$) if

$$\mu_j = h(j)\pi_j \quad \forall j.$$

Lemma 3.21

Let (P, π) be a reversible Markov chain. If $d\mu = h d\pi$, then $P^n h$ is the density of μP^n with respect to π.

4 Martingales

4.1 Definitions and Examples

Definition 4.1

Let $(X_n)_{n\geqslant 0}$ be a stochastic process. A process $(M_n)_{n\geqslant 0}$ is said to be a **martingale with respect to** $(X_n)_{n\geqslant 0}$ if $(M_n)_{n\geqslant 0}$ is adapted to $(X_n)_{n\geqslant 0}$ and, for each $n \geqslant 0$,

(i) $\mathbb{E}|M_n| < \infty$;

(ii) $\mathbb{E}[M_{n+1} \mid X_0, \ldots, X_n] = M_n$.

If the equality in *(ii)* is replaced by $\geqslant$ or $\leqslant$, then the process is said to be a **submartingale** or **supermartingale**, respectively. The phrase "adapted to" means that M_n is a measurable function $(X_0, \ldots, X_n)$ for each $n \geqslant 0$.

Proposition 4.2

If $(M_n)_{n\geqslant 0}$ is a martingale with respect to $(X_n)_{n\geqslant 0}$, then for all $m > n$,

$$\mathbb{E}[M_m \mid X_0, \ldots, X_n] = M_n.$$

If $(M_n)_{n\geqslant 0}$ is a submartingale or supermartingale, then the equality above is $\geqslant$ or $\leqslant$, respectively.

4.2 Stopping Times

Definition 4.3

A nonnegative integer-valued random variable T is a **stopping time** with respect to $(X_n)_{n\geqslant 0}$ if, for each $n \geqslant 0$, the occurrence of the event $\{T \leqslant n\}$ is determined entirely by $(X_0, \ldots, X_n)$. In other words, the indicator $1_{\{T\leqslant n\}}$ is a measurable function of $(X_0, \ldots, X_n)$.

Definition 4.4: Stopped process

Let $(X_n)_{n\geqslant 0}$ be a process, and T be a stopping time. If $(Y_n)_{n\geqslant 0}$ is adapted to $(X_n)_{n\geqslant 0}$, then the process $(Y_{T\wedge n})_{n\geqslant 0}$ is called the **stopped process**. Note that the stopped process satisfies $Y_{T\wedge n} = Y_n$ for $n \leqslant T$, and $Y_{T\wedge n} = Y_T$ for $n > T$.

4.2.1 Stopping times and martingales

Proposition 4.5

If $(M_n)_{n\geqslant 0}$ is a martingale and T is a stopping time, both with respect to $(X_n)_{n\geqslant 0}$, then the stopped process $(M_{T\wedge n})_{n\geqslant 0}$ is also a martingale with respect to $(X_n)_{n\geqslant 0}$.

Proposition 4.6

If $(M_n)_{n\geqslant 0}$ is a submartingale and T is a stopping time, both with respect to $(X_n)_{n\geqslant 0}$, then

$$\mathbb{E}[M_0] \leqslant \mathbb{E}[M_{T\wedge n}] \leqslant \mathbb{E}[M_n] \quad n \geqslant 0.$$

Proposition 4.7: Optimal Stopping Theorem

Let $(M_n)_{n\geqslant 0}$ be a submartingale and T be a stopping time, both with respect to $(X_n)_{n\geqslant 0}$. If there is a constant $k < \infty$ such that any one of the following hold

i.) $T \leqslant k$ a.s.; or

ii.) $|M_n| \leqslant k$ a.s. for each n, and $\Pr\{T < \infty\} = 1$; or

iii.) $\mathbb{E}[T] < \infty$ and $|M_n - M_{n-1}| \leqslant k$ a.s. for each $n \geqslant 1$,

then

$$\mathbb{E}[M_0] \leqslant \mathbb{E}[M_T].$$

The inequality above is an equality when $(M_n)_{n\geqslant 0}$ is a martingale.

Theorem 4.8: Wald's Identity

Let $(Y_n)_{n\geqslant 1}$ be adapted to $(X_n)_{n\geqslant 1}$. Assume Y_{n+1} is independent of $(X_1,\dots,X_n)$ for each $n \geqslant 1$, $\sup_{n\geqslant 1}\mathbb{E}|Y_n| < \infty$, and $\mathbb{E}[Y_n] = \mu$ for all $n \geqslant 1$. If $T \geqslant 1$ is a stopping time with respect to $(X_n)_{n\geqslant 1}$ satisfying $\mathbb{E}[T] < \infty$, then

$$\mathbb{E}\left[\sum_{n=1}^{T} Y_n\right] = \mu\mathbb{E}[T].$$

5 Poisson Processes

5.1 The Exponential Distribution

Definition 5.1: Exponential distribution

The **exponential distribution with rate** $\lambda > 0$, denote $\text{Exp}(\lambda)$, has density

$$f(t) = \begin{cases} \lambda e^{-\lambda t} & t \geqslant 0 \\ 0 & t < 0. \end{cases}$$

For $T \sim \text{Exp}(\lambda)$, the distribution function is given by

$$\Pr\{T \leqslant t\} = \begin{cases} 1 - e^{-\lambda t} & t \geqslant 0 \\ 0 & t < 0. \end{cases}$$

We note that $\mathbb{E}[T] = 1/\lambda$ and $\text{Var}(T) = 1/\lambda^2$. An important property of exponential random variables is the **memoryless property**. In particular, if $T \sim \text{Exp}(\lambda)$, then

$$\Pr\{T > t + s | T > t\} = \frac{\Pr\{T > t + s\}}{\Pr\{T > t\}} = \frac{e^{-\lambda(t+s)}}{e^{-\lambda t}} = e^{-\lambda s} = \Pr\{T > s\}.$$

Definition 5.2: Erlang distribution

If, $T_1, \ldots, T_k$ are i.i.d. exponential random variables with rate λ, then their sum $T = T_1 + \cdots + T_k$ has an **Erlang** distribution, with density

$$f_T(t) = \begin{cases} \lambda e^{-\lambda t} \frac{(\lambda t)^{k-1}}{(k-1)!} & t \geqslant 0 \\ 0 & t < 0. \end{cases}$$

5.2 Poisson Processes

Definition 5.3: Poisson Process

Let $\tau_1, \tau_2, \ldots$ be i.i.d. exponential random variables with rate $\lambda > 0$ and, for $n \geqslant 1$, define $T_n = \tau_1 + \tau_2 + \cdots + \tau_n$, with the convention that $T_0 = 0$. For each $t \geqslant 0$, define the random variable $N_t = \sup\{n \geqslant 0 : T_n \leqslant t\}$. The process $(N_t)_{t \geqslant 0}$ is called a **Poisson process** with rate λ.

This is best thought of as an example of a counting process. A **counting process** is a random process $(N_t)_{t \geqslant 0}$, such that (i) N_t is a non-negative integer for each time $t \geqslant 0$; (ii) the sample paths $t \mapsto N_t(\omega)$ are non-decreasing in t; and (iii) the sample paths $t \mapsto N_t(\omega)$ are right-continuous.

Definition 5.4: Poisson distribution

A random variable X is said to be Poisson distributed with mean $\lambda \geqslant 0$ ($X \sim \text{Poisson}(\lambda)$) if X has probability

mass function

$$\Pr\{X = k\} = e^{-\lambda}\frac{\lambda^k}{k!},\ k = 0, 1, 2, \ldots$$

Proposition 5.5

If $(N_t)_{t\geqslant 0}$ is a Poisson process with rate $\lambda \geqslant 0$, then for each $t \geqslant 0$, we have $N_t \sim \text{Poisson}(\lambda t)$.

Theorem 5.6

Let $(N_t)_{t\geqslant 0}$ be a Poisson process with rate λ. For any finite collection of distinct time instants $0 = t_0 < t_1 < \cdots < t_k$, the increments $(N_{t_1} - N_{t_0}), \ldots, (N_{t_k} - N_{t_{k-1}})$ are independent with $(N_{t_i} - N_{t_{i-1}}) \sim \text{Poisson}(\lambda(t_i - t_{i-1}))$ for each $1 \leqslant i \leqslant k$.

Theorem 5.7: Characterization of Poisson Processes

If $(N_t)_{t\geqslant 0}$ is a Poisson process, then the following hold:

1. $N_0 = 0$;
2. $N_t \sim \text{Poisson}(\lambda t)\ \forall t \geqslant 0$;
3. $(N_t)_{t\geqslant 0}$ has independent increments.

Conversely, if these properties hold for a counting process $(N_t)_{t\geqslant 0}$, then it is a Poisson process.

5.3 Conditioning on Arrivals

Definition 5.8

Let $X_1, X_2, \ldots, X_k$ be a collection of random variables. The **order statistics** $X_{(1)}, \ldots, X_{(k)}$ are the random variables defined by sorting the realizations of $X_1, X_2, \ldots, X_k$ into increasing order.

Theorem 5.9

Let $(N_t)_{t\geqslant 0}$ be a Poisson process with arrivals $(T_i)_{i\geqslant 1}$. Conditioned on the event $\{N_t = n\}$, the vector of arrival times $(T_1, \ldots, T_n)$ has the same distribution as that of order statistics $(U_{(1)}, \ldots, U_{(n)})$, where $U_i \sim \text{Unif}(0, t),\ 1 \leqslant i \leqslant n$ are independent.

Theorem 5.10

Let $(N_t)_{t\geqslant 0}$ be a Poisson process with rate λ and corresponding arrivals $(T_n)_{n\geqslant 1}$. For a Borel set $B \subset [0, \infty)$, let $|B|$ denotes its Lebesgue volume, and let $N(B)$ denote the number of arrivals in B; i.e.,

$$N(B) = \#\{n \geqslant 1 : T_n \in B\}.$$

If $B_1, B_2, \cdots \subset [0, \infty)$ are disjoint, bounded Borel sets, then $N(B_1), N(B_2), \ldots$ are independent, with $N(B_i) \sim$

Poisson($\lambda|B_i|$).

Theorem 5.11: Slivnyak's Theorem

Let $(N_t)_{t\geqslant 0}$ be a Poisson process with rate λ and let $x \in (0,\infty)$. Conditioned on one arrival at time x, the other arrivals form an (unconditional) rate-λ Poisson process.

6 Continuous-Time Markov Chains

6.1 Definitions and Constructions

Definition 6.1

A process $(X_t)_{t\geqslant 0}$ taking values in $\mathcal{S}$ is a temporally homogeneous **continuous-time Markov chain** if:

(i) given any initial state $X_0 = i \in \mathcal{S}$, the sample paths $t \mapsto X_t$ are a.s. right-continuous (with respect to the discrete topology on $\mathcal{S}$); and

(ii) for any choice of discrete time instants $0 \leqslant t_1 < \cdots < t_k < t \leqslant s$ and states $i_1, i_2, \ldots, i_k, i, j \in \mathcal{S}$, we have the Markov property

$$\Pr\{X_s = j \mid X_t = i, X_{t_k} = i_k, \ldots, X_{t_1} = i_1\} = \Pr\{X_{s-t} = j \mid X_0 = i\}.$$

Theorem 6.2

Let $(X_t)_{t\geqslant 0}$ be a continuous-time Markov chain. The transition probabilities satisfy

$$P^{s+t} = P^s P^t \text{ for all } s, t \geqslant 0,$$

and $\lim_{t\downarrow 0} P^t = I$. In other words, the transition probabilities $(P^t)_{t\geqslant 0}$ form a **Markov semigroup**.

Theorem 6.3

Let $(X_t)_{t\geqslant 0}$ be a continuous-time Markov chain with initial non-absorbing state $X_0 = i$. The holding time $T = \inf\{t \geqslant 0 : X_t \neq i\}$ has distribution $T \sim \text{Exp}(\lambda_i)$ for $\lambda_i \geqslant 0$ satisfying

$$P_{ii}^h = 1 - h\lambda_i + o(h).$$

Moreover, the next state X_T is independent of T and has distribution

$$p_{ij} \coloneqq \Pr\{X_T = j \mid X_0 = i\} = \lim_{h\downarrow 0} \frac{P_{ij}^h}{1 - P_{ii}^h}, \quad j \neq i.$$

Theorem 6.4

Let $(X_t)_{t\geqslant 0}$ be a continuous-time Markov chain with transition probabilities $(P^t)_{t\geqslant 0}$, starting in non-absorbing state $X_0 = i$, and let $T = \inf\{t \geqslant 0 : X_t \neq i\}$ denote the time of the first transition. Conditioned on T and $X_T = j$, the process $(X_{T+t})_{t\geqslant 0}$ is a continuous-time Markov chain with transition probabilities $(P^t)_{t\geqslant 0}$ and starting state j.

Definition 6.5

The transition probabilities $(p_{ij})_{i,j\in\mathcal{S}}$ (with $p_{ii} = 0$) define a discrete-time Markov chain, known as the **embedded chain**. The parameters $(\lambda_i)_{i\in\mathcal{S}}$ are called the **transition rates** for the Markov chain, λ_i is

precisely the rate at which the process transitions out of state i.

Lemma 6.6

Let $(p_{ij})_{i,j\in\mathcal{S}}$ be transition probabilities for a discrete-time Markov chain $(X_n)_{n\geqslant 0}$ starting in non-absorbing state $X_0 = i$.

(i) The random variable $N := \inf\{n \geqslant 0 : X_n \neq i\}$ is geometric with distribution

$$\Pr\{N = k \mid X_0 = i\} = p_{ii}^{k-1}(1 - p_{ii}), \quad k \geqslant 1$$

(ii) The random variable X_N is independent of N, and has distribution

$$\Pr\{X_N = j \mid X_0 = i\} = \frac{p_{ii}}{(1 - p_{ii})}, \quad j \neq i.$$

6.2 The Infinitesimal Generator

Definition 6.7

The **infinitesimal generator** for a continuous-time Markov chain $(X_t)_{t\geqslant 0}$ with transition rates $(\lambda_i)_{i\in\mathcal{S}}$ is a matrix Q with entries

$$q_{ij} := [Q]_{ij} = \begin{cases} \lambda_i p_{ij} & \text{for } j \neq i \\ -\lambda_i & \text{for } j = i, \end{cases}$$

where $(p_{ij})_{i,j\in\mathcal{S}}$ are the transition probabilities for the embedded chain. In particular, $\lambda_i = \sum_{j\neq i} q_{ij}$.

The numbers $(q_{ij})_{i,j\in\mathcal{S}}$ are called the **jump rates** for the Markov chain. Essentially, q_{ij} describes the rate at which the Markov chain with infinitesimal generator Q transitions from state i to state j $(j \neq i)$.

6.2.1 The Kolmogorov differential equations

Corollary 6.8

Theorem 6.3 implies the following.

For a continuous-time Markov chain with infinitesimal generator Q, the transition probabilities satisfy

$$P_{ii}^h = 1 - h\lambda_i + o(h)$$

and

$$P_{ij}^h = hq_{ij} + o(h), \quad j \neq i.$$

Theorem 6.9: Kolmogorov Differential Equations

Let $(P^t)_{t\geqslant 0}$ be the transition semigroup for a minimal continuous-time Markov chain with infinitesimal generator Q. The map $t \mapsto P^t$ is continuously differentiable on $[0,\infty)$, and is the (unique) minimal non-negative solution to the differential equations

$$\frac{d}{dt}P^t = QP^t;\ P^0 = I \qquad \text{(Kolmogorov Backward Equation)}$$

and

$$\frac{d}{dt}P^t = P^tQ;\ P^0 = I. \qquad \text{(Kolmogorov Forward Equation)}$$

Remark 6.10

A Markov semigroup $(P^t)_{t\geqslant 0}$ with generator Q always satisfies Kolmogorov's backwards equation. In contrast, the forward equation is not satisfied in general, but is satisfied by $(P_t)_{t\geqslant 0}$ corresponding to the minimal construction of a Markov chain with generator Q.

In the case of finite state space, or more generally bounded transition rates, the Kolmogorov differential equations have a unique solution.

Corollary 6.11

Let $(P^t)_{t\geqslant 0}$ be the transition semigroup for a continuous-time Markov chain with infinitesimal general Q. If $\sup_{i\in S}\lambda_i < \infty$, then $P^t = e^{tQ} \coloneqq \sum_{k\geqslant 0} t^k \frac{Q^k}{k!}$ for all $t \geqslant 0$.

6.2.2 Criteria for non-explosiveness

Definition 6.12: Explosiveness

Define the time of explosion

$$T_\infty \coloneqq \sup_{n\geqslant 1} T_n,$$
$$T_n = \sum_{j=1}^{n} \tau_j,$$

where T_n denotes the time of the nth transition. The time of explosion is essentially the time at which an infinite number of transitions have taken place. When a process makes an infinite number of transitions in a finite time, it is known as an **explosion**. We say that a Markov chain is **non-explosive** if $T_\infty = +\infty$ a.s.; otherwise, the chain is said to be **explosive**. Necessary and sufficient conditions for a Markov chain to be non-explosive can be stated in terms of its infinitesimal generator.

Theorem 6.13: Reuter's Condition

A Markov chain with infinitesimal generator Q is non-explosive if and only if the only non-negative bounded solution $\nu = (\nu_i)_{i \in S}$ to $\nu = Q\nu$ is $\nu = 0$.

6.3 Continuous-time Markov Limit Theorems

Definition 6.14

A continuous-time Markov chain is **irreducible** if the embedded chain is irreducible and has at least two states.

Definition 6.15: Stationary distribution

A **stationary distribution** for a continuous-time Markov chain with transition probabilities $(P^t)_{t \geqslant 0}$ is a probability (row) vector p satisfying $p = pP^t$ for all $t \geqslant 0$.

Theorem 6.16

Consider a continuous-time Markov chain with infinitesimal generator Q. A probability vector p satisfying $\sum_i p_i \lambda_i < \infty$ is a stationary distribution if and only if $pQ = 0$. Moreover, if the chain is irreducible then p is the unique stationary distribution.

6.3.1 Stationary distributions and embedded chains

Corollary 6.17

Consider and irreducible continuous-time Markov chain with generator Q. The following are equivalent.

1. The continuous-time chain has stationary distribution p satisfying $\sum_i p_i \lambda_i < \infty$.
2. The embedded chain has stationary distribution π satisfying $\sum_i \pi_i / \lambda_i < \infty$.

Moreover, if either (and therefore both) are true, then the stationary distributions are unique, and $\pi_k = C^{-1} p_k \lambda_k$ for all k, where $C = \sum_i p_i \lambda_i$.

7 Hypothesis Testing

7.1 Binary Hypothesis Testing

Definition 7.1: The setup

On the basis of observing a sample $\omega \in \Omega$, we would like to decide whether $(\Omega, \mathcal{F}, P_0)$ or $(\Omega, \mathcal{F}, P_1)$ is the better model. The former is called the **null hypothesis** H_0, and the latter is called the **alternate hypothesis** H_1.

A **test** is a function $\hat{H} : \Omega \to \{H_0, H_1\}$ that is measurable in the sense that $\hat{H}^{-1}(H_0) \in \mathcal{F}$. Associated with any test $\hat{H}$ are two fundamental error probabilities: the **Type I error rate** (or, false positive probability), and the **Type II error rate** (or, false negative probability). More precisely,

$$P_0\{\hat{H} = H_1\} =: \text{Type I error rate (or, false positive probability)}$$
$$P_1\{\hat{H} = H_0\} =: \text{Type II error rate (or, false negative probability).}$$

The **power** of a test $\hat{H}$ is the probability of avoiding a Type II error, and is therefore equal to $P_1\{\hat{H} = H_1\}$.

7.1.1 The likelihood ratio

We assume henceforth that $P_1 \ll P_0$.

Definition 7.2

The Radon-Nikodym theorem ensures there is a $\mathcal{F}-$measurable, P_0-a.s. unique function $\Lambda : \Omega \to [0, \infty)$ satisfying the "**change of measure**" identity

$$\mathbb{E}_{P_1}[1_A] = \mathbb{E}_{P_0}[\Lambda 1_A], \quad \text{for all } A \in \mathcal{F}.$$

The function Λ is called a Radon-Nikodym derivative (usually denoted by $\frac{dP_1}{dP_0}$), but in the context of hypothesis testing it is generally referred to as the **likelihood ratio** because it can be thought of as the relative likelihood of observing a sample ω under the different hypotheses H_1 and H_0. Λ is simply the ratio of densities, or if Ω is discrete, then Λ is the ratio of the probability mass functions.

7.1.2 Threshold tests and the error curve

Definition 7.3

Assume $P_1 \ll P_0$, and let $\eta \geqslant 0$. The threshold test with threshold η, denote $\hat{H}_\eta$, is defined according to

$$\hat{H}_\eta = \begin{cases} H_1 & \text{if } \Lambda(\omega) \geqslant \eta \\ H_0 & \text{if } \Lambda(\omega) < \eta. \end{cases}$$

Example 7.4: MAP and ML tests

The **maximum a posteriori test (MAP)** is a threshold test where we have a prior belief that H_0 is true with probability $\pi_0 < 1$ and H_1 is true with probability $\pi_1 = 1 - \pi_0$. The MAP test is the threshold test with threshold $\eta = \pi_0/\pi_1$, and has the property that it minimizes the total error rate among all tests. Under prior π, the total error probability for any test $\hat{H}$ satisfies

$$\Pr\left\{\hat{H} \text{ errors}\right\} = \pi_0 P_0\left\{\hat{H} = H_1\right\} + \pi_1 P_1\left\{\hat{H} = H_0\right\} \geqslant \Pr\left\{\hat{H}_{\text{MAP}} \text{ errors}\right\}.$$

The **maximum likelihood (ML) test** is defined to be the threshold test with $\eta = 1$. It minimizes the sum of Type I and Type II error rates, which follows from the previous example when $\pi_0 = \pi_1$.

Definition 7.5

The threshold tests $(\hat{H}_\eta)_{\eta \geqslant 0}$ define a function called the **error curve**, which plays a fundamental role in characterizing the best tradeoff between Type I and Type II error rates.

Assume $P_1 \ll P_0$ and let Λ denote the likelihood ration. The error curve $u : [0,1] \to \mathbb{R}$ is defined via

$$u(\theta) \coloneqq \sup_{\eta \geqslant 0}\left\{P_1\left\{\hat{H}_\eta = H_0\right\} + \eta\left(P_0\left\{\hat{H}_\eta = H_1\right\} - \theta\right)\right\}, \quad 0 \leqslant \theta \leqslant 1.$$

Note that, as the pointwise supremum of affine functions, u is a convex function on $[0,1]$.

7.1.3 The Neyman-Pearson lemma

We say that a test $\hat{H}$ **lies above the error curve** if

$$P_1\left\{\hat{H} = H_0\right\} \geqslant u(P_0\left\{\hat{H} = H_1\right\}),$$

and we say that $\hat{H}$ **lies on the error curve** if this is met with equality.

Theorem 7.6: Neyman-Pearson Lemma

Assume $P_1 \ll P_0$. All tests $\hat{H}$ lie above the error curve. Moreover, every threshold test $\hat{H}_\eta$ lies on the error curve. The Neyman–Pearson lemma ensures that threshold tests are optimal in the sense that they lie on the error curve, and any other test lies above.

Definition 7.7: Randomized threshold test

Fix parameters $\eta_0, \eta_1 \geqslant 0$ and $p \in [0,1]$. The corresponding (randomized) threshold test $\hat{H}$ is defined by taking $R \sim \text{Bernoulli}(p)$, and putting

$$\hat{H}(\omega) = 1_{R=0}\hat{H}_{\eta_0}(\omega) + 1_{R=1}\hat{H}_{\eta_1}(\omega).$$

As a result, by varying the parameters $\eta_0, \eta_1 \geqslant 0$ and $p \in [0,1]$, any point on the error curve is achievable by a (possibly randomized) threshold test. For a fixed $\theta \in [0,1]$, the **Neyman-Pearson rule** is defined

to be the (possibly randomized) threshold test that achieves Type II error probability $u(\theta)$ subject to the constant that Type I error probability does not exceed θ. In other words, subject to a Type I error constraint, the Neyman–Pearson rule is the most powerful test.

Definition 7.8: Sufficient statistic

A mapping $T : \omega \mapsto T(\omega)$ is said to be a **sufficient statistic** if there exists a function ν such that $\Lambda(\omega) = \nu \circ T(\omega)$ for all $\omega \in \Omega$.

7.2 Sequential Analysis

Definition 7.9

Let P_0, P_1 be probability distributions on, say, $\mathbb{R}$, and consider i.i.d. random variables $(X_n)_{n\geqslant 1}$, with common distribution Q. Let the null hypothesis be $H_0 : Q = P_0$, and the alternate hypothesis be $H_1 : Q = P_1$. Let $(\hat{H}_n)_{n\geqslant 1}$ be a sequence of tests adapted to $(X_n)_{n\geqslant 1}$, and T be a stopping time. This induces Type I and Type II error rates for the **sequential test** $\hat{H}_T$ equal to

$$\alpha \coloneqq P_0\left\{\hat{H}_T = H_1\right\}, \text{ and } \beta \coloneqq P_1\left\{\hat{H}_T = H_0\right\},$$

respectively, where we abuse notation and abbreviate

$$P_i\{\cdot\} = \Pr\{\cdot \mid H_i \text{ true}\}, \quad i = 0, 1.$$

7.2.1 Average sample requirements

Definition 7.10

For probability measures $u \ll v$ with corresponding likelihood ratio, the relative entropy between u and v is defined as

$$D(u \,\|\, v) \coloneqq \mathbb{E}_u[\log \Lambda] = \mathbb{E}_v[\Lambda \log \Lambda].$$

The entropy is greater than or equal to 0 due to convexity, with equality if and only if $u = v$. For two reals $p, q \in [0, 1]$, we abuse notation slightly and write

$$D(p \,\|\, q) \coloneqq p \log\left(\frac{p}{q}\right) + (1-p)\log\left(\frac{1-p}{1-q}\right).$$

Theorem 7.11

Let the above notation prevail, and let $\mathbb{E}_i[\cdot]$ denote the expectaion under hypothesis H_i for $i = 0, 1$. If $\alpha + \beta \leqslant 1$, it holds that

$$\mathbb{E}_0[T] \geqslant \frac{D(\alpha \,\|\, 1-\beta)}{D(P_0 \,\|\, P_1)}, \text{ and } \mathbb{E}_1[T] \geqslant \frac{D(1-\beta \,\|\, \alpha)}{D(P_1 \,\|\, P_0)}.$$

Note that T is the number of samples.

7.2.2 The sequential probability ratio test

Definition 7.12

Consider the setting where $P_0 \neq P_1$, and assume $P_1 \ll P_0$, with likelihood ratio $\Lambda = \frac{dP_1}{dP_0}$. Fix two thresholds $\eta_0 < \eta_1$. For i.i.d. observations $(X_n)_{n \geqslant 1}$ as before, define the sequence of likelihoods

$$L_n = \prod_{i=1}^{n} \Lambda(X_i), \quad n \geqslant 1.$$

Define the stopping time $T = \inf\{n \geqslant 1 : L_n \notin (\eta_0, \eta_1)\}$, and the corresponding **sequential probability ratio test**

$$\hat{H}_T = \begin{cases} H_0 & \text{if } L_T \leqslant \eta_0 \\ H_1 & \text{if } L_T \geqslant \eta_0. \end{cases}$$

This induces Type I and Type II error rates for the sequential test $\hat{H}_T$ equal to

$$\alpha \coloneqq P_0\left\{\hat{H}_T = H_1\right\}, \text{ and } \beta \coloneqq P_1\left\{\hat{H}_T = H_0\right\}.$$

Proposition 7.13

For thresholds $0 \leqslant \eta_0 < \eta_1$, the Type I/II error rates for the corresponding sequential probability ratio test $\hat{H}_T$ satisfy

$$\frac{\alpha}{1-\beta} \leqslant \frac{1}{\eta_1} \text{ and } \frac{\beta}{1-\alpha} \leqslant \eta_0.$$

Theorem 7.14

Fix thresholds $\eta_0 < \eta_1$, and let α, β denote the Type I/II error rates realized by the corresponding sequential probability ratio test $\hat{H}_T$. If $D(P_1 \,\|\, P_0) < \infty$ and $D(P_0 \,\|\, P_1) < \infty$, then

$$\mathbb{E}_0[T] \simeq \frac{D(\alpha \,\|\, 1-\beta)}{D(P_0 \,\|\, P_1)}, \text{ and } \mathbb{E}_1[T] = \frac{D(1-\beta \,\|\, \alpha)}{D(P_1 \,|\, P_0)}.$$

In this case, the approximations above are fairly accurate, demonstrating (approximate) optimality of the sequential probability ratio test. This assertion of optimality can be made more precise: among all tests with the same power, the sequential probability ratio test requires the fewest samples on average.

Lemma 7.15

Let $(Z_n)_{n\geqslant 1}$ be i.i.d. random variables. For given $a < b$, define the stopping time

$$K = \inf\left\{ n \geqslant 1 : \sum_{i=1}^{n} Z_i \notin (a, b) \right\}.$$

If $\Pr\{|Z_1| > 0\} > 0$, then $\mathbb{E}[K] < \infty$.

This lemma guarantees that any sequential probability ratio test will require finitely many samples on average.