Draft date: January 9, 2019

IFT-6561

APPENDIX

Pierre L'Ecuyer
Département d'Informatique et de Recherche Opérationnelle
Université de Montréal

A. A Review of Probability and Statistics

This appendix gives a condensed summary of basic concepts and results in probability, statistics, and stochastic modeling. Its aim is *not* to replace a good textbook. For more detailed treatments, the reader may consult, for example, Billingsley (1986), Chung (1974), Wolff (1989) for probability and Markov models and Hogg and Craig (1995), Rice (1995), Serfling (1980), Shao (1999) for statistics.

A.1 Probabilities

The behavior of a stochastic model depends, by definition, on the *realization* ω of a random phenomenon. This ω represents all sources of randomness in the model. It must belong to the set Ω of all possible realizations (or outcomes), called the *sample space*.

To define *probabilities* for these realizations, we must first select a family $\mathcal{F}$ of subsets of Ω for which the probabilities will be defined. The members of $\mathcal{F}$ are called the *measurable sets*, or *events* (not to be confounded with the *events* in discrete event simulation, where the same word has a different meaning). The set $\mathcal{F}$ must satisfy the conditions of a σ-field: (a) Ω itself must belong to $\mathcal{F}$, (b) if $B \in \mathcal{F}$ then its complement $\Omega \setminus B$ is also in $\mathcal{F}$, and (c) if $B_1, B_2, \ldots$ are in $\mathcal{F}$ then their (denumerable) union $\cup_{i=1}^{\infty} B_i$ is also in $\mathcal{F}$. The pair $(\Omega, \mathcal{F})$ is called a *measurable space*. Note that the smallest possible $\mathcal{F}$ contains only Ω and the empty set.

A *signed measure* $\mathbb{Q}$ is a function that assigns a value in $(-\infty, \infty] = \mathbb{R} \cup \{\infty\}$ to each event $B \in \mathcal{F}$, and which is countably additive: If the B_i's are disjoint sets in $\mathcal{F}$ then $\mathbb{Q}(B_1 \cup B_2 \cup \cdots) = \mathbb{Q}(B_1) + \mathbb{Q}(B_2) + \cdots$. The measure is *positive* if it can never take negative values. A *probability measure* is a positive measure $\mathbb{P}$ for which $\mathbb{P}(\Omega) = 1$. When $\mathbb{P}$ is a probability measure on $(\Omega, \mathcal{F})$, the mathematical structure $(\Omega, \mathcal{F}, \mathbb{P})$ is called a *probability space*.

Example A.1 Suppose we throw two dice of different colors and observe the two numbers showing up. Then Ω can be defined as the set of all 36 possible pairs that we can obtain, $\mathcal{F}$ can contain all subsets of Ω, and a probability measure on $\mathcal{F}$ can be defined by $\mathbb{P}(B) = |B|/36$ for all $B \subseteq \Omega$. This corresponds to the idea of two independent fair dice. □

If Ω is finite or denumerable, $\mathcal{F}$ can be taken as the family of all its possible subsets, but if Ω is non-denumerable (for example, an interval of the real line), then it turns out that one cannot take $\mathcal{F}$ as the set of all subsets of Ω, because many such subsets are too "weird" and cause trouble. This gives rise to technical subtleties studied by *measure theory* (Billingsley 1986). When $\Omega = \mathbb{R}$, $\mathcal{F}$ is usually taken as the *Borel σ-field* $\mathcal{B}$, defined as the smallest σ-field that contains all the intervals with rational end points.

In a complex stochastic simulation model, ω can be seen as an infinite sequence of random bits and the model's behavior can always be expressed as a function of the values taken by those bits. Another, perhaps more natural, way of interpreting ω in the context of simulation is as an infinite sequence of real numbers in the interval $[0, 1]$, say $U_0, U_1, U_2, \ldots$, such that for all integers $s > 0$, $\mathbf{U}_s = (U_0, U_1, \ldots, U_{s-1})$ behaves like a point selected at random uniformly in the unit hypercube $[0, 1]^s$. This means that if $B \subseteq [0, 1]^s$ and B is measurable, then $\mathbb{P}(\mathbf{U}_s \in B) = \text{volume}(B)$. The aim of random number generators in simulation is to *imitate* this type of behavior. The simulation program will take the ω provided by the generator and transform it (often in complicated ways) to get the desired result. Note that in simulation applications, we usually take the open interval $(0, 1)$ instead of the close interval $[0, 1]$, because the transformations of interest are sometimes infinite or undefined at 0 or at 1.

If $B \in \mathcal{F}$ and $\mathbb{P}[B] = 1$, we say that B occurs *with probability 1 (w.p.1)* or *almost surely (a.s.)*. We have $\mathbb{P}[B] = 1$ if and only if $\mathbb{P}[B^c] = 0$ where $B^c = \Omega \setminus B$ is the complement of B. However, $\mathbb{P}[B^c] = 0$ does not mean that B^c cannot happen or that B is sure to happen. For example, if Ω is the unit interval $[0, 1]$, $\mathcal{F}$ the Borel σ-field in $[0, 1]$, and $\mathbb{P}$ the *Lebesgue measure* on $[0, 1]$, for which the probability of an interval equals its length, then any single point $\omega \in [0, 1]$ has probability 0. However, ω must take some value and the corresponding event $\{\omega\}$ will then happen, even though its probability was zero.

If A and B are two events (in $\mathcal{F}$) such that $\mathbb{P}[B] > 0$, we define the *conditional probability* of A given B by

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

The events A and B are called *independent* when $\mathbb{P}(A \mid B) = \mathbb{P}(A)$, i.e., $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. In general, n events $B_1, B_2, \ldots, B_n$ are *independent* if and only if $\mathbb{P}(B_1 \cap B_2 \cap \cdots \cap B_n) = \mathbb{P}(B_1)\mathbb{P}(B_2)\cdots\mathbb{P}(B_n)$. They are *pairwise independent* (a weaker property) if $\mathbb{P}(B_i \cap B_j) = \mathbb{P}(B_i)\mathbb{P}(B_j)$ for all $i \neq j$. The probability of the union of several events obeys the *inclusion-exclusion formula*:

$$\mathbb{P}[B_1 \cup \cdots \cup B_n] = \sum_{i=1}^{n} \mathbb{P}[B_i] - \sum_{1 \le i < j \le n} \mathbb{P}[B_i \cap B_j] + \sum_{1 \le i < j < \le n} \mathbb{P}[B_i \cap B_j \cap B_k] - \cdots.$$

More generally, let $\mathcal{G} \subset \mathcal{F}$ be a σ-field contained in $\mathcal{F}$ and suppose that for each $B \in \mathcal{G}$, we know whether or not B has occurred. Informally, the probabilities that we have after we know all this information define the *probability*

distribution conditional on $\mathcal{G}$. The corresponding conditional probabilities are denoted by $\mathbb{P}(A \mid \mathcal{G})$ for each $A \in \mathcal{F}$. They are in fact random variables: their values depend on which events in $\mathcal{G}$ have occurred. Note that if $A \in \mathcal{G}$, then $\mathbb{P}(A \mid \mathcal{G})$ is always 0 or 1, because $\mathcal{G}$ "tells us" whether A has occurred or not.

Example A.2 In Example A.1, let B_j be the set of realizations for which the sum over the two dice is equal to j, for $j = 2, \ldots, 12$. Let $\mathcal{G}$ be the class of subsets defined as the union of any number of those B_j's (including none of them, in which case we have the empty set). Then the probabilities conditional on $\mathcal{G}$ represent the probabilities conditional on knowing the sum over the two dice. In particular, if A is the event that we have a 6 on the first die, then the reader can verify (as an exercise) that $\mathbb{P}[A \mid B_j] = 1/(13 - j)$ for $j = 7, \ldots, 12$, and is 0 for $j \leq 6$. To compute the probability of A conditional on $B_{10} \cup B_{11} \cup B_{12}$, observe that $B_{10} \cup B_{11} \cup B_{12}$ contains six of the 36 possible outcomes, and there is a 6 on the first die for three of them, so $\mathbb{P}[A \mid B_{10} \cup B_{11} \cup B_{12}] = 1/2$. □

A more rigorous definition of conditional probabilities is given later, based on the definition of conditional expectation. It permits one to condition on events having probability 0.

A.2 Integrals

Given two measurable spaces $(\Omega, \mathcal{F})$ and $(\Omega', \mathcal{F}')$, a function $f : \Omega \to \Omega'$ is a *measurable function* if for each $B \in \mathcal{F}'$, $\{\omega : f(\omega) \in B\} \in \mathcal{F}$. Measurability depends on the σ-fields that have been selected. Often, $(\Omega', \mathcal{F}') = (\mathbb{R}, \mathcal{B})$, the real space with the Borel sigma-field, and we call the measurable functions $\mathcal{F}$-measurable to emphasize that their measurability depends on $\mathcal{F}$.

The *Lebesgue integral* of a $\mathcal{F}$-measurable function $f : \Omega \to \mathbb{R}$, with respect to a positive and finite measure $\mathbb{Q}$ on $(\Omega, \mathcal{F})$, can be defined as follows. This integral is denoted

$$\int_\Omega f(\omega)\mathbb{Q}(\mathrm{d}\omega) \quad \text{or more simply} \quad \int f \mathrm{d}\mathbb{Q}.$$

If f is a linear combination of indicator functions,

$$f(\omega) = \sum_{i=1}^{n} a_i \mathbb{I}[\omega \in B_i]$$

where a_i is a constant and $B_i \in \mathcal{F}$ for each i, then

$$\int_\Omega f(\omega)\mathbb{Q}(\mathrm{d}\omega) = \sum_{i=1}^{n} a_i \mathbb{Q}[B_i].$$

If $f \geq 0$, i.e., f cannot take negative values, then it can be proved that there is an increasing sequence of functions $f_1 \leq f_2 \leq f_3 \leq \cdots$ such that each f_j is

a linear combination of indicator functions and $f(\omega) = \lim_{j\to\infty} f_j(\omega)$. We then define

$$\int_\Omega f(\omega)\mathbb{Q}(\mathrm{d}\omega) = \lim_{j\to\infty} \int_\Omega f_j(\omega)\mathbb{Q}(\mathrm{d}\omega),$$

which could be infinite but always exists. If f can take negative values, then we can decompose it as $f = f^+ - f^-$ where f^+ and f^- are both positive and $\mathcal{F}$-measurable. If both f^+ and f^- have finite integrals, then f is called *$\mathbb{Q}$-integrable* and we define

$$\int_\Omega f(\omega)\mathbb{Q}(\mathrm{d}\omega) = \int_\Omega f^+(\omega)\mathbb{Q}(\mathrm{d}\omega) - \int_\Omega f^-(\omega)\mathbb{Q}(\mathrm{d}\omega).$$

If $\{f_n,\, n \geq 0\}$ is a sequence of $\mathcal{F}$-measurable functions $f_n : \Omega \to \mathbb{R}$, there are many situations where we would like to interchange the limit and the integral as follows:

$$\lim_{n\to\infty} \int_\Omega f_n(\omega)\mathbb{Q}(\mathrm{d}\omega) = \int_\Omega \left(\lim_{n\to\infty} f_n(\omega)\right) \mathbb{Q}(\mathrm{d}\omega). \tag{A.1}$$

The following theorems provide sufficient conditions for the interchange to be valid:

Theorem A.1 (Monotone convergence Theorem). *If $f_n \geq f_{n-1} \geq \cdots \geq 0$ for all n, then (A.1) is valid.*

Theorem A.2 (Dominated convergence Theorem). *If $|f_n| \leq g$ for all n, where $g : \Omega \to \mathbb{R}$ is $\mathcal{F}$-measurable and $\int_\Omega g(\omega)\mathbb{Q}(\mathrm{d}\omega) < \infty$, then (A.1) is valid.*

A.3 Change of Measure and Densities

Suppose that $\mathbb{P}$ and $\mathbb{Q}$ are two measures on $(\Omega, \mathcal{F})$ such that $\mathbb{Q}$ dominates $\mathbb{P}$ in the sense that for all measurable sets $A \in \mathcal{F}$, $\mathbb{P}[A] > 0$ implies that $\mathbb{Q}(A) > 0$. Then we say that $\mathbb{P}$ is *absolutely continuous* with respect to $\mathbb{Q}$, and for $A \in \mathcal{F}$ we can write

$$\int_A \mathrm{d}\mathbb{P}(\omega) \;=\; \int_A [(\mathrm{d}\mathbb{P}/\mathrm{d}\mathbb{Q})(\omega)]\, \mathrm{d}\mathbb{Q}(\omega)$$

where $(\mathrm{d}\mathbb{P}/\mathrm{d}\mathbb{Q})(\omega)$ is the density (or Random-Nikodym derivative) of $\mathbb{P}$ with respect to $\mathbb{Q}$.

If $\mathbb{P}$ and $\mathbb{Q}$ have densities f and g with respect to the Lebesgue measure, then $(\mathrm{d}\mathbb{P}/\mathrm{d}\mathbb{Q})$ is the ratio of these densities. Likewise, if Ω is a discrete set and $\mathbb{P}$ and $\mathbb{Q}$ have densities with respect to the counting measure over Ω (for which the measure of a set is just its cardinality), then $(\mathrm{d}\mathbb{P}/\mathrm{d}\mathbb{Q})(\omega) = \mathbb{P}(\omega)/\mathbb{Q}(\omega)$, a ratio of measures.

A.4 Random Variables

Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a (real-valued) *random variable (r.v.)* is a $\mathcal{F}$-measurable function $X : \Omega \to \mathbb{R}$, which to each outcome $\omega \in \Omega$ assigns a real number $X(\omega)$. Usually, this real number (before its value is known) is simply denoted by X. Recall that *$\mathcal{F}$-measurable* means that $\{\omega : X(\omega) \in B\} \in \mathcal{F}$ for each $B \in \mathcal{B}$ where $\mathcal{B}$ is the Borel σ-field on $\mathbb{R}$. Then, X defines a probability measure $\tilde{\mathbb{P}}$ on $(\mathbb{R}, \mathcal{B})$ via $\tilde{\mathbb{P}}(B) = \mathbb{P}(\{\omega : X(\omega) \in B\})$, usually denoted by $\mathbb{P}(X \in B)$ for each $B \in \mathcal{B}$. This determines the *probability distribution* of X. In this book, we use the notation "$X \sim$ ..." to mean "the probability distribution of X is ..." and $X \sim Y$ to mean "X and Y have the same probability distributions.".

Example A.3 In Example A.1, where we throw two independent dice, we have $\Omega = \{(1,1),(1,2),\ldots,(6,6)\}$. Let X be the sum of values on the two dice. This X is a random variable that can take the values $2, 3, \ldots, 12$. We have $\mathbb{P}(X = x) = \mathbb{P}(\{\omega = (\omega_1, \omega_2) : \omega_1 + \omega_2 = x\})$. For example, $\mathbb{P}(X = 5) = \mathbb{P}(\{(1,4),(2,3),(3,2),(4,1)\}) = 4/36 = 1/9$. □

Example A.4 In a telephone call center, the waiting time of each customer and the duration of each call can be modeled as random variables that can take any value in the real interval $[0, \infty)$. For such a random variable X, the sample space could be taken as $(0,1)$ and X could be defined as $X = \varphi(\omega)$ for an appropriate transformation $\varphi : (0,1) \to [0,\infty)$. The inversion method to generate random variables operates that way. □

The function F defined by

$$F(x) = \mathbb{P}[X \le x] \qquad \text{for all } x \in \mathbb{R}$$

is called the *cumulative distribution function* (cdf) of X. This function is always nondecreasing and goes from 0 to 1 (unless X can be $\pm\infty$ with positive probability, which happens sometimes). It defines the probability distribution of X in a unique way.

♣ Add a picture.

A random variable X has an *absolutely continuous* cdf with respect to the Lebesgue measure if we can write

$$F(x) = \int_{-\infty}^{x} f(y)\mathrm{d}y,$$

for some function $f : \mathbb{R} \to [0,\infty)$ called the *density* of X (with respect to the Lebesgue measure). An equivalent condition for the existence of f is that $\mathbb{P}[X \in A] = 0$ for each $A \in \mathcal{B}$ of Lebesgue measure zero. For an absolutely continuous random variable, we always have $f(x) \ge 0$, $\mathbb{P}[X = x] = 0$ for all x,

$$\mathbb{P}[a \le X \le b] = F(b) - F(a) = \int_{a}^{b} f(x)\mathrm{d}x,$$

$\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1$, and $f(x) = F'(x)$ when the derivative exists. It is important to understand that $f(x)$ is *not* a probability. It is customary to just say that X has a continuous distribution, even though one can construct examples where the cdf is continuous but not absolutely continuous. Examples of continuous distributions include the uniform, exponential, normal, and chi-square distributions (see Chapter 2).

A (real-valued) random variable is called *discrete* (or is said to have a *discrete distribution*) if it takes its values in a *denumerable* subset of the real numbers, say $\{x_0, x_1, x_2, \ldots\}$. This set is often the set of non-negative integers, $\{0, 1, 2, \ldots\}$. For a discrete random variable, the function p defined by $p(x_i) = \mathbb{P}[X = x_i]$ is called the *probability mass function* (pmf) of X and we have $F(x) = \sum_{x_i \le x} p(x_i)$. The Bernoulli, binomial, Poisson, and geometric distributions are examples of discrete distributions (see Chapter 2). Note that the pmf is also the density of the probability measure $\mathbb{P}$ with respect to the *counting measure*, which gives a weight of 1 to each possible value, so its measure of a set is just the cardinality of that set. In this sense, p can also be viewed as a pdf.

Some random variables X are neither purely discrete nor purely continuous. For example, there could be a probability mass at some points and a density elsewhere. In most interesting (one-dimensional) cases, X can be written as $X = qX_1 + (1-q)X_2$ where X_1 is discrete, X_2 is continuous, and $0 \le q \le 1$.

The *reliability function* (or *survival function*) of X is defined by

$$\bar{F}(x) = \mathbb{P}[X > x].$$

If X is continuous, the *failure rate* is defined by

$$r(x) = f(x)/(1 - F(x)).$$

For a small $\epsilon > 0$, we have

$$\mathbb{P}[X < x + \epsilon | X > x] = \frac{\mathbb{P}[x < X < x + \epsilon]}{\mathbb{P}[X > x]} \approx \frac{f(x)}{1 - F(x)}\,\epsilon = r(x)\epsilon.$$

That is, if X denotes the age of first failure of a system, then $r(x)\epsilon$ represents (approximately) the probability that a failure occurs in the next ϵ units of time given that the system has survived until age x.

A.5 Mathematical Expectation and Variance

The *mathematical expectation* (or *theoretical average*) of a real-valued random variable X is defined by the integral

$$\mathbb{E}[X] = \int_{\Omega} X(\omega)\mathbb{P}(\mathrm{d}\omega).$$

If X is *discrete* with mass function p, this general expression boils down to

$$\mathbb{E}[X] = \sum_{i=0}^{\infty} x_i p(x_i),$$

and if X is *continuous* with density f, it becomes

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x) \mathrm{d}x.$$

If $g : \mathbb{R} \to \mathbb{R}$ is a measurable function, then $Y = g(X)$ is also a random variable and

$$\mathbb{E}[Y] = \mathbb{E}[g(X)] = \int_{\Omega} g(X(\omega)) \mathbb{P}(\mathrm{d}\omega).$$

If X has cdf F, the latter integral is sometimes denoted by

$$\int_{-\infty}^{\infty} g(x) \mathrm{d}F(x).$$

This notation is the old-fashioned *Riemann-Stieltjes integral*. In the case where X has a density f, this integral is equivalent to the ordinary *Riemann integral*

$$\int_{-\infty}^{\infty} g(x) f(x) \mathrm{d}x$$

(provided that gf is Riemann-integrable.) The Riemann-Stieltjes integral is more general than the Riemann integral because (for instance) it covers the case where F has jumps (which correspond to masses of probability at some points). The Lebesgue integral (see Section A.2) is more general than these two.

Proposition A.3 *If X cannot take negative values, then*

$$\mathbb{E}[X] = \int_{0}^{\infty} (1 - F(x)) \mathrm{d}x.$$

Proof. Since $X = \int_0^\infty \mathbb{I}[X > x] \mathrm{d}x$, we have

$$\mathbb{E}[X] = \mathbb{E}\left[\int_0^\infty \mathbb{I}[X > x] \mathrm{d}x\right] = \int_0^\infty \mathbb{E}[\mathbb{I}[X > x]] \mathrm{d}x = \int_0^\infty \mathbb{P}[X > x] \mathrm{d}x.$$

The *variance* of X is defined by

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

It is sometimes denoted $\sigma^2(X)$. We always have $\mathrm{Var}[X] \geq 0$. Physically, the variance represents the moment of inertia of the pdf (or pmf) of X with respect to its mean. Its square root is called the *standard deviation* and the ratio $\sqrt{\mathrm{Var}[X]}/\mathbb{E}[X] = \sigma(X)/\mathbb{E}[X]$ is the *coefficient of variation*, also called the *relative error*.

It is easy to verify that if $Y = aX + b$ then $\mathbb{E}[Y] = a\mathbb{E}[X] + b$ and $\mathrm{Var}[Y] = a^2\mathrm{Var}[X]$.

Example A.5 The normal distribution with mean μ and variance σ^2, denoted $\mathsf{N}(\mu, \sigma^2)$ or $N(\mu, \sigma^2)$, is a continuous distribution with density $f(x) = (\sigma\sqrt{2\pi})^{-1} \exp[(x-\mu)^2/(2\sigma^2)]$ for $-\infty < x < \infty$. This density is shown in Figure A.1, in which σ indicates the standard deviation. The figure also shows the corresponding cdf, denoted by Φ. The $N(0,1)$ distribution is called the *standard normal*. □

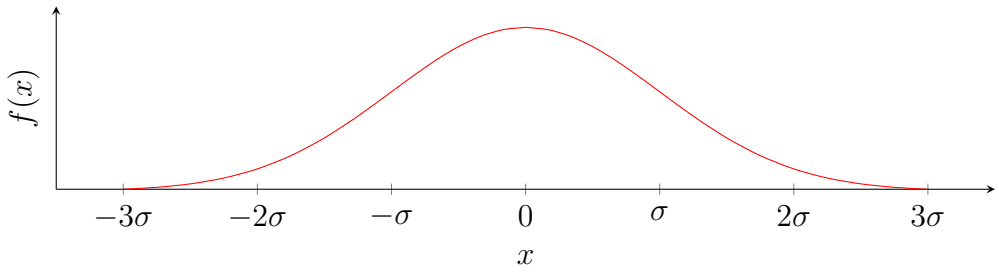

Fig. A.1. Density of the normal distribution with mean $\mu = 0$

More generally, for a random variable X with mean μ and variance σ^2, the *kth moment* of X and *kth centered moment* of X are defined as $\mathbb{E}[X^k]$ and $\mathbb{E}[(X-\mu)^k]$, respectively. The third and fourth moments of $(X-\mu)/\sigma$ are the *skewness coefficient* (or *coefficient of asymmetry*), $\nu = \mathbb{E}[(X-\mu)^3]/\sigma^3$, and the *kurtosis coefficient*, $\kappa = \mathbb{E}[(X-\mu)^4]/\sigma^4$. We have $\nu = 0$ when the probability mass (or density) function is symmetric with respect to its mean (as for the normal or Student-t distributions), $\nu > 0$ if it is skewed to the right (as for the exponential or chi-square distribution; see Figure A.2), and $\nu < 0$ if it is skewed to the left. The kurtosis measures the thickness of the tails. For example, the normal distribution has $\kappa = 3$, while the Student-t has $\kappa > 3$ and κ decreases with the number of degrees of freedom.

The *moment-generating function* (mgf) of a random variable X is defined by

$$M_X(\theta) = \mathbb{E}\left[e^{\theta X}\right] = \int_{-\infty}^{\infty} e^{\theta x} \mathrm{d}F(x), \qquad \text{for all } \theta \in \mathbb{R},$$

when this expectation exists. When this function exists in a neighborhood of $\theta = 0$, it determines the entire distribution of X in a unique way, and we obtain the jth moment of X by taking its jth derivative evaluated at 0 (whence its name):

$$\mathbb{E}[X^j] = \left.\frac{d^j M_X(\theta)}{d\theta^j}\right|_{\theta=0}.$$

In fact, all these moments exist if and only if $M_X(\theta)$ exists in some open interval that contains 0. If X is continuous with density f, then $M_X(\theta) = \int_{-\infty}^{\infty} e^{\theta x} f(x)\mathrm{d}x$,

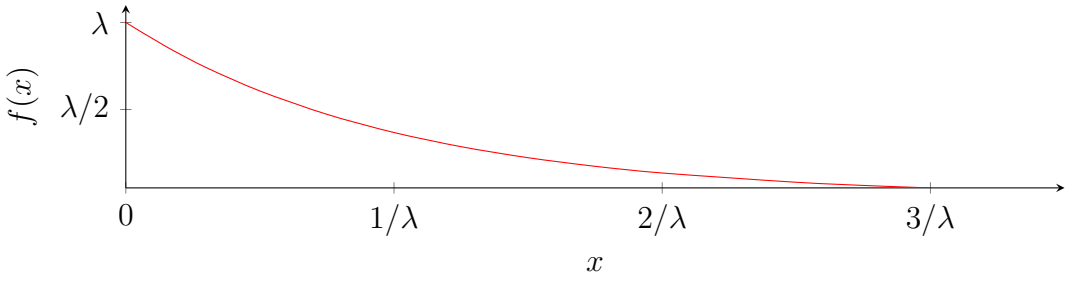

Fig. A.2. The exponential density with mean $1/\lambda$ is skewed to the right (i.e., $\nu > 0$)

and $M_X(-\theta)$ is also the (two-sided) *Laplace transform* of X. In that case,

$$f_\theta(x) = e^{\theta x} f(x)/M_X(\theta) \tag{A.2}$$

is also a probability density function, called an *exponentially twisted* version of $f(x)$. Such *exponential twisting*, for carefully selected θ, is widely used for importance sampling (see Chapter 6).

The natural logarithm of $M_X(\theta)$ (when it exists) is the *cumulant-generating function* of X:

$$\Psi_X(\theta) = \ln(M_X(\theta))$$

The jth derivative of $\Psi_X(\theta)$ evaluated at $\theta = 0$, when it exists, is the *jth cumulant* of X,

$$\kappa_j = \Psi_X^{(j)}(0) = \left.\frac{d^j \Psi_X(\theta)}{d\theta^j}\right|_{\theta=0}.$$

The first two cumulants κ_1 and κ_2 are the mean and the variance. If X is continuous with density f, then the exponentially twisted random variable with density f_θ defined in (A.2) has mean $\Psi'(\theta)$, variance $\Psi''(\theta)$, and similarly for higher moments.

The moment generating function may not exist in some cases, but the complex-valued *characteristic function* of X, defined as

$$\varphi_X(\theta) = M_X(\mathrm{i}\theta) = \mathbb{E}[e^{\mathrm{i}\theta X}]$$

for all $\theta \in \mathbb{R}$, where $\mathrm{i} = \sqrt{-1}$, always exists. The jth moment of X can also be obtained by evaluating the jth derivative of φ_X at 0: $\mathbb{E}[X^j] = (-\mathrm{i})^j \varphi_X^{(j)}(0)$.

For a discrete random variable X with probability mass function $p(x)$ for $x = 0, 1, 2, \ldots$, one often works with the *probability generating function* (or *z-transform*) of X, defined as

$$G(z) = M_X(\ln z) = \mathbb{E}[z^X] = \sum_{x=0}^{\infty} p(x) z^x.$$

A random variable X with cdf F has a *heavy-tail distribution* (to the right) if $1 - F(x)$ converges more slowly than e^{-ax} for any $a > 0$ when $x \to \infty$, or equivalently if its generating function $M(t)$ is infinite for any $t > 0$. For example, the Pareto distribution has a heavy-tail, whereas the Weibull distribution with shape parameter $\alpha < 1$ is heavy-tailed.

A.6 Conditional expectation

Let X be a random variable defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, let $\mathcal{G} \subset \mathcal{F}$ be a σ-field contained in $\mathcal{F}$ (also called a *sub-σ-field of $\mathcal{F}$*), and let $\mathbb{P}'$ be the restriction of $\mathbb{P}$ to $\mathcal{G}$. Assume $\mathbb{E}[X] < \infty$. The *conditional expectation of X given $\mathcal{G}$*, denoted $\mathbb{E}[X \mid \mathcal{G}]$, is *any* random variable on $(\Omega, \mathcal{G}, \mathbb{P}')$ that satisfies

$$\int_B \mathbb{E}[X \mid \mathcal{G}] \mathrm{d}\mathbb{P}' = \int_B X \mathrm{d}\mathbb{P}$$

for all $B \in \mathcal{G}$. There always exists at least one version of this random variable and any two versions are equal with probability 1. In other words, the conditional expectation is defined only w.p.1. The conditional expectation can be interpreted loosely as the expectation of X after we have the information about the occurrence (or not) of all the events in $\mathcal{G}$.

As special cases, if $\mathcal{G} = \mathcal{F}$, then $\mathbb{E}[X \mid \mathcal{G}] = \mathbb{E}[X \mid \mathcal{F}] = X$ w.p.1 because $\mathbb{P}' = \mathbb{P}$ (intuitively, X is $\mathcal{F}$-measurable so the information in $\mathcal{F}$ must tell us the value of X), whereas if $\mathcal{G}$ contains only Ω and the empty set, this $\mathcal{G}$ tells us nothing about the value of X and we have $\mathbb{E}[X \mid \mathcal{G}] = \mathbb{E}[X]$ (a constant).

Conditional probabilities can be seen as special cases of conditional expectations: For any event $B \in \mathcal{F}$, the *conditional probability of B given $\mathcal{G}$* is $\mathbb{P}[B \mid \mathcal{G}] = \mathbb{E}[\mathbb{I}[B] \mid \mathcal{G}]$.

If Y is a random variable defined on $(\Omega, \mathcal{F}, \mathbb{P})$, let $\mathcal{G}(Y)$ be the smallest σ-field $\mathcal{G}$ (necessarily contained in $\mathcal{F}$) with respect to which Y is $\mathcal{G}$-measurable. This $\mathcal{G}(Y)$ is called the *σ-field generated by Y*. The *conditional expectation of X given Y*, denoted $\mathbb{E}[X \mid Y]$, is defined as $\mathbb{E}[X \mid \mathcal{G}(Y)]$. We can also define the conditional probabilities $\mathbb{P}[X \in A \mid Y]$ as equal to $\mathbb{P}[X \in A \mid \mathcal{G}(Y)]$ for each $A \in \mathcal{F}$.

Example A.6 In Example A.2, the given σ-field $\mathcal{G}$ is $\mathcal{G}(Y)$ where Y is the sum on the two dice. If X is the value on the first dice, then $\mathbb{E}[X \mid Y]$ is a random variable that takes the value $Y/2$. □

Example A.7 Suppose you wait in a first-in first-out single-server queue. Let X_1 be the time until the customer in front of you starts its service, X_2 the service time of that customer, and suppose that these two random variables are continuous. Your total waiting time is $X = X_1 + X_2$. Suppose you observe X_1 and you are interested in the conditional probability that $X > c$ given that $X_1 = x_1$, for some constant c. This probability $\mathbb{P}[X > c \mid X_1 = x_1]$ cannot be

written as $\mathbb{P}[X > c, X_1 = x_1]/\mathbb{P}[X_1 = x_1]$, because $\mathbb{P}[X_1 = x_1] = 0$. However, we have $\mathbb{P}[X > c \mid X_1 = x_1] = \mathbb{P}[X_2 > c - x_1]$, which can be computed easily if we know the cdf of X_2. □

The next result follows directly by taking $B = \Omega$ in the definition.

Proposition A.4 (Unbiasedness of conditional expectation). *If $\mathcal{G} \subset \mathcal{F}$ is a σ-field and $\mathbb{E}[X] < \infty$, then*

$$\mathbb{E}[\mathbb{E}[X \mid \mathcal{G}]] = \mathbb{E}[X].$$

The following variance decomposition is used frequently in the book, in particular for the analysis of variance reduction methods.

Proposition A.5 (Variance decomposition). *For any σ-field $\mathcal{G} \subset \mathcal{F}$, we have*

$$\mathrm{Var}[X] = \mathrm{Var}[\mathbb{E}[X \mid \mathcal{G}]] + \mathbb{E}[\mathrm{Var}[X \mid \mathcal{G}]].$$

Proof.

$$\begin{aligned}
\mathrm{Var}[X] &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \\
&= \mathbb{E}[\mathbb{E}[X^2 \mid \mathcal{G}]] - (\mathbb{E}[\mathbb{E}[X \mid \mathcal{G}]])^2 \\
&= \mathbb{E}[\mathrm{Var}[X \mid \mathcal{G}]] + \mathbb{E}[(\mathbb{E}[X \mid \mathcal{G}])^2] - (\mathbb{E}[\mathbb{E}[X \mid \mathcal{G}]])^2 \\
&= \mathbb{E}[\mathrm{Var}[X \mid \mathcal{G}] + \big(\mathbb{E}[X \mid \mathcal{G}]\big)^2] - \big(\mathbb{E}[\mathbb{E}[X \mid \mathcal{G}]]\big)^2 \\
&= \mathbb{E}[\mathrm{Var}[X \mid \mathcal{G}]] + \mathrm{Var}[\mathbb{E}[X \mid \mathcal{G}]].
\end{aligned}$$

♣ Stochastic order. See Wolff 1989.

A.7 Joint distribution and independence

The notions of mass and density functions, expectation, variance, etc., can be generalized to *vectors* or random variables, also called *random vectors*.

Let $X_1, \ldots, X_d$ be d random variables defined on the same probability space and let $\mathbf{X} = (X_1, \ldots, X_d)^\mathsf{t}$, where the t means "transposed" (our vectors are *column* vectors). The vector $\mathbf{X}$ is said to have a d-dimensional *multivariate distribution*, with *joint cdf* $F : \mathbb{R}^d \to [0, 1]$, if for any $\mathbf{x} = (x_1, \ldots, x_d)^\mathsf{t} \in \mathbb{R}^d$,

$$F(\mathbf{x}) = \mathbb{P}[\mathbf{X} \leq \mathbf{x}] = \mathbb{P}[X_1 \leq x_1, \ldots, X_d \leq x_d].$$

The jth *marginal* cdf of F is F_j, defined by $F_j(x_j) = F(\infty, \ldots, \infty, x_j, \infty, \ldots, \infty) = \mathbb{P}[X_j \leq x_j]$.

The random variables $X_1, \ldots, X_d$ are *mutually independent* if and only if $F(x_1, \ldots, x_d) = F_1(x_1) \cdots F_d(x_d)$ for all $(x_1, \ldots, x_d) \in \mathbb{R}^d$. They are *pairwise independent* (a weaker condition) if and only if $F(x_i, x_j) = F_i(x_i)F_j(x_j)$ for all pairs $i \neq j$. If $X_1, \ldots, X_d$ are mutually independent, then their sum $Y = X_1 + \cdots + X_d$ has moment-generating function $M_Y(\theta) = M_{X_1}(\theta) \cdots M_{X_d}(\theta)$ and cumulant generating function $\Psi_Y(\theta) = \Psi_{X_1}(\theta) + \cdots + \Psi_{X_d}(\theta)$. Moreover, for any measurable functions $g_1, \ldots, g_d$, where $g_j : \mathbb{R} \to \mathbb{R}$,

$$\mathbb{E}[g_1(X_1) \cdots g_d(X_d)] = \mathbb{E}[g_1(X_1)] \cdots \mathbb{E}[g_d(X_d)].$$

If $X_1, \ldots, X_d$ are *discrete*, their *joint probability mass function* p is defined by

$$p(x_1, \ldots, x_d) = \mathbb{P}[X_1 = x_1, \ldots, X_d = x_d]$$

for all values of $x_1, \ldots, x_d$ that these random variables can take. The *marginal* mass function of X_j is defined by

$$p_j(x_j) = \mathbb{P}[X_j = x_j].$$

One can easily prove that the discrete random variables $X_1, \ldots, X_d$ are mutually *independent* if and only if

$$p(x_1, \ldots, x_d) = p_1(x_1) \cdots p_d(x_d)$$

for all values $x_1, \ldots, x_d$.

If $X_1, \ldots, X_d$ are *continuous*, their *joint density function* f is defined via

$$F(x_1, \ldots, x_d) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_d} f(y_1, \ldots, y_d) \mathrm{d}y_1 \cdots \mathrm{d}y_d.$$

The *marginal* density of X_j is f_j, defined by

$$F_j(x) = \int_{-\infty}^{x} f_j(y) \mathrm{d}y.$$

The continuous random variables $X_1, \ldots, X_d$ are mutually *independent* if and only if

$$f(x_1, \ldots, x_d) = f_1(x_1) \cdots f_d(x_d)$$

for all $(x_1, \ldots, x_d) \in \mathbb{R}^d$. This also holds if and only if the characteristic function of $\mathbf{X}$ is the product of the characteristic functions of the X_j's.

Let $Z = g(X_1, \ldots, X_d)$ where $g : \mathbb{R}^d \to \mathbb{R}$ is a measurable function, so that Z is a random variable. If $X_1, \ldots, X_d$ are continuous random variables, then

$$\mathbb{E}[Z] = \mathbb{E}[g(X_1, \ldots, X_d)] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, \ldots, x_d) f(x_1, \ldots, x_d) \mathrm{d}x_1 \cdots \mathrm{d}x_d.$$

This also holds in the discrete case if we replace the integrals by sums.

Theorem A.6 (Jensen's inequality). *If $\mathbf{X}$ is a random vector in $\mathbb{R}^d$ with finite expectation and $h : \mathbb{R}^d \to \mathbb{R}$ is a convex measurable function, then*

$$\mathbb{E}[h(\mathbf{X})] \geq h(\mathbb{E}[\mathbf{X}]).$$

A.8 Covariance and correlation

The *covariance* between two random variables X and Y is defined by

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

If X and Y are independent, then $\mathrm{Cov}(X, Y) = 0$, but the converse is not necessarily true.

The (Pearson) linear *correlation coefficient* between X and Y is

$$\rho(X, Y) = \mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{(\mathrm{Var}(X)\mathrm{Var}(Y))^{1/2}} = \frac{\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]}{\sigma(X)\sigma(Y)}. \quad \text{(A.3)}$$

It always satisfies $-1 \leq \rho(X, Y) \leq 1$ and can be viewed as a standardized version of the covariance. It measures the *linear dependence* between X and Y. We say that X and Y are *uncorrelated* if $\rho(X, Y) = 0$, *positively corelated* if $\rho(X, Y) > 0$, and *negatively correlated* if $\rho(X, Y) < 0$.

The *covariance matrix* $\boldsymbol{\Sigma} = \mathrm{Cov}[\mathbf{X}]$ of a random vector $\mathbf{X} = (X_1, \ldots, X_d)^{\mathsf{t}}$ is the matrix whose elements are $\sigma_{ij} = \mathrm{Cov}[X_i, X_j]$, and the *correlation matrix* $\mathbf{R}$ is the one whose elements are $\rho(X_i, X_j)$. We can write $\mathrm{Cov}[\mathbf{X}] = \mathbb{E}[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^{\mathsf{t}}]$ where $\boldsymbol{\mu} = \mathbb{E}[\mathbf{X}] = (\mathbb{E}[X_1], \ldots, \mathbb{E}[X_d])^{\mathsf{t}}$. Any covariance matrix must be symmetric and nonnegative definite, because for any vector $\mathbf{a} \in \mathbb{R}^d$, $\mathbf{a}^{\mathsf{t}}\boldsymbol{\Sigma}\mathbf{a} = \mathrm{Cov}(\mathbf{a}^{\mathsf{t}}\mathbf{X}) \geq 0$. Any correlation matrix must have all its diagonal elements equal to 1. Moreover, if $\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{b}$ where $\mathbf{A}$ is a $d \times d$ matrix and $\mathbf{b}$ is a d-dimensional vector, then $\mathbb{E}[\mathbf{Y}] = \mathbf{A}\boldsymbol{\mu} + \mathbf{b}$ and

$$\mathrm{Cov}[\mathbf{Y}] = \mathbb{E}[\mathbf{A}(\mathbf{X} - \boldsymbol{\mu})(\mathbf{A}(\mathbf{X} - \boldsymbol{\mu}))^{\mathsf{t}}] = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^{\mathsf{t}}. \quad \text{(A.4)}$$

As a special case, by taking $\mathbf{A}$ as the diagonal matrix with diagonal elements $c_1, \ldots c_d$, we obtain:

Proposition A.7 *If $X_1, \ldots, X_d$ are arbitrary random variables and $c_1, \ldots, c_d$ are constants, then*

$$\mathbb{E}[c_1X_1 + \cdots + c_dX_d] = c_1\mathbb{E}[X_1] + \cdots + c_d\mathbb{E}[X_d]$$

and

$$\begin{aligned} \mathrm{Var}[c_1X_1 + \cdots + c_dX_d] &= \sum_{i=1}^{d}\sum_{j=1}^{d} c_ic_j\mathrm{Cov}(X_i, X_j) \\ &= \sum_{i=1}^{d} c_i^2\mathrm{Var}(X_i) + 2\sum_{i=1}^{d}\sum_{j=i+1}^{d} c_ic_j\mathrm{Cov}(X_i, X_j). \end{aligned}$$

A.9 Change of Variables

A change of variables is often very convenient to transform an integral or a probability distribution to an equivalent one that is much easier to handle. We define it in the d-dimensional real space, but it also works more generally. Let $\varphi : \mathbb{R}^d \to \mathbb{R}^d$ be a one-to-one differentiable mapping having a continuous inverse, and let B a measurable subset of $\mathbb{R}^d$. Then one can write

$$\int_B f(\mathbf{x})\mathrm{d}\mathbf{x} = \int_{\varphi^{-1}(B)} f(\varphi(\mathbf{y}))|\det(J(\mathbf{y}))|\mathrm{d}\mathbf{y}$$

where $J(\mathbf{y})$ is the *Jacobian* of the transformation, defined as the $d \times d$ matrix whose element (i, j) contains the derivative of the ith coordinate of $\mathbf{x} = \varphi(\mathbf{y})$ with respect to the jth coordinate of $\mathbf{y}$.

If $\mathbf{Y}$ is a continuous random vector in $\mathbb{R}^d$, the random vector $\mathbf{X} = \varphi(\mathbf{Y})$ has density f if and only if $\mathbf{Y}$ has density

$$g(\mathbf{y}) = f(\varphi(\mathbf{y}))|\det(J(\mathbf{y}))|.$$

As a special case, if φ defines a linear transformation via $\varphi(\mathbf{y}) = \mathbf{A}\mathbf{y}$ for some matrix $\mathbf{A}$, then $J(\mathbf{y}) = \mathbf{A}$, so $|\det(J(\mathbf{y}))| = |\det(\mathbf{A})|$, a constant.

A.10 Convergence of random variables

Let $X_1, X_2, \ldots$ be an infinite sequence of random variables and X another random variable (or a constant). Here, all random variables are assumed to be defined on the same probability space. Saying that this sequence converges to X can have different meanings, which are not equivalent. We point out some of them.

We say that *$X_n \to X$ with probability 1* (or *almost surely*), denoted $X_n \overset{\text{w.p.1}}{\to} X$ or $X_n \to X$ w.p.1, if

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

We say that *$X_n \to X$ in probability*, denoted $X_n \overset{\text{p}}{\to} X$, if for all $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \epsilon) = 0.$$

We say that *$X_n \to X$ in distribution*, denoted $X_n \Rightarrow X$, if

$$\lim_{n\to\infty} F_n(x) = F(x).$$

at all points $x \in \mathbb{R}$ where F is continuous, where F and F_n are the cdf's of X and X_n. Convergence in distribution is also called *weak convergence*. For $1 \le p < \infty$, we say that *$X_n \to X$ in the $\mathcal{L}_p$ norm* if $\mathbb{E}[|X|^p] < \infty$, $\mathbb{E}[|X_n|^p] < \infty$ for each n, and

$$\lim_{n\to\infty} \mathbb{E}[|X_n - X|^p] = 0.$$

When $p = 2$, this is called *convergence in quadratic mean*, or *mean-square convergence*. If $X_n \to X$ *in the* $\mathcal{L}_p$ *norm* then $X_n \to X$ *in the* $\mathcal{L}_q$ *norm* for all $q \leq p$.

Each of these modes of convergence extends directly to random vectors: convergence of vectors is equivalent to convergence of all components of the vector.

The next proposition summarizes some well-established relationships between these modes of convergence. They are illustrated by the diagram of Figure A.3.

Proposition A.8 *Convergence with probability 1 implies convergence in probability, which implies convergence in distribution (the weakest form among those mentioned here). Convergence in the $\mathcal{L}_p$ norm for some $p \geq 1$ also implies convergence in probability. However, convergence with probability 1 and convergence in the $\mathcal{L}_p$ norm do not imply each other in any way.*

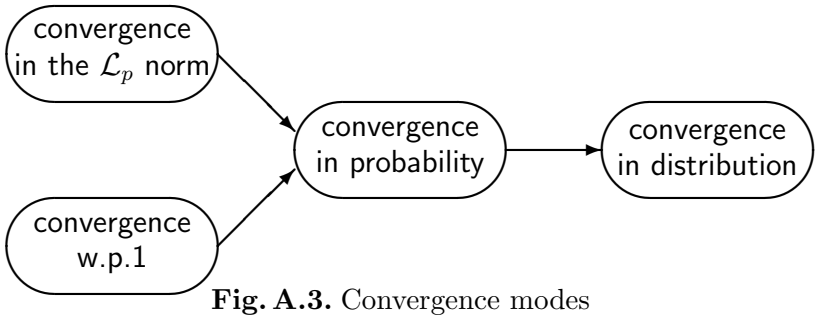

Fig. A.3. Convergence modes

For a more detailed treatment and several additional results, see, e.g., Galambos (1995). For example, convergence in probability implies the existence of a subsequence of $\{X_n\}$ that converges w.p.1, and with some additional conditions it implies convergence in the $\mathcal{L}_p$ norm (see Proposition A.13).

Theorem A.9 (Weak convergence criterion). *We have $X_n \Rightarrow X$ if and only if*

$$\lim_{n\to\infty} \mathbb{E}[g(X_n)] = \mathbb{E}[g(X)]$$

for any bounded continuous function $g : \mathbb{R} \to \mathbb{R}$, if and only if

$$\lim_{n\to\infty} \varphi_{X_n}(\theta) = \varphi_X(\theta),$$

where φ_X denotes the characteristic function of X.

Theorem A.10 (Continuous mapping theorem). *If a sequence of random vectors* $\{\mathbf{X}_n = (X_{n,1}, \ldots, X_{n,d}), n \geq 1\}$ *converges in distribution to a random vector* $\mathbf{X} = (X_1, \ldots, X_d)$ *when* $n \to \infty$, *then for any function* $g : \mathbb{R}^d \to \mathbb{R}$ *such that the probability that* X *falls exactly on a discontinuity of* g *is zero, we have* $g(\mathbf{X}_n) \Rightarrow g(\mathbf{X})$.

By taking $g(x, y, z) = yx + z$ in the previous result, we obtain:

Theorem A.11 (Slutsky's theorem). *If* $X_n \Rightarrow X$, $Y_n \Rightarrow a$, *and* $Z_n \Rightarrow b$ *for some constant* a *and* b, *and all these random variables are defined on the same probability space, then* $Y_n X_n + Z_n \Rightarrow aX + b$.

Theorem A.12 (Anscombe's theorem). *Suppose* $X_n \Rightarrow X$ *and* $N_n \overset{\text{w.p.1}}{\to} \infty$, *where* N_n *is a positive integer and* $N_{n+1} \geq N_n$ *for all* n. *If* $N_n/n \Rightarrow c$ *for some constant* $c > 0$, *then* $X_{N_n} \Rightarrow X$. *The result also holds if the condition "*$N_n/n \Rightarrow c$*" is replaced by: "*$\{X_n, n \geq 0\}$ *and* $\{N_n, n \geq 0\}$ *are independent sequences."*

Definition A.1 A family of random variables $\{Y_i, i \in I\}$, where I is an arbitrary set, is *uniformly integrable* if

$$\lim_{k\to\infty} \sup_{i\in I} \mathbb{E}\left[|Y_i| \, \mathbb{I}[|Y_i| > k]\right] = 0.$$

A sufficient condition for uniform integrability is that $\sup_{i\in I} \mathbb{E}\left[|Y_i|^{1+\delta}\right] < \infty$ for some constant $\delta > 0$. □

Proposition A.13 *Suppose* $X_n \Rightarrow X$. *Then* $X_n \to X$ in the $\mathcal{L}_p$ norm *if and only if* $\{|X_n - X|^p, n \geq 0\}$ *is uniformly integrable.*

A.11 Convergence of probability measures

Convergence of sequences of random variables is a special case of the notion of convergence of sequences of measures in general; see, e.g., Billingsley (1968). This general theory can provide more general central limit theorems, among other things. We recall the following concept:

Definition A.2 The *total variation distance* between two probability measures $\mathbb{P}$ and $\mathbb{Q}$ on $(\Omega, \mathcal{F})$ is defined as

$$d(\mathbb{P}, \mathbb{Q}) = \sup_{B\in\mathcal{F}} |\mathbb{Q}(B) - \mathbb{P}(B)| \, .$$

A sequence of measure $\{\mathbb{P}_n, n \geq 0\}$ *converges in total variation* to a measure $\mathbb{P}$ if $\lim_{n\to\infty} d(\mathbb{P}_n, \mathbb{P}) = 0$. □

A.12 Point Estimation

Suppose we want to estimate an unknown quantity ν by some random variable X with expectation $\mu = \mathbb{E}[X]$ and variance $\sigma^2 = \mathrm{Var}[X]$. The difference $\beta = \mu - \nu$ is called the *bias* of the estimator X. If $\beta = 0$, the estimator is *unbiased*.

If $X_1, \ldots, X_n$ are independent realizations of X, then $X_1, \ldots, X_n$ is called a *sample* of *independent and identically distributed* random variables (or an i.i.d. *sample*). We define the *sample mean* (or *empirical mean*) by

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

and the *sample variance* by

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \frac{1}{n-1}\sum_{i=1}^{n} X_i^2 - \frac{n(\bar{X}_n)^2}{n-1}.$$

It is easy to verify that

$$\mathbb{E}[\bar{X}_n] = \mu \quad \text{and} \quad \mathbb{E}[S_n^2] = \sigma^2,$$

so $\bar{X}_n$ and S_n^2 are unbiased estimators of μ and σ^2.

The variance of $\bar{X}_n$ is σ^2/n and it can be estimated by S_n^2/n. If $(X_1, Y_1), \ldots, (X_n, Y_n)$ is an i.i.d. sample of (X, Y), then an unbiased estimator of $\mathrm{Cov}[X, Y]$ is given by

$$\widehat{\mathrm{Cov}}[X, Y] = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)(Y_i - \bar{Y}_n).$$

An infinite sequence of estimators $\{Y_n, n \geq 1\}$ is often denoted simply by Y_n. For example, we sometimes use $\bar{X}_n$ and S_n^2 to denote infinite sequences indexed by n. With this abuse of notation, when $n \to \infty$, we say that Y_n is *asymptotically unbiased* if $\mathbb{E}[Y_n - \mu] \to 0$, *consistent* if $Y_n \to \mu$ in probability, and *strongly consistent* if $Y_n \overset{\text{w.p.1}}{\to} \mu$.

Example A.8 If $\mu = \mathbb{E}[X_i]$ and $\sigma^2 = \mathrm{Var}[X_i] < \infty$, then $\bar{X}_n$ is strongly consistent for μ and S_n^2 is strongly consistent for σ^2. □

The two theorems that follow are key results in probability theory. The first one says that the average of i.i.d. random variables converges strongly to their expectation when the latter is finite, but tells nothing about the convergence speed. The second result provides information about the convergence speed. Proofs can be found in Billingsley (1986), pages 290 and 367.

Theorem A.14 (Strong law of large numbers). *If* $X_1, X_2, \ldots$ *are* i.i.d. *with* $\mathbb{E}[X_i] = \mu$ *finite, then* $\bar{X}_n \overset{\text{w.p.1}}{\to} \mu$ *when* $n \to \infty$.

Theorem A.15 (Central-limit theorem (CLT)). *If* $X_1, X_2, \ldots$ *are* i.i.d., $\mathbb{E}[X_i] = \mu$, *and* $\mathrm{Var}[X_i] = \sigma^2 < \infty$, *then*

$$\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \;\Rightarrow\; \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \;\Rightarrow\; Z \sim N(0,1) \qquad \text{when } n \to \infty,$$

i.e., for all $x \in \mathbb{R}$,

$$\mathbb{P}\left(\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \le x\right) \to \Phi(x)$$

where Φ *is the cdf of a* $N(0,1)$ *random variable.*

More generally, if $\mathbf{X}_1, \mathbf{X}_2, \ldots$ *are* i.i.d. *random vectors with* $\mathbb{E}[\mathbf{X}_i] = \boldsymbol{\mu}$ *and* $\mathrm{Cov}[\mathbf{X}_i] = \boldsymbol{\Sigma}$ *(finite), then*

$$\sqrt{n}(\bar{\mathbf{X}}_n - \boldsymbol{\mu}) \;\Rightarrow\; \mathbf{Z} \sim N(\mathbf{0}, \boldsymbol{\Sigma}),$$

the multivariate normal distribution with mean $\mathbf{0}$ *and covariance matrix* $\boldsymbol{\Sigma}$.

Theorem A.15 guarantees convergence of the standardized average to the normal distribution, but how fast does this convergence occurs? The following version of the Berry-Esseen theorem, proved in Katz (1963), bounds the error made when we approximate the cdf F_n of the Student statistic

$$T_n = \sqrt{n}(\bar{X}_n - \mu)/S_n,$$

defined by $F_n(x) = \mathbb{P}[T_n \le x]$, by the standard normal cdf Φ. It shows that this error converges uniformly as $O(n^{-1/2})$. The result also holds if we replace S_n by σ.

Theorem A.16 (Berry-Esseen inequality). *Under the assumptions of Theorem A.15, if* $\mathbb{E}[|X_i - \mu|^3] = \beta_3$, *there is a constant* $c \le 0.7056$ *such that*

$$\sup_{x \in \mathbb{R},\, n \ge 2} \sqrt{n}|F_n(x) - \Phi(x)| \le c\,\frac{\beta_3}{\sigma^3}.$$

This result can be generalized in many ways, for example to the case where the X_i's are not identically distributed; then, σ^2 and β_3 in the bound must be replaced by the average variance and the average third absolute centered moment, and $c \le 6$ (Feller 1971).

A.13 Confidence intervals

The accuracy of $\bar{X}_n$ as an estimator of μ is often assessed via a *confidence interval* (CI) for μ. A random interval $[I_1, I_2]$ is a CI at (confidence) *level* $1-\alpha$ (or a $100(1-\alpha)\%$ CI) for μ if $\mathbb{P}[I_1 \le \mu \le I_2] = 1 - \alpha$. The boundaries I_1 and I_2 of the CI, and its *width* $I_2 - I_1$, are random variables.

Often, the CI is constructed for a given *target* or *nominal* level $1-\alpha$, but its true *coverage probability* $\mathbb{P}[I_1 \le \mu \le I_2]$ may differ from $1-\alpha$ and is often unknown. The difference $\mathbb{P}[I_1 \le \mu \le I_2] - (1-\alpha)$ is the *coverage error*.

Ideally, a good CI should have (a) (approximately) the correct coverage and (b) small values of $\mathbb{E}[I_2 - I_1]$ and $\mathrm{Var}[I_2 - I_1]$.

A CI $(I_{n,1},\, I_{n,2})$ that depends on the sample size n is *asymptotically valid* if its coverage error converges to 0 when $n \to \infty$.

For a fixed confidence level $1-\alpha$, the CLT tells us that for n large enough, if $z_{1-\alpha/2} = \Phi^{-1}(1-\alpha/2)$ (i.e., $\Phi(z_{1-\alpha/2}) = 1-\alpha/2$), then

$$\mathbb{P}[|\bar{X}_n - \mu| \le z_{1-\alpha/2} S_n/\sqrt{n}] \approx 1-\alpha.$$

In this case, a CI at (approximate) level $1-\alpha$ is given by

$$[I_{n,1},\, I_{n,2}] = [\bar{X}_n - z_{1-\alpha/2} S_n/\sqrt{n},\ \bar{X}_n + z_{1-\alpha/2} S_n/\sqrt{n}].$$

For example, for $\alpha = 0.05$, we have $z_{1-\alpha/2} \approx 1.96$.

When $n \to \infty$, the width of the CI is asymptotically proportional to $\sigma/\sqrt{n}$, so it converges as $O(n^{-1/2})$.

When n is small, the central-limit theorem cannot be invoked, but if the X_i's are i.i.d. normal, then we can use the following:

Theorem A.17 *(Convergence to Student and chi-square distributions). If $X_1, \ldots, X_n$ are* i.i.d. *$N(\mu, \sigma^2)$, then*
(i) $\bar{X}_n$ and S_n^2 are independent;
(ii) $(n-1)S_n^2/\sigma^2$ has the chi-square distribution with $n-1$ degrees of freedom;
(iii) $\sqrt{n}(\bar{X}_n - \mu)/S_n$ has the Student-t distribution with $n-1$ degrees of freedom.

Part (iii) of this theorem permits one to compute a CI for μ at (exact) level $1-\alpha$:

$$[\bar{X}_n - t_{n-1,1-\alpha/2} S_n/\sqrt{n},\ \bar{X}_n + t_{n-1,1-\alpha/2} S_n/\sqrt{n}]$$

where $\mathbb{P}[T_{n-1} \le t_{n-1,1-\alpha/2}] = 1-\alpha/2$ and T_{n-1} has the Student distribution with $(n-1)$ degrees of freedom, which is approximately $N(0,1)$ when n is large. Part (ii) can be used to compute a CI for σ^2: select x_1 and x_2 such that

$$\mathbb{P}[x_1 < \chi^2_{n-1} < x_2] = 1-\alpha,$$

and put

$$[I_{n,1}, I_{n,2}] = [(n-1)S_n^2/x_2,\ (n-1)S_n^2/x_1].$$

We have

$$\begin{aligned}
\mathbb{P}[I_{n,1} \le \sigma^2 \le I_{n,2}] &= \mathbb{P}[(n-1)S_n^2/x_2 \le \sigma^2 \le (n-1)S_n^2/x_1] \\
&= \mathbb{P}[x_1 \le (n-1)S_n^2/\sigma^2 \le x_2] \\
&= 1-\alpha.
\end{aligned}$$

It is important to recall that all of this is valid only if the X_i's have the normal distribution. Otherwise, these intervals can be used as approximations if the X_i's are independent and have a distribution that is not too far from the normal (in particular, it should not be too asymmetric).

A.14 Large deviations

♣ To be done.

A.15 Statistical tests of hypotheses

In a statistical test of hypothesis, one selects a *null hypothesis* $\mathcal{H}_0$, which could be any hypothesis made about the stochastic model, and a random variable Y, called the *test statistic*, whose distribution is known (or can be well approximated) when $\mathcal{H}_0$ is true. Then, a realization of Y is obtained and we judge if it seems reasonable that this value of Y was generated from the theoretical distribution of Y under $\mathcal{H}_0$. If not, i.e., if the value is much too large or much too small, then we say that we *reject* $\mathcal{H}_0$.

Suppose the realization of Y is y. We then define the *p-value* of the test as $p = \mathbb{P}[Y \geq y \mid \mathcal{H}_0]$. This p-value is in fact a random variable whose outcome is revealed when we perform the experiment for the test. If Y has a continuous distribution, then the p-value is uniformly distributed in the interval $(0, 1)$ under $\mathcal{H}_0$. A popular practice is to select a priori a *significance level* $\alpha > 0$ (e.g., 0.05 or 0.01) and reject $\mathcal{H}_0$ when $p < \alpha$ (for a single-sided test) or when p is outside the interval $[\alpha/2, 1 - \alpha/2]$ (for a two-sided test). The *power of the test* is then the probability β of rejecting $\mathcal{H}_0$ when it is false; this probability depends on what is actually true. The probability of rejecting $\mathcal{H}_0$ when it is true is α. Usually, $\beta < 1$ and $\alpha > 0$, so there can be a wrong decision in either direction.

Perhaps a better approach is to avoid fixing a threshold α a priori and simply report the p-value. This provides more information. In any case, whatever is the p-value, a statistical test *never proves* that $\mathcal{H}_0$ is true or false. It only provides statistical evidence against it, or for it. On the other hand, the evidence against $\mathcal{H}_0$ is sometimes very strong, for example if the p-value is smaller than 10^{-15}, which happens frequently when applying statistical tests to random number generators.

Example A.9 Suppose we throw n balls randomly and independently in r boxes, where each ball has probability $\tilde{p}_j$ of falling in box j, for $j = 1, \ldots, r$. We are not sure about the $\tilde{p}_j$'s, but we have reasons to believe that $\tilde{p}_j = p_j$ for each j, where the p_j's are fixed, and want to test this hypothesis. Let X_j be the number of balls falling in box j in our experiment and let $o_j = np_j = \mathbb{E}[X_j \mid \mathcal{H}_0]$. The *chi-square test statistic* in this context is

$$Y = \sum_{j=1}^{r} \frac{(X_j - o_j)^2}{o_j}.$$

Its distribution under $\mathcal{H}_0$ is approximately chi-square with $r - 1$ degrees of freedom if o_j is large enough (e.g., > 5) for all j (see, e.g., Read and Cressie 1988). So if Y takes the value y, we can compute the p-value (approximately)

by $p = 1 - F(y)$ where F is the distribution function of a chi-square random variable with $r - 1$ degrees of freedom. □

A.16 Stochastic processes

A *stochastic process* is a family $\{Y_t, t \in I\}$ of random variables (or vectors) defined on the same probability space. The index is often interpreted as the time. The process is *continuous-time* if I is continuous (e.g., $I = [0, \infty)$), and *discrete-time* if I is discrete (e.g., $I = \{0, 1, 2, \ldots\}$). When I is continuous, we often denote Y_t by $Y(t)$.

A stochastic process is *Markovian* if, conditionally on its present value Y_t at time t, its future is independent of its past. More precisely, this means that for any random variable X that can be written as a function of $\{Y_s,\ s > t\}$, the distribution of X conditional on $\{Y_s,\ s \le t\}$ is the same as that conditional on Y_t. In other words, the process is Markovian if Y_t contains enough information to generate its future without looking further in the past. A process can always be made Markovian by putting enough information in the state Y_t. Of course, this would generally enlarge the size of the state space. When the index set I is the set of non-negative integers $\{0, 1, \ldots\}$, the Markov process is called a (discrete-time) *Markov chain*.

For a stochastic process $\{Y_j,\ j = 0, 1, \ldots\}$, *$\sigma$-field generated by* $Y_0, \ldots, Y_t$, denoted $\mathcal{F}_t = \sigma(Y_0, \ldots, Y_t)$, represents all the information that can be deduced by observing the trajectory of the process up to step t. Similarly, for a continuous-time process $\{Y_t,\ t \ge 0\}$, the *σ-field $\mathcal{F}_t$ generated by* $\{Y_s,\ 0 \le s \le t\}$ represents all the information that can be deduced from the observation of $\{Y_s,\ 0 \le s \le t\}$. In both cases, the family $\{\mathcal{F}_t,\ t \ge 0\}$ is called a *filtration*. It can be viewed as a filter that leaks information on the sample path as time goes on. A random variable is *$\mathcal{F}_t$-measurable* if its value can always be computed from the information contained in $\mathcal{F}_t$ (without looking in the future, in particular).

A random variable T (discrete or continuous) is a *stopping time* with respect to the filtration $\{\mathcal{F}_t,\ t \ge 0\}$ if T is $\mathcal{F}_T$-measurable, i.e., if the value of T is known by the time it is reached. For a rigorous mathematical coverage of these topics, see, e.g., Billingsley (1986).

A.17 Renewal processes

Let $Y_1, Y_2, \ldots$ be a sequence of i.i.d. random variables, $S_0 = 0$, and $S_n = Y_1 + \cdots + Y_n$ for $n \ge 1$. We can interpret the S_n's as renewal epochs for a continuous-time process, and the Y_j's as times between renewals, as follows. Let $N(t) = \max\{n \ge 0 \mid S_n \le t\}$ denote the number of renewals during the time interval $[0, t]$. The process $\{N(t),\ t \ge 0\}$ is called a *renewal process*. As a special case, if each Y_j is exponential with mean $1/\lambda$, the renewal process is a stationary

Poisson process with rate λ. In that case, for all t, $N(t)$ is a Poisson random variable with mean λt.

At a given epoch t, the times of the last renewal and of the next renewal are $S_{N(t)}$ and $S_{N(t)+1}$, respectively. Call the time interval $(S_{j-1}, S_j]$ the jth renewal cycle and suppose a cost X_j is incurred during that cycle, where the X_j's are i.i.d. and X_j is independent of $\{Y_\ell, \ell \neq j\}$. The total cost for the first n cycles is $V_n = \sum_{j=1}^{n} X_j$, and the total cost until time t is $V(t) = V_{N(t)}$. Let $v(t) = \mathbb{E}[V(t)]$. (The costs may also be interpreted as rewards instead.)

Theorem A.18 (Renewal reward theorem.) *Suppose that* $0 < \mathbb{E}[Y_j] < \infty$ *and* $\mathbb{E}[|X_j|] < \infty$. *Then,*

$$\bar{v} \stackrel{\text{def}}{=} \lim_{t\to\infty} \frac{\mathbb{E}[V_{N(t)}]}{t} = \lim_{t\to\infty} \frac{\mathbb{E}[V_{N(t)+1}]}{t} = \frac{\mathbb{E}[X_j]}{\mathbb{E}[Y_j]}; \tag{A.5}$$

and

$$\bar{v} \stackrel{\text{w.p.1}}{=} \lim_{t\to\infty} \frac{V_{N(t)}}{t} = \lim_{t\to\infty} \frac{V_{N(t)+1}}{t}. \tag{A.6}$$

A proof can be found in Wolff (1989). Eq. (A.5) is called the *expected value version* of the renewal reward theorem, while (A.6) is the *sample path version*. Note that all expressions in (A.5) are deterministic, whereas the expressions in (A.6) (except for $\bar{v}$) are random variables. The special case of this theorem where $X_j = 1$ for all j is the *elementary renewal theorem*. The above analysis and the theorem also hold if Y_1 and X_1 have a different distribution than the other Y_j's and X_j's. This occurs when the process does not start at the beginning of a renewal cycle.

Theorem A.19 (Wald identity.) *Let* $X_1, X_2, \ldots$ *be independent random variables with common mean* $\mathbb{E}[X_j] = \mathbb{E}[X_1]$, *suppose* $\sup_{j\geq 1} \mathbb{E}[|X_j|] < \infty$, *and let* N *be a stopping time with respect to filtration generated by* $\{X_j, j \geq 1\}$ *(this means that* $\{N \geq n\}$ *is independent of* $(X_n, X_{n+1}, \ldots)$ *and* $\mathbb{P}[N < \infty] = 1$*) and such that* $\mathbb{E}[N] < \infty$. *Then*

$$\mathbb{E}\left[\sum_{j=1}^{N} X_j\right] = \mathbb{E}[X_1]\,\mathbb{E}[N].$$

More general versions of this theorem allow dependence between the X_i's and N does not necessarily have to be a stopping time, but some other technical conditions are needed.

A.18 Markov chains

A.18.1 Discrete-time Markov chains

Discrete-time Markov chains (DTMC) are an important class of Markovian processes for which the time index is $I = \{0, 1, 2, \ldots\}$. We consider the case of a *stationary* (or *time-homogeneous*) DTMC $\{Y_n,\, n \geq 0\}$ with *denumerable* state space $\mathcal{Y} = \{0, 1, 2, \ldots\}$ (we can just enumerate the states that way). The (one-step) *transition probabilities* are

$$p_{i,j} = \mathbb{P}[Y_{n+1} = j \mid Y_n = i]$$

for all $i, j \in \mathcal{Y}$ (they do not depend on n) and the k-step transition probabilities are defined as

$$p_{i,j}^{(k)} = \mathbb{P}[Y_{n+k} = j \mid Y_n = i].$$

The *transition probability matrix* is the matrix $\mathbf{P}$ whose element (i, j) is $p_{i,j}$. One can show that the matrix whose element (i, j) is $p_{i,j}^{(k)}$ is $\mathbf{P}^k$, the matrix $\mathbf{P}$ raised to the power k. Let $p_j = \mathbb{P}[Y_0 = j]$ (the initial-state probabilities) and $p_j^{(k)} = \mathbb{P}[Y_k = j]$. Define the row vectors $\mathbf{p} = (p_0, p_1, \ldots)$ and $\mathbf{p}^{(k)} = (p_0^{(k)}, p_1^{(k)}, \ldots)$. Then we have $\mathbf{p}^{(k)} = \mathbf{p}\mathbf{P}^k$ (here we use row vectors instead of column vectors for convenience and compatibility with standard notation).

Two state i and j are said to *communicate* if $p_{i,j}^{(k)} > 0$ for some $k > 0$ and $p_{j,i}^{(\ell)} > 0$ for some $\ell > 0$. The DTMC is called *irreducible* if all states communicate. A state j is said to be *recurrent* if when we are in that state, the probability that we return to it is 1, and *positive recurrent* if the expected number of steps required for this return, denoted ν_j, is finite. For an irreducible chain, if one state is positive recurrent, then all states are positive recurrent.

Theorem A.20 (Steady-state probabilities). *If the DTMC is irreducible, then all states are positive recurrent if and only if there is a unique probability (row) vector $\boldsymbol{\pi} = (\pi_0, \pi_1, \ldots)$ (whose elements are all non-negative and sum to 1) that satisfies the linear system*

$$\boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi}. \tag{A.7}$$

Moreover, $\pi_j = 1/\nu_j$ for all j.

We now assume that the DTMC is irreducible and positive recurrent. The vector $\boldsymbol{\pi}$ is called the vector of steady-state probabilities (or the stationary distribution). If we select the initial state Y_0 with the initial probabilities $p_j^{(0)} = \mathbb{P}[Y_0 = j] = \pi_j$ for all j, then we have $p_j^{(k)} = \mathbb{P}[Y_k = j] = \pi_j$ for all k an j, and π_j represents the fraction of the steps that we spend in state j, in the long run.

If we partition the state space in two sets, say B and $\bar{B}$, the fraction of the transitions that go from B to $\bar{B}$ must be the same as that going from $\bar{B}$ to B, in the long run. In particular, the fraction of transitions leaving any given state must be the same as that going into that state. However, if we pick any two

states i and j, there could be much more transitions going from i to j than from j to i, or vice-versa. When the two are equal in the long run, that is, when we have

$$\pi_i p_{i,j} = \pi_j p_{j,i} \tag{A.8}$$

for all pairs of states (i, j), we say that the chain is *reversible*. Such a chain can be simulated backward while keeping the same probability distribution over sample paths, if we initialize the state from the steady-state distribution and look at the sample paths in reverse when simulating backward. Any probability vector $\boldsymbol{\pi}$ on $\mathcal{Y}$ that satisfies (A.8) is said to be *reversible* for the chain. An important result states that any reversible probability vector must be equal to the vector of steady-state probabilities. This is often useful to show that a given vector $\boldsymbol{\pi}$ is a solution to (A.7): it suffices to show that it is reversible (Häggström 2002).

A state i is *periodic* with period d if $p_{i,i}^{(n)} = 0$ whenever n is not a multiple of d, and d is the largest integer for which this property holds. When the period is $d = 1$, the state is called *aperiodic*. In an irreducible DTMC, all states have the same period. If state i has period d, then $\lim_{k\to\infty} p_{i,i}^{(kd)} = d\pi_i$.

For the case where $d = 1$ for an irreducible positive recurrent chain, we have $\lim_{k\to\infty} p_{i,j}^{(k)} = \lim_{k\to\infty} p_j^{(k)} = \pi_j$ for all (i, j), and the total variation distance between $\mathbf{p}^{(k)}$ and $\boldsymbol{\pi}$ converges at a geometric (or exponential) rate, regardless of $\mathbf{p}^{(0)}$; that is,

$$\sup_{i,j} \left| p_{i,j}^{(k)} - \pi_j \right| \leq K\rho^k \tag{A.9}$$

for some constants $K > 0$ and $\rho < 1$. This geometric convergence actually occurs under a more general condition, called the *Doeblin condition*: There exists an integer k_0, a constant $\delta > 0$, and a probability distribution $\mathbf{p} = (p_0, p_1, \ldots)$ on the states, such that $p_{i,j}^{(k_0)} \geq \delta p_j$ for all states i, j. If the chain is irreducible and this condition holds, then we have a geometric convergence of the total variation distance, with $\rho = (1 - \delta)^{1/k_0}$.

Suppose that at step n, we incur a cost (or reward) C_n whose distribution may depend on (Y_{n-1}, Y_n) but not on n, and not on other states conditionally on (Y_{n-1}, Y_n), and the random variables $C_1, C_2, \ldots$ are independent conditionally on $\{Y_n, n \geq 0\}$. Let $c_{i,j} = \mathbb{E}[C_n \mid Y_{n-1} = i, Y_n = j]$ and $c_i = \mathbb{E}[C_n \mid Y_{n-1} = i] = \sum_{j=0}^{\infty} p_{i,j} c_{i,j}$. Let i be an arbitrary (positive recurrent) state, let $Y_0 = i$, and let $T_{ii} = \inf\{n > 0 \mid Y_n = i\}$. We have $\mathbb{E}[T_{ii}] = \nu_i = 1/\pi_i$. We can view each return to i as marking the end of a regenerative cycle. Then, the renewal reward theorem gives us:

Theorem A.21 (Renewal reward theorem). *Suppose* $Y_0 = i$ *and* $\mathbb{E}\left[\sum_{n=1}^{T_{ii}} |C_n|\right] < \infty$. *Then, the average cost per step in the long run obeys:*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} C_j \overset{\text{w.p.1}}{=} \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} \mathbb{E}[C_j] = \sum_{i=0}^{\infty} \pi_i c_i \overset{\text{def}}{=} \nu. \tag{A.10}$$

The leftmost expression in (A.10) is a random variable and the other expressions are constant (i.e., deterministic). The last sum is the *steady-state average cost*. It does not depend on i.

Suppose now that there is a set of *absorbing states* $A \subset \{0, 1, \ldots\}$ such that the chain stops at step $T_A = \inf\{n > 0 \mid Y_n \in A\}$. The total expected cost until absorption, when starting in state $Y_0 = i$, is

$$v_i \stackrel{\text{def}}{=} \mathbb{E}\left[\sum_{n=1}^{T_A} C_n \mid Y_0 = i\right]$$

for $i \notin A$ and $v_i = 0$ for $i \in A$. In particular, if the one-step cost is 1 when we reach a state in some fixed set $A' \subset A$, and 0 otherwise, then v_i represents the probability of reaching A' before $A \setminus A'$, when starting from i.

Theorem A.22 *If $\mathbb{P}[T_A < \infty \mid Y_0 = i] = 1$ for all i, then the v_i's are the unique finite solution to the linear system*

$$v_i = \sum_{j=0}^{\infty} p_{i,j}[c_{i,j} + \mathbb{I}(j \notin A)\, v_j] = c_i + \sum_{j \notin A} p_{i,j} v_j, \qquad i = 0, 1, \ldots$$

A.18.2 Continuous-time Markov chains

A stationary *continuous-time Markov chain* (CTMC) with state-space $\mathcal{Y} = \{0, 1, 2, \ldots\}$ is a continuous-time process $\{Y(t),\, t \geq 0\}$ for which

$$\mathbb{P}\left[Y(s+t) = j \mid \{Y(u),\, 0 \leq u \leq s\}\right] = \mathbb{P}\left[Y(s+t) = j \mid Y(s)\right]$$

and does not depend on s. This implies that the process is Markovian, and we can define

$$P_{i,j}(t) = \mathbb{P}[Y(s+t) = j \mid Y(s) = i].$$

These probabilities satisfy the *Chapman-Kolmogorov* equations:

$$P_{i,j}(s+t) = \sum_{k=0}^{\infty} P_{i,k}(s) P_{k,j}(t),$$

together with $\sum_{j=1}^{\infty} P_{i,j}(t) = 1$. We assume from now on that our CTMC is a pure jump process, with a strictly positive and finite sojourn time whenever it visits a state. Under this assumption, it can be shown that each sojourn time to a state i is exponential with some rate $\lambda_i > 0$. This λ_i is the jump rate out of that state. Moreover, the next state is j with some probability $p_{i,j}$, independently of the sojourn time. So if $0 = T_0 < T_1 < T_2 < \ldots$ are the jump times of the CTMC, and if we put $Y_n = Y(T_n)$ for $n \geq 0$, then $\{Y_n,\, n \geq 0\}$ is a DTMC with transition probabilities $p_{i,j}$. We call it the *embedded DTMC*. We also have that conditionally on $\{Y_n,\, n \geq 0\}$, the sojourn times $\{T_n - T_{n-1},\, n \geq 0\}$ are independent random variables and $T_n - T_{n-1}$ is exponential with rate $\lambda_{Y_{n-1}}$.

When in state i, the jump rate of the CTMC is λ_i and its jump rate to state j is $\lambda_{i,j} = \lambda_i p_{i,j}$. Define $\lambda_{i,i} = -\lambda_i$. The matrix A whose element (i,j) is $\lambda_{i,j}$ is known as the *infinitesimal generator* of the CTMC.

We say that the CTMC is *irreducible* if for all pairs of states (i,j), $P_{i,j}(t) > 0$ for some $t > 0$. This holds if and only if the DTMC is irreducible. The CTMC is *regular* if for any initial state $Y(0)$ and any finite time interval, the number of transitions in that time interval is finite w.p.1. A sufficient condition for regularity is that $\lambda_i \leq \lambda$ for all i, for some constant $\lambda < \infty$. When this condition holds, we can simulate the CTMC by generating jump times from a stationary Poisson process with rate λ. A jump that occurs when the process is in state i provokes a true transition with probability λ_i/λ (to state j with probability $\lambda_{i,j}/\lambda$), and is ignored (or considered as a transition to state i) with probability $1 - \lambda_i/\lambda$. This technique is known as *uniformization* of the chain.

The CTMC is *positive recurrent* if from any state i, the expected time to return to i is finite. The *key renewal theorem* states that for a positive recurrent CTMC, for all pairs (i,j),

$$\lim_{t\to\infty} P_{i,j}(t) = q_j,$$

where q_j is a constant that also represents the fraction of the time that the CTMC spends in state j, in the long run. The overall rate of occurrence of jumps from i to j is then $q_i\lambda_{i,j}$, and the rate of occurrence of jumps out of state i is $q_i\lambda_i$. The jump rate *into* state i must also equal this value, because the absolute difference between these two numbers of jumps is at most 1. More generally, if we split the state space in two, say $B \subset \mathcal{Y}$ and $\bar{B} = \mathcal{Y} \setminus B$, then the overall jump rate from B to $\bar{B}$ must be the same as that from $\bar{B}$ to B. By equaling those rates, we get the *balance equation*:

$$\sum_{i\in B}\sum_{j\in\bar{B}} q_i\lambda_{i,j} = \sum_{i\in\bar{B}}\sum_{j\in B} q_i\lambda_{i,j}.$$

To find the state probabilities q_i, we can select a collection of subsets B in a way that the corresponding balance equations are easy to solve and provide a unique solution, and then solve this linear system of equations. This method is often used to derive explicit formulas for the q_i's.

Suppose that costs are accumulated at rate $\gamma(i)$ when the CTMC is in state i. Then the total cost during the time interval $[0,t]$ is $V(t) = \int_0^t \gamma(Y(t))\mathrm{d}t$. If $\sum_{i=0}^{\infty} |\gamma(i)|q_i < \infty$, then the average cost per unit of time in the long run obeys

$$\lim_{t\to\infty} \frac{V(t)}{t} \overset{\text{w.p.1}}{=} \sum_{i=0}^{\infty} q_i\gamma(i).$$

Suppose now that a cost $\kappa(i)$ is incurred each time we jump to state i. Then the total cost during $[0,t]$ is $V(t) = \sum_{n=1}^{\infty} \kappa(Y_n)\mathbb{I}[T_n \leq t]$. If $\sum_{i=0}^{\infty} |\kappa(i)|\lambda_i q_i < \infty$, then the average cost per unit of time in the long run obeys

$$\lim_{t\to\infty} \frac{V(t)}{t} \overset{\text{w.p.1}}{=} \sum_{i=0}^{\infty} \kappa(i)q_i\lambda_i.$$

A *birth-and-death* process is a CTMC with state space $\mathcal{Y} = \{0, 1, 2, \ldots\}$, with jump rates $\lambda_{j,j+1} = \lambda_j$, $\lambda_{j,j-1} = \mu_j$, and $\lambda_{i,j} = 0$ when j is neither $i-1$ or $i+1$, or $j < 0$. Here, the embedded DTMC has period 2. By writing the balance equations $\lambda_j q_j = \mu_{j+1} q_{j+1}$ for all $j \geq 0$, we find that $q_j = b_j q_0$, where $b_0 = 1$ and $b_j = \lambda_0 \cdots \lambda_{j-1} / [\mu_1 \cdots \mu_j]$ for $j \geq 1$. Since the q_j's must sum to 1, we have $q_0 = 1/\sum_{j=0}^{\infty} b_j$, provided that this sum is finite (otherwise, the process is not positive recurrent).

♣ Add figure.

A.18.3 Markov chains on general state spaces

♣ To be done. See Meyn and Tweedie (1993). Stability. Harris recurrence. Ergodicity. Geometric ergodicity.

A.19 Queueing notation and formulas

When customers arrive at a service facility faster than the servers can handle them, a queue of waiting customers may form in front of the facility. Usually, the queue is due to randomness in the arrival and service times, and is only temporary (it eventually disappears, then reappears, and so on). A *queueing system* is comprised of a customer population, an arrival process for the customers, and a service facility with one or more servers. When all servers are busy, arriving customers must wait in the queue. The queueing discipline is often *first-come first-served* (FCFS), but there are many other possibilities such as *last-come first-served* (LCFS), shortest-job first, ordering by priorities, service in random order, and so on. In a single-server queue, FCFS and LCFS are equivalent to *first-in first-out* (FIFO) and *last-in first-out* (LIFO), respectively, but this is not necessarily true when there are many servers, because the order in which customers leave may differ from the order in which they start service. The *capacity* of the system (maximal number in queue or in service), and the size of the population (total number of potential customers), could be infinite or finite. When the system is filled to its capacity, all arrivals are lost. If the capacity equals the number of servers, there is never a queue and we speak of a *loss system*. Some models consider possible *abandonment* of customers already in the queue (*reneging*) or upon arrival when they think the queue is too long (*balking*).

A standard nomenclature to specify queueing systems (or models) is the Kendall notation, in which $A/B/s/K$ means a system where the interarrival times between successive customers has distribution type A, service times have distribution type B, there are s identical servers, and the system has capacity K (assumed to be infinite when the parameter K is omitted). When specifying the interarrival and service time distributions, G means a general (unspecified) distribution, GI means general and independent (used for the interarrival

times, since the independence is usually implicit for the service times), M means Markovian (which means i.i.d. and exponentially distributed), and D means deterministic (a constant).

Classical references on queueing models include Gross and Harris (1998), Kleinrock (1975), Kleinrock (1976), Wolff (1989).

Example A.10 An $M/M/s$ queue has i.i.d. exponential interarrival times, i.i.d. exponential service times, s servers, and infinite capacity. An $M/M/1$ queue is a special case, with a single server. A $GI/G/1$ queue has i.i.d. interarrival times with arbitrary distribution, i.i.d. service times with arbitrary distribution, and a single server. □

For a queueing system operating in steady-state, we use the following notation:

λ	average *arrival rate* (inverse of the mean time between arrivals)
μ	mean *service rate* (inverse of the mean service time)
r	$= \lambda/\mu$: *traffic intensity*, or *load* (in Erlangs)
s	number of identical servers
ρ	$= \lambda/(s\mu)$: *average server utilization* (fraction of the time a server is busy, in the case where no customer is lost)
w	*average waiting time* in the queue, per customer
q	*average queue length* (number of waiting customers), with respect to time
ℓ	average number of customers in the system (waiting or being served)
W	waiting time of a random customer, in steady-state
δ	$= \mathbb{P}[W > 0]$: *delay probability* (fraction of customers not served immediately)

A key relationship is *Little's formula* (see Kleinrock 1975, pages 187–190):

$$q = \lambda w. \tag{A.11}$$

It says that the average number of customers in the queue is the arrival rate multiplied by the average time spent in the queue. The intuitive interpretation is that over a very long time interval of length t, the total waiting time of all customers during that interval is on the one hand the integral of the queue length over that interval, which is approximately qt, and on the other hand it is approximately the total waiting time of all customers who arrived during that interval, which is approximately $\lambda t w$. It is important to observe that this formula applies to any subsystem of the queueing system, and to any subnetwork in the case of a queueing network. In general, we can replace λ by the arrival rate to the subsystem, w by the average time a customer spends in the subsystem, and q by the average number in the subsystem. In particular, if we consider the customers in the entire system (waiting or being served), we get the variant

$$\ell = \lambda(w + 1/\mu).$$

The $M/M/s$ queue evolves as a birth-and-death process whose state is the number of customers in the system, and its steady-state probabilities are easily obtained using the general formulas for birth-and-death processes. We get the *Erlang-C* formula for the delay probability:

$$\delta = \mathbb{P}[W > 0] = \frac{\dfrac{r^s}{(s-1)!(s-r)}}{\dfrac{r^s}{(s-1)!(s-r)} + \displaystyle\sum_{i=0}^{s-1} \frac{r^i}{i!}}. \tag{A.12}$$

When $\lambda \to \infty$, $s \to \infty$, and $(1-\rho)\sqrt{s} \to \beta$ for $0 < \beta < 1$, then (Halfin and Whitt 1981)

$$1/\delta \to 1 + \beta\Phi(\beta)/\phi(\beta), \tag{A.13}$$

where Φ and ϕ are the standard normal distribution and density functions. The right side of (A.13) is the *Halfin-Whitt approximation*. It is often used to approximate the value of s required for a given δ; this approximation gives the *square root safety staffing formula* $s \approx r + \beta\sqrt{r}$.

Given that a customer must wait, in the $M/M/s$ queue, its conditional waiting time in queue is exponential with mean $1/(s\mu - \lambda)$. The probability of waiting more than x, for $x \geq 0$, is then

$$\mathbb{P}[W > x] = \mathbb{P}[W > 0] \exp[-(s\mu - \lambda)x]. \tag{A.14}$$

We also have $w = \delta/[\mu s(1-\rho)]$. When $s = 1$ (the $M/M/1$ queue), this simplifies to $\delta = \rho = \lambda/\mu$ and $w = \rho/[\mu(1-\rho)]$.

For an $M/M/s/K$ queue, the probability q_j that there are j customers in the system is

$$\begin{aligned} q_j &= q_0 r^j/j! \quad \text{for } j = 1, \ldots, s, \\ q_j &= q_0 \binom{r}{s}^{j-s} r^s/s! \quad \text{for } j = s+1, \ldots, K, \end{aligned}$$

where

$$\frac{1}{q_0} = \sum_{j=0}^{s} \frac{r^j}{j!} + \frac{r^s}{s!} \sum_{j=1}^{K-s} \binom{s}{r}^j.$$

The fraction of customers who are lost is $\lambda(1 - q_K)$. The average queue length is

$$q = \sum_{j=s+1}^{K} (j-s) q_j.$$

For an $M/G/1$ queue, the delay probability is still $\delta = \rho$, and w can be computed by the Pollaczek-Khintchine (P-K) mean value formula for the average waiting time (Kleinrock 1975):

$$w = \frac{\lambda \mathbb{E}[S^2]}{2(1-\rho)} = \frac{\rho}{\mu(1-\rho)} \frac{1+c_s^2}{2}, \tag{A.15}$$

where c_s is the coefficient of variation of a customer's service time. If we multiply this equation by λ and use Little's law, we obtain the P-K formula for q.

- ♣ Abandonments: Erlang-A formulas
- ♣ Jackson networks

B. Bachmann-Landau (or big $\mathcal{O}$ and little o) notation

Given two real-valued functions $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$, we say that

$$f(x) = \mathcal{O}(g(x)) \text{ as } x \to \infty$$

if and only if there is a finite constant $K > 0$ and a real number x_0 such that $|f(x)| \leq K\,|g(x)|$ for all $x > x_0$. Sometimes, the functions are defined only on a subset of the real numbers; then we consider only the values of x that belong to that subset. In particular, when x can only non-negative integer values, we usually replace x by n and say $f(n) = \mathcal{O}(g(n))$.

Similarly, for an arbitrary constant a (usually $a = 0$),

$$f(x) = \mathcal{O}(g(x)) \text{ as } x \to a$$

if and only if there is a finite constant $K > 0$ and a real number δ such that $|f(x)| \leq K\,|g(x)|$ whenever $|x - a| < \delta$.

Usually, the "$x \to \infty$" or "$x \to a$" is clear from the context and is omitted. For example, $f(n) = 42/n + O(1/n^2)$ means that $f(n) - 42/n$ is upper bounded by a quantity that converges at least as fast as K/n^2 for some constant K when $n \to \infty$.

As another example, if we say that algorithm takes $\mathcal{O}(n \log n)$ time and $\mathcal{O}(n^2)$ space in the worst-case, this means that there is a constant K and some integer n_0 such that for any problem instance of size $n \geq n_0$ (the meaning of "size" is problem-dependent), the algorithm will solve the problem in less than $Kn \log n$ units of time and using less than Kn^2 units (e.g., bytes) of memory.

If $f(x) = \mathcal{O}(g(x))$ and $g(x) = \mathcal{O}(f(x))$, then we say that $f(x) = \Theta(g(x))$.

We say that

$$f(x) = o(g(x)) \text{ as } x \to \infty$$

if and only if $\limsup_{x\to\infty} |f(x)/g(x)| = 0$. That is, if for any $\epsilon > 0$, there is a real number x_0 such that $|f(x)| \leq \epsilon\,|g(x)|$ for all $x > x_0$. For an arbitrary constant a,

$$f(x) = o(g(x)) \text{ as } x \to a$$

if and only if $\limsup_{|x-a|\to 0} |f(x)/g(x)| = 0$. The interpretation is that $|f(x)|$ converges to 0 much faster than $g(x)$, or increases much slower than $|g(x)|$, when $x \to \infty$ (or when $x \to a$).

Some authors prefer to write $f(x) \in \mathcal{O}(g(x))$ instead of $f(x) = \mathcal{O}(g(x))$, and this makes more sense because $\mathcal{O}(g(x))$ is actually the *set* of functions f for which the property holds. The same applies to Θ and o. But even though $f(x) = \mathcal{O}(g(x))$ is an abuse of notation, it is more standard and well encrusted.

References

Billingsley, P. 1968. *Convergence of Probability Measures*. New York, NY: John Wiley.

Billingsley, P. 1986. *Probability and Measure*. second ed. New York, NY: Wiley.

Chung, K. L. 1974. *A Course in Probability Theory*. second ed. New York, NY: Academic Press.

Feller, W. 1971. *An Introduction to Probability Theory and Its Applications, Vol. 2*. second ed. New York, NY: Wiley.

Galambos, J. 1995. *Advanced Probability Theory*. second ed. New York, NY: Marcel Dekker.

Gross, D and C. M. Harris. 1998. *Fundamentals of Queueing Theory*. Third ed. New York, NY: Wiley.

Häggström, O. 2002. *Finite Markov Chains and Algorithmic Applications*. Cambridge, U.K.: Cambridge University Press.

Halfin, S and W. Whitt. 1981. Heavy-traffic limits for queues with many exponential servers. *Operations Research*, 29(3):567–588.

Hogg, R. V. and A. F. Craig. 1995. *Introduction to Mathematical Statistics*. 5th ed. Prentice-Hall.

Katz, M. 1963. A note on the Berry-Esseen theorem. *Annals of Mathematical Statistics*, 34:1007–1008.

Kleinrock, L. 1975. *Queueing Systems, Vol. 1*. New York, NY: Wiley.

Kleinrock, L. 1976. *Queueing Systems, Vol. 2*. New York, NY: Wiley.

Meyn, S. P and R. L. Tweedie. 1993. *Markov Chains and Stochastic Stability*. Springer-Verlag.

Read, T. R. C. and N. A. C. Cressie. 1988. *Goodness-of-Fit Statistics for Discrete Multivariate Data*. Springer Series in Statistics, New York, NY: Springer-Verlag.

Rice, J. A. 1995. *Mathematical Statistics and Data Analysis*. second ed. Belmont, California: Duxbury Press.

Serfling, R. J. 1980. *Approximation Theorems for Mathematical Statistics*. New York, NY: Wiley.

Shao, J. 1999. *Mathematical Statistics*. New York, NY: Springer.

Wolff, R. W. 1989. *Stochastic Modeling and the Theory of Queues*. New York, NY: Prentice-Hall.

Index