Online Appendix B

Preliminaries in Probability and Analysis

It is remarkable that a science which began with the consideration of games of chance should have become the most important object of human knowledge . . . . The most important questions in life are, for the most part, only problems in probability.
—Pierre-Simon Laplace, Marquis de Laplace (1749–1827) in *Théorie Analytique des Probabilités*

I cannot believe that God would choose to play dice with the universe.
—Albert Einstein (1879–1955)

I would suggest that nobody—not even God—would know what a phrase like playing dice would mean in this context.
—Niels Henrik David Bohr (1885–1962), reply to Einstein in 1949 on the occasion of Einstein's 70th birthday, continuing their famous discussion on the basis of quantum mechanics

It is so easy to see far and discover when standing on the shoulders of giants, who before us have developed prior knowledge.
—Sir Isaac Newton (1642–1727) as quoted in [233]

There is randomness and hence uncertainty in mathematics, just as there is in physics.
—Paul Davis

This online appendix provides a practical common background for necessary applied probability concepts for continuous and discrete random variables. These concepts include conservation of probability, expectation, variance, higher moments, and basic distributions of interest. Also treated are applied analysis concepts of discontinuity and nonsmoothness for deterministic processes, i.e., regular functions of time, as they affect regular calculus concepts of Taylor approximations, asymptotics, and optimality principles. There is more in this appendix than many readers would be expected to know, so it should at least be browsed for familiarity and consulted as a reference.

B.1 Distributions for Continuous Random Variables

Uppercase variables, such as $X = X(\omega)$, denote random variables, which are, in general, functions of some underlying random parameter or variable ω defined on some standard sample space Ω. For notational simplicity, the dependence on the underlying or background random variable $\omega \in \Omega$ will often be suppressed. Variables in lower case letters, such as x, denote the actual sample variables or realizations associated with the random variables and are used as the dummy variables in integrals.

B.1.1 Probability Distribution and Density Functions

Definition B.1. *The symbol Φ denotes the corresponding **probability distribution** such that*

$$\Phi(x) \equiv \text{Prob}[X \leq x] \tag{B.1}$$

in the case of a distribution on $-\infty < X < +\infty$. Here, the notation Prob *denotes the probability function for the probability of occurrence of events on a subset as the ratio relative to all events in the sample space. Elsewhere many other notations are used, such as the minimal **P** and **Pr**.*

If the **distribution is proper**, then $\Phi(+\infty) = 1$, i.e., we say probability is conserved. Also, $\Phi(-\infty) = +0$ and Φ is obviously continuous as long as the probability distribution contains no jumps in value. However, later in this book, we will consider more general random processes, in continuous time, that are composed of continuous processes as well as processes with jump discontinuities, possibly a countably infinite number of jumps. Thus, in general, we have the following.

Properties B.2. ***Continuous Distribution Functions, $\Phi(x)$.***

- Φ *is nondecreasing, since probabilities must be nonnegative.*
- Φ *is continuous by properties of integrals with nonnegative integrands (assuming there are no probability point masses, i.e., discrete components).*
- $\Phi(-\infty) = +0$ *by properties of integrals and $X > -\infty$.*
- $\Phi(+\infty) = 1$ *if Φ is a **proper distribution**.*

- $\Phi(x+y) = \Phi(x) + \text{Prob}[x < X \leq x+y],\ y > 0$ *by the additivity of probability over disjoint sets, which here are* $(-\infty, x]$ *and* $(x, x+y]$.

Definition B.3. *The symbol* ϕ *will denote a* ***probability density*** *such that*

$$\phi(x)dx = \text{Prob}[x < X \leq x + dx] \tag{B.2}$$

in terms of the probability for the continuous random variable X.

Properties B.4. ***Relation Between Distribution and Density.***

- *By the additivity of probability and definition of the distribution function,*

$$\phi(x)dx = \text{Prob}[x < X \leq x + dx] = \Phi(x + dx) - \Phi(x).$$

- *Thus, for infinitesimal* dx *and* Φ *differentiable,*

$$\phi(x)dx = \Phi'(x)dx,$$

so

$$\phi(x) = \Phi'(x). \tag{B.3}$$

The differentiability of the distribution Φ *is not considered a serious restriction here, since differentiability in the generalized sense will be considered in Section* B.12.

- *The relationship between the distribution function and the density in integral form is*

$$\Phi(x) \equiv \text{Prob}[X \leq x] \equiv \int_{-\infty}^{x} \phi(y)dy \tag{B.4}$$

in the case of a differentiable distribution on $-\infty < X < +\infty$.

- *Another more general form is*

$$\Phi(x) \equiv \text{Prob}[X \leq x] \equiv \int_{-\infty}^{x} d\Phi(y),$$

which is called a ***Stieltjes integral****. In abstract formulations, the differential is written* $d\Phi(y) = \Phi(dy)$ *as shorthand notation for* $\Phi((y, y+dy])$ *in the half-open interval notation here.*

- *Sometimes it is useful to transform the random variable* X *to a more convenient random variable* Y*, where* $X = \psi(Y)$*, for example. Consequently, for clarity of notation, let* $\phi(x) = \phi_X(x)$ *and similarly* $\Phi(x) = \Phi_X(x)$*, adding an extra subscript to mark which random variable pertains to a given density or distribution function since the argument* x *is only a dummy variable. Thus, the* ***change of distribution for a change of random variable*** *on the interval* $(x_1, x_2]$ *is written*

$$\begin{aligned}\Phi_X(x_2) - \Phi_X(x_1) &= \int_{x_1}^{x_2} \phi_X(x)dx \\ &= \int_{y_1}^{y_2} \phi_Y(y)dy = \Phi_Y(y_2) - \Phi_Y(y_1),\end{aligned} \tag{B.5}$$

where

$$\phi_Y(y) = \phi_X(x)\left|\frac{dx}{dy}\right| = \phi_X(x)\left|\psi'(y)\right| \tag{B.6}$$

provided $\psi(y)$ is a differentiable monotonic function on (y_1, y_2), i.e., either $\psi'(y) > 0$ or $\psi'(y) < 0$, where, in either case, the limits of integration are given by

$$y_1 = \min[\psi^{-1}(x_1), \psi^{-1}(x_2)]$$

and

$$y_2 = \max[\psi^{-1}(x_1), \psi^{-1}(x_2)].$$

B.1.2 Expectations and Higher Moments

In general, there are basic definitions for averaged quantities in the case of continuous distributions.

Definition B.5. *The **mean** or **expectation** of any continuously distributed random variable X is*

$$\mu \equiv \mathrm{E}[X] \equiv \int_{-\infty}^{+\infty} x\phi(x)dx \tag{B.7}$$

provided the above integral converges absolutely. The symbol E *is the expectation operator. Similarly, the expectation of a function of X, $f(X)$, is*

$$\mathrm{E}[f(X)] \equiv \int_{-\infty}^{+\infty} f(x)\phi(x)dx \tag{B.8}$$

provided the integral converges absolutely.

Properties B.6. ***Expectations.***

- *The expectation operator is a linear operator,*

$$\mathrm{E}[c_1X_1 + c_2X_2] = c_1\mathrm{E}[X_1] + c_2\mathrm{E}[X_2], \tag{B.9}$$

provided the expectations exist, for random variables X_i and constants c_i, for $i = 1:2$ (using MATLAB notation for the range of i).

Definition B.7. *The **variance** or **mean square deviation** or **second central moment** for any continuously distributed random variable X is*

$$\sigma^2 \equiv \mathrm{Var}[X] \equiv \mathrm{E}[(X - \mathrm{E}[X])^2] = \int_{-\infty}^{+\infty} (y-\mu)^2\phi(y)dy \tag{B.10}$$

*provided the integral converges absolutely. The deviation and the central moments are defined relative to the mean μ. The square root of the variance σ is called the **standard deviation**.*

While the mean and the variance are the most often used moments of the distribution, i.e., of the density, sometimes some of the higher moments are useful for further characterizing the distribution.

Definition B.8. *The **third central moment** is defined here in the normalized form called the **skewness coefficient*** [83] *for the random variable X:*

$$\eta_3[X] \equiv \mathrm{E}[(X - \mathrm{E}[X])^3]/(\mathrm{Var}[X])^{3/2} \tag{B.11}$$

such that the distribution is negatively skewed, symmetric, or positively skewed, if $\eta_3[X]$ is negative, zero, or positive, respectively (zero being the skew of the normal distribution as discussed in Subsection B.1.4*).*

Definition B.9. *The **fourth central moment** is a measure of **kurtosis** (peakedness) and is defined here in the normalized form called the **kurtosis coefficient*** [83] *for the random variable X:*

$$\eta_4[X] \equiv \mathrm{E}[(X - \mathrm{E}[X])^4]/(\mathrm{Var}[X])^2 \tag{B.12}$$

*such that the distribution is **platokurtic** or **leptokurtic** if the **coefficient of excess kurtosis** ($\eta_4[X] - 3$) is negative or positive, respectively. (The value 3 is the value of $\eta_4[X]$ for the normal distribution, discussed in Subsection* B.1.4*.)*

The property of kurtosis, from the Greek word for convexity, signifies more at the crown (as seen from the density) for a distribution with peakedness in the case of leptokurtic and a distribution with flatness in the case of platokurtic. The kurtosis property together with skewness is of particular interest in mathematical finance for characterizing nonnormal properties of real market distributions.

The little book on statistical distributions of Evans, Hastings, and Peacock [83] concisely lists principal formulae for skewness, kurtosis, and many other properties for 40 distributions. The book has more useful and easy-to-find information in it than other books on distributions, including those requiring several volumes.

B.1.3 Uniform Distribution

The most fundamental continuous probability distribution is the uniform distribution.

Definition B.10. *The **uniform density** on the finite interval $[a, b]$ is defined as*

$$\phi_u(x; a, b) \equiv \begin{Bmatrix} 1/(b-a), & x \in [a, b] \\ 0, & x \notin [a, b] \end{Bmatrix}. \tag{B.13}$$

Definition B.11. *The **uniform distribution** is defined by integrating the uniform density,*

$$\Phi_u(x; a, b) \equiv \int_{-\infty}^{x} \phi_u(y; a, b)dy = \begin{Bmatrix} 0, & x \le a \\ (x-a)/(b-a), & a \le x \le b \\ 1, & b \le x \end{Bmatrix}, \tag{B.14}$$

$-\infty < x < +\infty$*, so that $\Phi_u(x; a, b) = 1$ for $b \le x < +\infty$, conserving total probability.*

Hence, the basic moments and other properties easily follow from simple integration.

Properties B.12. ***Uniform Distribution Moments.***

- ***Conservation of probability:*** $\mathrm{E}_u[1] = 1$.
- ***Mean:***

$$\mu = \mathrm{E}_u[X] = \int_a^b x\phi_u(x; a, b)dx = (b + a)/2. \tag{B.15}$$

- ***Variance:***

$$\sigma^2 = \mathrm{Var}_u[X] = \int_a^b (x - E_u[X])^2\phi_u(x; a, b)dx = (b - a)^2/12. \tag{B.16}$$

- ***Uniform domain correspondence to mean and variance:*** $a = \mu - \sqrt{3}\sigma$ *and* $b = \mu + \sqrt{3}\sigma$.
- ***Coefficient of skew:*** $\eta_3 = 0$.
- ***Coefficient of kurtosis:*** $\eta_4 = 1.8$ *or* $\eta_4 - 3 = -1.2$ *is the excess value over the normal value.*

Hence, the uniform distribution is platokurtic, signifying its obvious flatness compared to normal.

An important use of the uniform distribution is the numerical simulation of the distributions that can be transformed from the uniform distribution. The most basic random number generator is the standard uniform random number generator. The standard uniform random number generator is usually based on a deterministic generator called the linear congruential generator [230, 97] that is defined as nonzero on the open interval (0, 1) instead of the closed interval [0, 1] as for the theoretical distribution $\phi_u(x; 0, 1)$, which is more convenient for numerical purposes and the end points do not contribute to the expectation integral anyway. Most computing systems, such as MATLAB [210], Maple [1], or Mathematica [285], and programming languages have a built-in uniform random number generator but must be used with care considering that they use deterministic operations such as modular arithmetic, multiplication, and division. These random number generators are more properly called **pseudo-random number generators** since they generate only approximations to random numbers, which exist only exactly in theory. Pseudo-random numbers should be carefully tested before using them in any computation. For instance, the MATLAB uniform generator is called `rand` (*note that the MATLAB functions and code fragments are typeset in* `typewriter` *style*) and can simulate an approximation to a scalar, vector, or more general array of random numbers. Figure B.1 illustrates the histograms of a row vector with N simulations of uniform deviates for $\phi_u(x; 0, 1)$ using the form

```
x = rand(N, 1)
```

or more generally

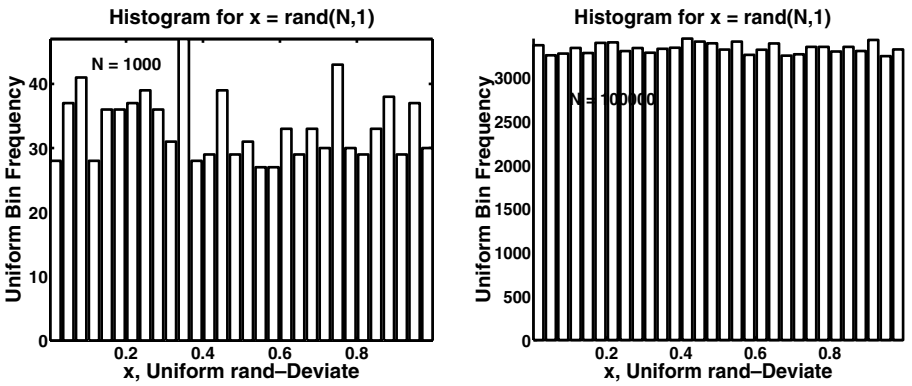
```
y = a + (b − a) ∗ rand(N, 1),
```

which simulates an N-vector sample uniform on (a, b) in MATLAB. Other computing systems may use a programming loop with N iterations. The approximate distribution

displays the bin-centered histogram function `hist(x)`. Scaling the bin frequencies upon normalizing by the average bin count `N/nbins`, where **`nbins`** is the number of bins, here 30 bins, would produce a scaled histogram more appropriate for approximating probability density, $\phi_u(x; 0, 1)$, of the theoretical uniform distribution. Thus, if f_i is the frequency associated with the ith bin $[x_i, x_i + \Delta x)$ `for i = 1 : nbins`, in MATLAB loop notation, of width Δx, then

$$\sum_{i=1}^{\texttt{nbins}} f_i = N \text{ or } \frac{1}{N} \sum_{i=1}^{\texttt{nbins}} f_i = 1,$$

the latter in normalized form.



(a) Sample size $N = 10^3$.

(b) Sample size $N = 10^5$.

Figure B.1. *Histograms of simulations of uniform distribution on* $(0, 1)$ *using MATLAB* [210] *for two different sample sizes N.*

Clearly, the larger sample size simulation with $N = 100{,}000$ in Subfigure B.1(b) is a much better approximation of the uniform approximation then the much cruder representation with $N = 1000$ in Subfigure B.1(a). The relative error for the sample mean is -0.24% for $N = 1000$ and -0.43% for $N = 100{,}000$.

Note that the error in the sample mean did increase slightly with sample size, but these are only single samples, and it would not be realistic to draw any general conclusions from this case. These are just approximations to random samples, although it would be reasonable to expect that the average over repeated samples would be lower the higher the sample size, provided that the selected random number generator is sufficiently robust. Computing more pairs of samples using the same sizes, $N = 1000$ and $N = 100{,}000$, with different random states would demonstrate that the sample mean error would likely be smaller, but not necessarily. The relative errors for the sample standard deviation (square root of the sample variance) are 0.95% for $N = 1000$ and -0.20% for $N = 100{,}000$, which is more reasonable.

The sample variance is obtained from the MATLAB function `var(x)`, which is normalized by number of degrees of freedom $(N - 1)$ for the best estimate of the variance,

correcting for conditioning due to the mean value, which in MATLAB is the function `mean(x)`.

For more sophisticated distribution validation tests, chi-square (χ^2) or better, Kolmogorov–Smirnoff [230] tests can be used. The two samples displayed in Figure B.1 illustrate the problem of single samples requiring the averaging of several independent replications using a different random number generator initialization, called a **random seed** but now called a **`state`** in MATLAB (e.g., `rand('state', j)` sets `rand` in the `j`th `state`), so the error systematically decreases with sample size. Otherwise, the user can take a larger sample size. See Online Appendix C, Section C.1, for the MATLAB figure code.

In this appendix, we present empirical representations of distributions by histograms derived from random number generation, rather than the purely mathematical graphs of the probability density as portrayed in probability and statistics texts. This is to emphasize that the distributions derived from real environments are not as ideal as the exact mathematical density functions. Another reason is to emphasize that sometimes computations are necessary when no exact solutions are available or useful when exact solutions are too complicated, beyond the expertise of the entry-level graduate student or advanced undergraduate student.

B.1.4 Normal Distribution and Gaussian Processes

A continuous distribution of interest for Gaussian processes and other applications is given in terms of the normal probability density, the derivative of the normal or Gaussian probability distribution.

Definition B.13. *The **normal density** with mean $\mu = \mathrm{E}_n[X]$ and $\sigma^2 = \mathrm{Var}_n[X]$ is defined as*

$$\phi_n(x;\mu,\sigma^2) \equiv \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad -\infty < x < +\infty\,,\ \sigma > 0, \tag{B.17}$$

*where ϕ_n denotes the normal density function with argument x and parameters $\{\mu, \sigma^2\}$ following the semicolon. Here, X is called the **normal random variate.***

Definition B.14. *The **normal distribution** is defined here through the density as*

$$\Phi_n(x;\mu,\sigma^2) \equiv \int_{-\infty}^{x} \phi_n(y;\mu,\sigma^2)dy\,, \quad -\infty < x < +\infty, \tag{B.18}$$

so that $\Phi_n(+\infty;\mu,\sigma^2) = 1$, conserving total probability.

Remark B.15. *The normal distribution can be computed using MATLAB, Maple, or Mathematica computing systems, but the common special function that can be used, without resorting to special packages, is the **error function complement**,*

$$\mathrm{erfc}(x) = 1 - \mathrm{erf}\,(x) = \frac{2}{\sqrt{\pi}} \int_x^{\infty} e^{-t^2} dt, \tag{B.19}$$

so that the normal distribution can be computed from these two identities

$$\Phi_n(x;\mu,\sigma^2) = \frac{1}{2}\text{erfc}\left(\frac{\mu - x}{\sqrt{2}\sigma}\right) \tag{B.20}$$

$$= 1 - \frac{1}{2}\text{erfc}\left(\frac{x - \mu}{\sqrt{2}\sigma}\right). \tag{B.21}$$

Properties B.16. ***Normal Distribution Skew and Kurtosis.***

- *The normal distribution is* ***skewless****, since the coefficient of skew is* $\eta_3[X] = 0$.
- *The normal distribution has no* ***excess kurtosis****, since the coefficient of excess kurtosis is* $(\eta_4[X] - 3) = 0$*, where* 3 *is the coefficient of kurtosis of the normal distribution.*

As with the uniform distribution, the normal distribution is a theoretical idealization that is very useful in the analysis of stochastic problems. However, for practical computations, numerical simulations are usually necessary. Since the normal density function is an exponential of a quadratic, direct transformation from a uniform random generator is not possible. However, the usual normal random number generating algorithm, called Box–Muller [230, 97], cleverly applies the uniform random generator to a polar coordinate version of a two-dimensional normal distribution, reminiscent of the classic technique of converting a normal probability integral on the infinite domain from one-dimension to two dimensions and polar coordinates to get exact integral values. In some computing systems there is a special built-in function for a normal random generator. In MATLAB [210] the function is called **`randn`**, also having vector or array capabilities in the vector form `x = randn(N, 1)` for a $N-$vector sample. (More generally, `y = mu + sigma * randn(N, 1)` would simulate the density $\phi_n(y;\,\texttt{mu},\,\texttt{sigma}^2)$, where `mu` is the specified mean and `sigma` is the specified standard deviation.) The simulated normal density is illustrated by the histogram in Figure B.2 using two sample sizes, $N = 1000$ and 100,000. Clearly, the larger sample size in Subfigure B.2(b) gives a better qualitative representation of the theoretical bell-shaped curve of the normal density $\phi_n(x;\,0,\,1)$. The percent relative errors in the mean and standard deviation are, respectively, -1.53% and -0.35% for $N = 1000$, while the errors are 1.31% and -0.083% for the $N = 100{,}000$ sample size. See Online Appendix C, Section C.2, for the MATLAB figure code.

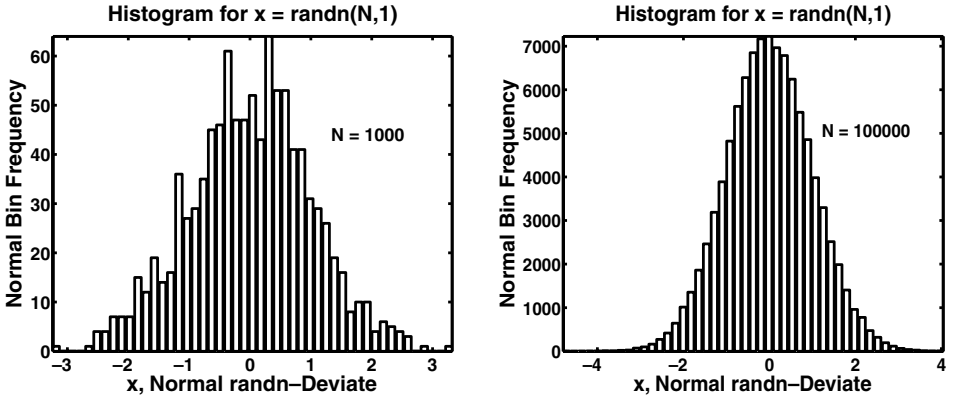
B.1.5 Simple Gaussian Processes

For later use, we will let $W(t)$ denote what is called a standard, mean zero Wn zero Wiener process with distribution

$$\Phi_{W(t)}(x) = \Phi_n(x;0,t), \quad -\infty < x < +\infty\,, \quad t > 0, \tag{B.22}$$

with corresponding probability density

$$\phi_{W(t)}(x) = \phi_n(x;0,t), \quad -\infty < x < +\infty\,, \quad t > 0. \tag{B.23}$$



(a) Sample size $N = 10^3$.

(b) Sample size $N = 10^5$.

Figure B.2. *Histograms of simulations of the standard normal distribution with mean* 0 *and variance* 1 *using MATLAB* [210] *with* 50 *bins for two sample sizes N. The histogram for the large sample size of* $N = 10^5$ *in* (b) *exhibits a better approximation to the theoretical normal density* $\phi_n(x; 0, 1)$.

A simple **Gaussian process** with linear mean growth in time,

$$X = G(t) = \mu t + \sigma W(t), \tag{B.24}$$

has mean $\mathrm{E}[X] = \mu t$ and variance $\mathrm{Var}[X] = \sigma^2 t$, so that the distribution of this process is

$$\Phi_{G(t)}(x) = \Phi_n(x; \mu t, \sigma^2 t) = \frac{1}{\sqrt{2\pi\sigma^2 t}} \int_{-\infty}^{x} e^{-\frac{(y-\mu t)^2}{2\sigma^2 t}} dy \tag{B.25}$$

on $-\infty < x < +\infty$, $t > 0$. The standard Wiener and Gaussian processes are also called diffusion processes, so they form models of the diffusion part of the jump-diffusion processes that are the main topic in this book. To see the connection between the stochastic Gaussian process and the deterministic diffusion process, let

$$u(x, t) = \Phi_{G(t)}(x)$$

and take partial derivatives of $u(x, t)$ with respect to t and x to derive the diffusion equation with **drift** $(-\mu)$ and **diffusion coefficient** $(\sigma^2/2)$,

$$u_t(x, t) = -\mu u_x(x, t) + \tfrac{\sigma^2}{2} u_{xx}(x, t), \quad -\infty < x < +\infty\,, \quad t > 0, \tag{B.26}$$

where the subscripts on u_t, u_x, and u_{xx} denote partial derivatives and the equation is called a partial differential equation (PDE).

Remarks B.17.

- *Here we use the term Gaussian process as it is used in applied mathematics, science, and engineering, i.e., for processes that are normally distributed. (For a more abstract view of Gaussian processes, see Mikosch* [209].*)*
- *There will be much more on the Wiener and Gaussian processes later, since they form the basic process for building the diffusion component of the jump-diffusion processes.*

B.1.6 Lognormal Distribution

Often in applications, such as in many linear financial models, the exponential of a normally distributed random variable arises and the distribution of this exponential is called a **lognormal distribution** since its logarithm produces the normally distributed exponent.

Theorem B.18. *Let*

$$X_{ln} = \exp(\mu + \sigma X_n) \tag{B.27}$$

be a ***lognormal variate*** *and let X_n be a standard normal variate, i.e., with zero mean and unit variance, Then* ***lognormal density*** *with mean $\mu_{ln} = \mathrm{E}[X_{ln}]$ and $(\sigma_{ln})^2 = \mathrm{Var}[X_{ln}]$ can be written in terms of the normal density ϕ_n* (B.17) *such that*

$$\phi_{ln}\left(x; \mu_{ln}, (\sigma_{ln})^2\right) \equiv x^{-1}\phi_n\left(\ln(x); \mu, \sigma^2\right) \quad 0 < x < +\infty\,,\ \sigma > 0, \tag{B.28}$$

where ϕ_{ln} denotes the lognormal density function with argument x and parameters $\{\mu_n, (\sigma^2)_n\} = \{\mu, \sigma^2\}$ follow the semicolon. If $x = 0$, then define ϕ_{ln} as the limiting case:

$$\phi_{ln}\left(0; \mu_{ln}, (\sigma_{ln})^2\right) \equiv \phi_{ln}\left(0^+; \mu_{ln}, (\sigma_{ln})^2\right) = 0. \tag{B.29}$$

Proof. Let the realization variable satisfy $x > 0$, recall that $\sigma > 0$, and that the natural logarithm is an increasing function. Consider the corresponding lognormal distribution definition, subsequently manipulated into the normal distribution:

$$\begin{aligned}
\Phi_{ln}\left(x; \mu_{ln}, (\sigma_{ln})^2\right) &= \mathrm{Prob}\left[X_{ln} \le x\right] && \text{(B.30)}\\
&= \mathrm{Prob}\left[\exp(\mu + \sigma X_n) \le x\right] && \text{(B.31)}\\
&= \mathrm{Prob}\left[X_n \le (\ln(x) - \mu)/\sigma\right] && \text{(B.32)}\\
&= \Phi_n((\ln(x) - \mu)/\sigma; 0, 1) && \text{(B.33)}\\
&= \Phi_n(\ln(x); \mu, \sigma^2). && \text{(B.34)}
\end{aligned}$$

The last step follows a normal distribution or density identity that allows transforming from the standard normal to nonstandard normal with mean μ and variance σ^2 (see Exercise 9 on p. B70). Upon taking the derivatives of the first and the last of this chain of equations and using the chain rule to handle the logarithmic argument of the normal distribution, the relationship between the densities is

$$\begin{aligned}
\phi_{ln}\left(x; \mu_{ln}, (\sigma_{ln})^2\right) &= (\Phi_{ln})'\left(x; \mu_{ln}, (\sigma_{ln})^2\right)\\
&= x^{-1}(\Phi_n)'\left(\ln(x); \mu, \sigma^2\right)\\
&= x^{-1}\phi_n\left(\ln(x); \mu, \sigma^2\right).
\end{aligned}$$

Note that as $x \to 0^+$, then

$$x^{-1} \exp\left(-(\ln(x) - \mu)^2 / \left(2\sigma^2\right)\right) \to 0^+,$$

since the exponential approaches zero much faster than the reciprocal of x approaches infinity. Thus, since the singularity at zero is removable, we define the exception value of the lognormal density at zero to be

$$\phi_{ln}\left(0; \mu_{ln}, (\sigma_{ln})^2\right) \equiv \phi_{ln}\left(0^+; \mu_{ln}, (\sigma_{ln})^2\right) = 0. \quad \square$$

In the above analytical manipulation of distribution probabilities, the general principles are embodied in the following lemma.

Lemma B.19. ***General Probability Inversion.***
Let X and Y be two random variables with continuous densities $\phi_X(x)$ and $\phi_Y(y)$, respectively. Further, let the dependence between them be given by $X = g(Y)$, where $g(y)$ is continuously differentiable and increasing so that an inverse function f exists, i.e., $y = f(x) = g^{-1}(x)$. Then the corresponding distributions are related by

$$\begin{aligned} \Phi_X(x) &= \text{Prob}[X \le x] = \text{Prob}[g(Y) \le x] \\ &= \text{Prob}[Y \le f(x)] = \Phi_Y(f(x)) \end{aligned} \tag{B.35}$$

and the densities are related by

$$\phi_X(x) = f'(x)\phi_Y(f(x)). \tag{B.36}$$

If, instead, g is strictly decreasing, then

$$\Phi_X(x) = \text{Prob}[Y \ge f(x)] = 1 - \Phi_Y(f(x)) \tag{B.37}$$

and

$$\phi_X(x) = -f'(x)\phi_Y(f(x)). \tag{B.38}$$

Proof. Since f is the inverse function of g, then with $x = g(y)$ and $y = f(x)$, $g(f(x)) = x$ and $g'(y)f'(x) = 1$, using the chain rule and the derivatives are reciprocals of each other. Further, the increasing property of g means f is also increasing, the signs of the derivatives must be the same. So if $x_1 \le x_2$, then $f(x_1) \le f(x_2)$, and the direction of an inequality is preserved upon application of f. In the g decreasing case, the direction is reversed. Thus, (B.35) has been demonstrated in the increasing case. The decreasing case is similar, except for the change in inequality direction and a minor point in converting from probability to distribution function. The probability complement equivalent of $\text{Prob}[Y \ge f(x)]$ would strictly be $1 - \text{Prob}[Y < f(x)]$, but since the densities are continuous the probabilities assigned to an isolated point are zero, i.e., $\text{Prob}[Y < f(x)] = \text{Prob}[Y \le f(x)]$.

The densities follow upon differentiating by the chain rule

$$\Phi'_X(x) = \phi_X(x) = f'(x)\Phi'_Y(f(x)) = f'(x)\phi_Y(f(x))$$

in the increasing case, and the decreasing case is similar except for the minus sign in the density (B.38), which also preserves the nonnegativity of the density, since $-f'(x) > 0$ in the negative case. The factor $\pm f'(x) > 0$ is the density conversion factor in either case. □

Properties B.20. ***Lognormal Distribution Moments.***

- ***Mean:***
$$\mu_{ln} = \mathrm{E}_{ln}[X] = e^{\mu+\sigma^2/2}.$$

- ***Variance:***
$$\sigma_{ln} = \mathrm{Var}_{ln}[X] = (\mu_{ln})^2 \left(e^{\sigma^2} - 1\right).$$

- ***Inverse, normal from lognormal:***
$$\sigma^2 = \ln\left(1 + \sigma_{ln}/\left(\mu_{ln}\right)^2\right)$$
and
$$\mu = \ln\left(\mu_{ln}\right) - \frac{1}{2}\sigma^2.$$

- ***Coefficient of skewness:***
$$\eta_3^{(ln)}[X] = \left(e^{\sigma^2} + 2\right)\sqrt{e^{\sigma^2} - 1}.$$

- ***Coefficient of kurtosis:***
$$\eta_4^{(ln)}[X] = \left(e^{4\sigma^2} + 2e^{3\sigma^2} + 3e^{2\sigma^2} - 3\right).$$

Remark B.21. *The mean formula is justified using the logarithmic transformation, $y = (\ln(x) - \mu)/\sigma$, from lognormal back to normal along with completing the square method in the exponent,*

$$\begin{aligned}\mathrm{E}_{ln}[X] &= \int_0^\infty \frac{\exp(-(\ln(x) - \mu)^2/(2\sigma^2))}{x\sqrt{2\pi\sigma^2}} x\,dx \\ &= \frac{1}{\sqrt{2\pi}} e^{\mu} \int_{-\infty}^{+\infty} e^{-y^2/2} e^{\sigma y}\,dy \\ &= \frac{1}{\sqrt{2\pi}} e^{\mu+\sigma^2/2} \int_{-\infty}^{+\infty} e^{-(y-\sigma^2)^2/2}\,dy = e^{\mu+\sigma^2/2}.\end{aligned}$$

Then the rest of the moments rely on the same techniques.

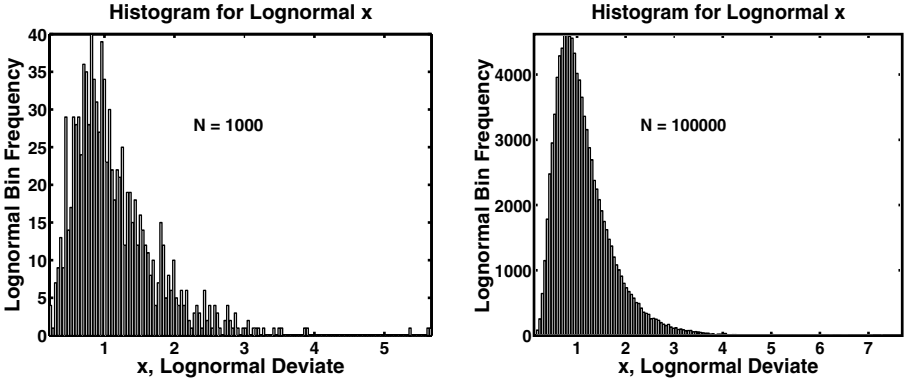
The simulation of the lognormal distribution relies on the fact (B.27) that the lognormal variate is the exponential of a normal variate, i.e., $X_{ln} = \exp(\mu + \sigma X_n)$. Thus the

MATLAB approximation will be the set of simulations

```
y = mu*ones(N,1) + sigma*randn(N,1);
x = exp(y);
```

where again `randn(N,1)` is MATLAB's normal random generator for a sample size of N while the `ones(N,1)` function produces an N-vector of ones preserving the vector form when adding the constant `mu`, with similar constructs in Maple and Mathematica. Equation (B.28) for the density implies that the proper lognormal density will be obtained in theory.

The MATLAB graphical histogram output for two sample sizes, $N = 1000$ and 100,000, both sorted into `nbins=` 150, is given in Figure B.3. The selected normal parameters are $\mu_n = \mu =$ `mu` $= 0.0$ and $\sigma_n = \sigma =$ `sigma` $= 0.5$, corresponding to lognormal parameters $\mu_{ln} \simeq 1.133$ and $\sigma_{ln} \simeq 0.3646$. The percent relative errors in the lognormal mean and standard deviation are respectively -0.56% and -0.60% for $N = 1000$, while the relative errors are -0.085% and -0.30% for the for $N = 100{,}000$ sample size. Again, the larger sample size Figure B.3(b) gives a better qualitative representation of the theoretical shape of the lognormal density $\phi_{ln}(x; \mu_{ln}, \sigma_{ln})$. Both subfigures confirm that the density goes to zero as $x \to 0^+$. See Online Appendix C, Section C.3, for the MATLAB figure code.



(a) Sample size $N = 10^3$.

(b) Sample size $N = 10^5$.

Figure B.3. *Histograms of simulations of the lognormal distribution with mean* $\mu_n = 0$ *and variance* $\sigma_n = 0.5$ *using MATLAB* [210] *normal distribution simulations,* `x = exp(mu*ones(N,1) + sigma*randn(N,1))` *with* 150 *bins for two sample sizes. The histogram for the large sample size of* $N = 10^5$ *in* (b) *exhibits a better approximation to the theoretical lognormal density* $\phi_n(x; 0, 1)$ *than the one in* (a).

B.1.7 Exponential Distribution

The continuous exponential density is closely related to the interarrival time of a Poisson process (discussed in Chapter 1).

Definition B.22. *The **exponential density** is given for some exponential random variate τ_e by*

$$\phi_e(t;\mu) \equiv \frac{1}{\mu} e^{-t/\mu}\,, \quad 0 \le t < +\infty\,, \ \mu > 0, \tag{B.39}$$

*with mean μ, so the exponential distribution is called a **one-parameter distribution**. The explicit form of the **exponential distribution** is*

$$\Phi_e(t;\mu) = \text{Prob}[\tau_e \le t] = \begin{Bmatrix} 1 - e^{-t/\mu}\,, & t \ge 0 \\ 0\,, & t < 0 \end{Bmatrix}. \tag{B.40}$$

Properties B.23. ***Exponential Distribution Moments.***

- ***Conservation of probability**: $\text{E}_e[1] = 1$.*
- ***Mean**: $\mu = \text{E}_e[X]$ by selection of the exponential parameter.*
- ***Variance**: $\sigma^2 = \text{Var}_e[X] = \mu^2$ so the standard deviation is also μ.*
- ***Coefficient of skew**: $\eta_3 = 2$, positive relative to the mean on $[0, \infty)$.*
- ***Coefficient of kurtosis**: $\eta_4 = 9$ or $\eta_4 - 3 = 6$ is the excess value over the normal value.*

Hence, the exponential distribution defines a one-parameter family of distributions with the same mean and standard deviation but also positively skewed by virtue of the semi-infinite domain and leptokurtic with clear pointedness.

Since the **exponential distribution** has such a simple form it can easily be transformed into the uniform distribution for use in practical simulations. Using the **fundamental law** of **transformation** of **probabilities** [230] or as the **inverse transformation method** [97] for transforming the exponential density $\phi_e(x_e;\mu)$ to the standard $(0, 1)$ uniform density $\phi_u(x_u; 0, 1)$,

$$\phi_u\,(x_u; 0, 1) = \phi_e\,(x_e;\mu) \left| \frac{dx_e}{dx_u} \right|. \tag{B.41}$$

The Jacobian sign negative, $dx_e/dx_u < 0$ is chosen because it leads to a faster computational form by eliminating a constant of integration, i.e.,

$$x_e = -\mu \ln\,(x_u)\,, \tag{B.42}$$

which when inverted is

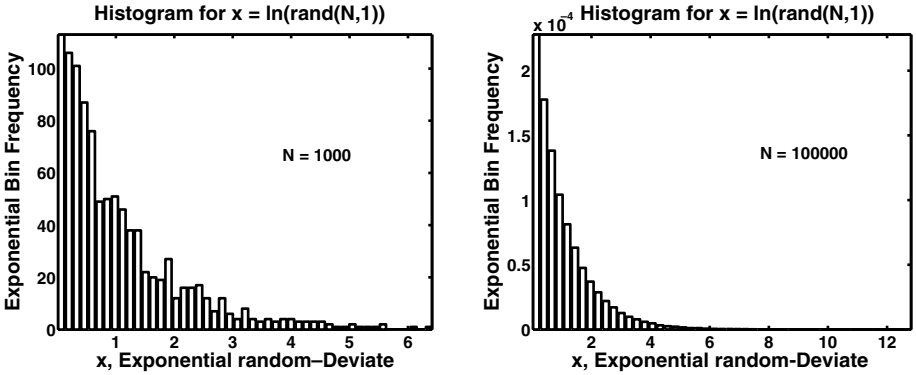
$$x_u = \exp\,(-x_e/\mu)\,. \tag{B.43}$$

A prime prerequisite for random simulations is that the distribution is covered in the transformation, but the order of the covering does not matter so we have

$$\begin{aligned} \Phi_e(x_e;\mu) &= \text{Prob}\,[0 \le X_e \le x_e] \\ &= \text{Prob}\,[\exp\,(-x_e/\mu) \le X_u \le 1] \\ &= 1 - \Phi_u\,(\exp\,(-x_e/\mu)\,;0,1)\,, \end{aligned}$$

which works although the uniform distribution here is covered from right to left starting from 1 while the exponential distribution is covered from left to right starting from $x_e = 0$. The interested reader can check that the general expectation $\mathrm{E}_e[f(X_e)] = \mathrm{E}_u[f(-\mu \ln(X_u))]$ is equivalent for any integrable function f (see Exercise 12).

Hence, `x = −mu * log(rand(N, 1))` leads to a MATLAB exponential random generator producing N-vector output, where `log` is the MATLAB natural logarithm function and `mu` is the input for the mean. The MATLAB graphical output for two sample sizes, $N = 1000$ and 100,000, is given in Subfigures B.4(a) and B.4(b), respectively. The percent relative errors in the mean and standard deviation are, respectively, 7.94% and −0.71% for $N = 1000$, while the errors are 3.81% and −0.54% for the $N = 100{,}000$ sample size. See Online Appendix C, Section C.4, for the MATLAB figure code called `exponential103fig1.m`.



(a) Sample size $N = 10^3$.

(b) Sample size $N = 10^5$.

Figure B.4. *Histograms of simulations of the standard exponential distribution, with mean taken to be* `mu` $= 1$, *using MATLAB's* `hist` *function* [210] *with* 50 *bins for two sample sizes N, generated by* `x = −mu * log(rand(N, 1))` *in MATLAB. The histogram for the large sample size of* $N = 10^5$ *in* (b) *exhibits a better approximation to the standard theoretical exponential density* $\phi_e(x; 1)$.

Remarks B.24.

- *Alternatively, a more direct exponential to uniform transformation could have been selected,*

$$\hat{x}_u = 1 - \exp\left(-\hat{x}_e/\mu\right)$$

with inverse

$$\hat{x}_e = -\mu \ln\left(1 - \hat{x}_u\right), \tag{B.44}$$

but that would not be as numerically efficient for large sample sizes N as (B.42) *which is more often used, since* (B.42) *requires one less floating point operation, not needing to subtract the uniform random sample from* 1 *per sample in* (B.44).

Typically random sample sizes are huge, so good representations of the distribution can be obtained.

- *The probabilistic view of the difference between the two exponential to uniform transformations follows from Lemma* B.19 *on general probability inversion. In the direct case,* $\hat{g}(y) = -\mu \ln(1-y)$ *and* $\hat{f}(x) = 1-\exp(-x/\mu)$*, so* $g'(y) = +\mu/(1-y) > 0$ *for* $0 < y < 1$*. Thus,*

$$\Phi_{\widehat{X}_e}(x) = \Phi_{\widehat{X}_u}(1 - \exp(-x/\mu))$$

by (B.35) *and*

$$\phi_{\widehat{X}_e}(x) = \frac{1}{\mu}\exp(-x/\mu)\phi_{\widehat{X}_u}(1 - \exp(-x/\mu))$$

by (B.36)*, which implies* $\phi_{\widehat{X}_u}(1 - \exp(-x/\mu)) = 1$ *since its coefficient is* $\phi_{\widehat{X}_e}(x)$*. In the more useful case,* $g(y) = -\mu \ln(y)$ *and* $f(x) = \exp(-x/\mu)$*, so* $g'(y) < 0$ *and*

$$\phi_{X_e}(x) = +\frac{1}{\mu}\exp(-x/\mu)\phi_{X_u}(\exp(-x/\mu))$$

by (B.38) *and again* $\phi_{X_u}(\exp(-x/\mu)) = 1$.

B.2 Distributions of Discrete Random Variables

In general, averaged quantities for discrete distributions involve sums rather than integrals used in the continuous distributions. (*Note: The use of the term distribution is different for discrete and continuous cases*.)

Definition B.25. *Let the* ***discrete distribution*** *be*

$$\pi_k = \text{Prob}[X = x_k] \tag{B.45}$$

for some countable set of values $\mathcal{X} = \{x_k | k = 0 : m\}$*, where m could be infinite. (The* $0 : m$ *is MATLAB loop notation.)*

Definition B.26. ***Colon or Loop Notation.***
For compactness, the range of a discrete set will be in the MATLAB colon or loop notation [210, 143] *with* $k = m_1 : m_2$ *denoting that the index k ranges from integers* m_1 *to* m_2 *in steps of unity* (1)*, meaning the same as the loosely defined* $k = m_1, m_1+1, \ldots, m_2-1, m_2$*, assuming* $m_1 < m_2$*. In the case of nonunit steps* Δm*, then* $k = m_1 : \Delta m : m_2$ *is used instead of* $k = m_1, m_1 + \Delta m, \ldots, m_2 - \Delta m, m_2$*, assuming the range* $m_2 - m_1$ *is a positive integer multiple of* Δm.

Properties B.27. ***Discrete Distributions*** $\boldsymbol{\pi_k}$**.**

- ***Nonnegativity****:* $\pi_k \geq 0$.
- ***Conservation of probability****:*

$$\sum_{k=0}^{m} \pi_k = 1. \tag{B.46}$$

The basic definitions in the discrete distribution case for averaged quantities are listed as follows.

Definitions B.28.

- *The **mean** or **expectation** of the **discrete set** $\mathcal{X} = \{x_k | k = 0 : m\}$ is*

$$\mu = \mathrm{E}[\mathcal{X}] \equiv \sum_{k=0}^{m} x_k \pi_k \tag{B.47}$$

for any discretely distributed random variable provided the sum converges absolutely.

- *Similarly, the **expectation of a function $f(X)$ of X** is*

$$\mathrm{E}[f(X)] \equiv \sum_{k=0}^{m} f(x_k)\pi_k \tag{B.48}$$

provided the sum converges absolutely.

Definition B.29. *The **variance** or **mean square deviation** of the discrete set $\mathcal{X}$ is*

$$\mathrm{Var}[\mathcal{X}] \equiv \mathrm{E}[(\mathcal{X} - \mathrm{E}[\mathcal{X}])^2] = \sum_{k=0}^{m} (x_k - \mu)^2 \pi_k \tag{B.49}$$

for any discretely distributed random variable provided the sum converges absolutely, where the set difference $(\mathcal{X} - \mu) \equiv \{x_k - \mu | k = 0 : m\}$ for fixed μ.

B.2.1 Poisson Distribution and Poisson Process

Another important distribution is a **discrete distribution** and is called the Poisson distribution. It is useful for modeling jumps, especially for the jump component of jump-diffusions.

Definition B.30. *The **Poisson distribution** with Poisson variate ν and single Poisson parameter Λ is given by the probabilities*

$$p_k(\Lambda) \equiv \mathrm{Prob}[\nu = k] = e^{-\Lambda} \frac{(\Lambda)^k}{k!} \tag{B.50}$$

for $k = 0, 1, 2, \ldots$ and $\Lambda \geq 0$, expressed as a simple Poisson distribution with continuous parameter Λ which serves as both mean,

$$\mathrm{E}[\nu] = \Lambda, \tag{B.51}$$

and variance,

$$\mathrm{Var}[\nu] = \Lambda, \tag{B.52}$$

of this one-parameter discrete distribution.

The mean and variance can be conveniently computed from the properties of the exponential series,

$$\sum_{k=0}^{\infty} \frac{u^k}{k!} = e^u = \exp(u), \ -\infty < u < +\infty, \tag{B.53}$$

together with its derivatives, such as its first derivative form

$$\sum_{k=0}^{\infty} k \frac{u^k}{k!} = u \frac{d}{du} e^u,$$

which can be used to compute the mean property from

$$\mathrm{E}[\nu] = e^{-\Lambda} \sum_{k=0}^{\infty} k \frac{(\Lambda)^k}{k!}$$

to derive (B.51), and its second derivative form

$$\sum_{k=0}^{\infty} k^2 \frac{u^k}{k!} = \left(u \frac{d}{du} \right)^2 e^u,$$

which can be used with the mean to compute the variance property from

$$\mathrm{Var}[\nu] = e^{-\Lambda} \sum_{k=0}^{\infty} (k - \Lambda)^2 \frac{(\Lambda)^k}{k!}$$

to derive (B.52) upon expanding the square in the sum.

From (B.50), it is simple to deduce that $p_k(0^+) = \delta_{k,0}$, where $\delta_{k,0}$ is defined as follows.

Definition B.31.

$$\delta_{i,j} = \begin{Bmatrix} 1 \ \text{if} \ \ j = i \\ 0 \ \text{if} \ \ j \neq i \end{Bmatrix} \tag{B.54}$$

is the ***Kronecker delta*** *or discrete delta function.*

Figure B.5 the Poisson distribution versus the Poisson counting variable k for four values of the Poisson parameter, $\Lambda = 0.2, 1.0, 2.0$ and 5.0. See Online Appendix C, Section C.5, for the MATLAB figure code. For the smaller parameter value, $\Lambda = 0.2$, the distribution resembles a discretized version of the exponential distribution, while as Λ increases to 2.0 the distribution is beginning to resemble the normal distribution around the peak. For large values of the parameter Λ it can be shown (see Feller [84]) that the Poisson distribution has a normal approximation.

For later use, let $P(t)$ denote the simple Poisson process with linear time-dependent parameter $\Lambda = \lambda t$ as a jump process with unit jumps, hence also characterized as a counting process. It can be shown (see Çinlar [56], for instance) that the $P(t)$ discrete distribution is

$$p_k(\lambda t) \equiv \mathrm{Prob}[P(t) = k] = e^{-\lambda t} \frac{(\lambda t)^k}{k!}. \tag{B.55}$$

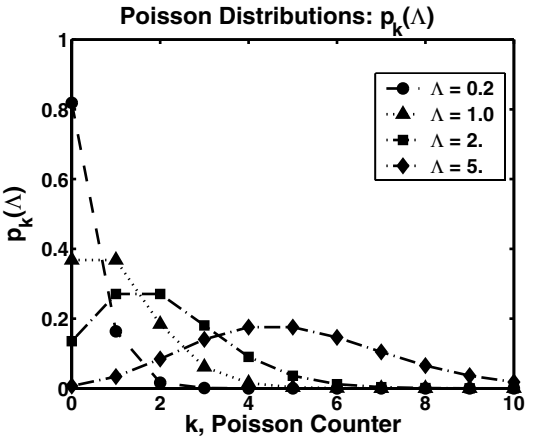


Figure B.5. *Poisson distributions with respect to the Poisson counter variable k for parameter values $\Lambda = 0.2, 1.0, 2.0,$ and 5.0. These represent discrete distributions, but discrete values are connected by dashed, dotted, and dash-dotted lines only to help visualize the distribution form for each parameter value.*

If the random variable T_k is the time of the kth Poisson unit jump for $k = 0 : +\infty$, then time between jumps or interarrival time can be shown to be distributed exponentially,

$$\begin{aligned}\text{Prob}[T_{k+1} - T_k \leq t \mid T_k] &= 1 - \text{Prob}[T_{k+1} - T_k > t \mid T_k] \\ &= 1 - \text{Prob}[P(T_k + t) - P(T_k) = 0 \mid T_k] \\ &= 1 - \text{Prob}[P(t) = 0] \\ &= 1 - e^{-\lambda t} = \Phi_e(t; 1/\lambda), \end{aligned} \tag{B.56}$$

in the first step using conservation of probability to write the probability in terms of one minus the complement, in the second step using the fact that the probability that the interarrival time $\Delta T_k = T_{k+1} - T_k > t$ is the same as the probability that Poisson increment $P(T_k + t) - P(T_k) = 0$, in the third step using the stationarity property that $P(s + t) - P(s)$ and $P(t)$ have the same distribution (to be discussed later), and finally using (B.55) with $k = 0$.

Remark B.32. *The Poisson process is presented in the main chapters of the text, since it serves as the basic process for building the jump component of the jump-diffusion processes.*

B.3 Joint and Conditional Distribution Definitions

In many parts of this book, several properties of joint and conditional distributions will be useful and are summarized for two random variables here. These random variables can be combinations of discrete and continuous random variables, e.g., discrete for jump variables or continuous for diffusion variables. The definition forms are the forms that are useful in this text, but they are not necessarily the most general definitions. Many can be easily

generalized from a couple to multiple random variables. For more general information see the long-standard reference of Feller [85] or the works of Karlin and Taylor [162, 265].

Definitions B.33. ***Jointly Distributed Random Variables.***

- *The **joint probabilities** or **joint distribution functions** of two random variables X and Y depend on whether the random variables are discrete or continuous, leading to three cases:*

  1. ***Two jointly distributed discrete random variables**, X and Y, have the **joint probability** or **joint distribution** function*

  $$\pi_{X,Y}(x_i, y_j) \equiv \text{Prob}[X = x_i, Y = y_j] \tag{B.57}$$

  for specified discrete values x_i and y_j for integers i and j (in general, the discrete sets are assumed to be countable or denumerable) and such values will be assumed with the qualifications given here.

  2. ***Two jointly distributed continuous random variables**, X and Y, have the **joint probability** or **joint distribution** function*

  $$\Phi_{X,Y}(x, y) \equiv \text{Prob}[X \leq x, Y \leq y]. \tag{B.58}$$

  3. ***Two jointly distributed mixed continuous and discrete random variables**, X and Y, have the hybrid **joint probability** or **joint distribution** function*

  $$\Phi_{X,Y}(x, y_j) \equiv \text{Prob}[X \leq x, Y = y_j] \tag{B.59}$$

  for some discrete value y_j.

- *The **joint densities**, if they exist, of two jointly distributed random variables X and Y, are defined as follows.*

  1. ***Two jointly distributed discrete random variables**, X and Y, do not have a **joint density** in the usual way, but for an applied formulation, the generalized functions can be used. (See Section* B.12 *on p.* B52*.)*

  2. ***Two jointly distributed continuous random variables**, X and Y, have the **joint density** if the partial derivatives exist,*

  $$\phi_{X,Y}(x, y) = \frac{\partial^2 \Phi_{X,Y}}{\partial x \partial y}(x, y), \tag{B.60}$$

  and then can be used to calculate the joint distribution using the integral formula

  $$\Phi_{X,Y}(x, y) = \int_{-\infty}^{x} d\xi \int_{-\infty}^{y} d\eta \, \phi_{X,Y}(\xi, \eta). \tag{B.61}$$

3. ***Two jointly distributed mixed continuous and discrete random variables**, X and Y, have the **joint density** if only the x-partial derivative exists,*

$$\phi_{X,Y}(x, y_j) = \frac{\partial \Phi_{X,Y}}{\partial x}(x, y_j). \tag{B.62}$$

*This is a **hybrid density distribution** rather than a strict joint density, but then it can be used to calculate the joint distribution,*

$$\Phi_{X,Y}(x, y_j) = \int_{-\infty}^{x} d\xi \; \phi_{X,Y}(\xi, y_j), \tag{B.63}$$

for some discrete value y_j.

- *The **marginal distributions** in one of two random variables X and Y are defined by summing or integrating over the other random variable:*

1. ***Two jointly distributed discrete random variables**, X and Y, have the **marginal distributions***

$$\pi_X(x_i) = \sum_{j=1}^{\infty} \pi_{X,Y}(x_i, y_j), \tag{B.64a}$$

$$\pi_Y(y_j) = \sum_{i=1}^{\infty} \pi_{X,Y}(x_i, y_j). \tag{B.64b}$$

2. ***Two jointly distributed continuous random variables**, X and Y, have the **marginal distributions***

$$\Phi_X(x) = \lim_{y \to +\infty} \Phi_{X,Y}(x, y) = \int_{-\infty}^{x} d\xi \int_{-\infty}^{+\infty} d\eta \; \phi_{X,Y}(\xi, \eta), \tag{B.65a}$$

$$\Phi_Y(y) = \lim_{x \to +\infty} \Phi_{X,Y}(x, y) = \int_{-\infty}^{y} d\eta \int_{-\infty}^{+\infty} d\xi \; \phi_{X,Y}(\xi, \eta), \tag{B.65b}$$

provided the limits exist.

3. ***Two jointly distributed mixed continuous and discrete random variables**, X and Y, have the **marginal distributions***

$$\Phi_X(x) = \int_{-\infty}^{x} d\xi \sum_{j=1}^{\infty} \phi_{X,Y}(\xi, y_j), \tag{B.66a}$$

$$\pi_Y(y_j) = \int_{-\infty}^{+\infty} d\xi \; \phi_{X,Y}(\xi, y_j), \tag{B.66b}$$

provided the limit exists.

- *The **marginal densities** of two random variables, X and Y, are defined as*

1. ***Two jointly distributed discrete random variables**, X and Y, do not have **marginal densities** in the usual way, but for an applied formulation, the generalized functions can be used. (See Section* B.12 *on p.* B52.*)*
2. ***Two jointly distributed continuous random variables**, X and Y, have the **marginal densities***

$$\phi_X(x) = \int_{-\infty}^{+\infty} d\eta\, \phi_{X,Y}(x, \eta), \tag{B.67a}$$

$$\phi_Y(y) = \int_{-\infty}^{+\infty} d\xi\, \phi_{X,Y}(\xi, y). \tag{B.67b}$$

3. ***Two jointly distributed mixed continuous and discrete random variables**, X and Y, have the **marginal density** for the continuous random variable X,*

$$\phi_X(x) = \sum_{j=1}^{\infty} \phi_{X,Y}(x, y_j), \tag{B.68}$$

and the marginal distribution $\pi_Y(y_j)$ is given in (B.66b)*.*

- *The **expectation function** $f(X, Y)$ **of joint random variables**, X and Y, is defined as follows:*

1. ***Two jointly distributed discrete random variables**, X and Y, have the **joint expectation** of $f(X, Y)$, providing the sums or integrals exist:*

$$\mathrm{E}_{X,Y}[f(X, Y)] = \sum_{i=1}^{+\infty} \sum_{j=1}^{+\infty} f(x_i, y_j)\pi_{X,Y}(x_i, y_j). \tag{B.69}$$

2. ***Two jointly distributed continuous random variables**, X and Y, have the **joint expectation** of $f(X, Y)$*

$$\mathrm{E}_{X,Y}[f(X, Y)] = \int_{-\infty}^{+\infty} d\xi \int_{-\infty}^{+\infty} d\eta\, f(\xi, \eta)\phi_{X,Y}(\xi, \eta). \tag{B.70}$$

3. ***Two jointly distributed mixed continuous and discrete random variables**, X and Y, have the **joint expectation***

$$\mathrm{E}_{X,Y}[f(X, Y)] = \int_{-\infty}^{+\infty} d\eta \sum_{j=1}^{\infty} f(\xi, y_j)\phi_{X,Y}(\xi, y_j), \tag{B.71}$$

where $\phi_{X,Y}(x, y_j)$ is the hybrid density distribution given by (B.62)*.*

- *The **covariance of two jointly distributed random variables**, X and Y, for all three cases, is defined as*

$$\mathrm{Cov}[X, Y] \equiv \mathrm{E}_{X,Y}[(X - \mathrm{E}_X[X])(Y - \mathrm{E}_Y[Y])], \tag{B.72}$$

provided the expectations exist. Hence,

$$\mathrm{Cov}[X, Y] = \mathrm{E}_{X,Y}[X \cdot Y] - \mathrm{E}_X[X] \cdot \mathrm{E}_Y[Y]. \tag{B.73}$$

- *The **variance of a sum or difference of two random variables**, X and Y,*

$$\mathrm{Var}[X \pm Y] = \mathrm{Var}_X[X] \pm 2\mathrm{Cov}[X, Y] + \mathrm{Var}_Y[Y], \tag{B.74}$$

by writing the variance of the sum-difference as expectations and collecting terms into a covariance using (B.72) *and using the definition of variance twice for the remaining terms, i.e.,*

$$\begin{aligned}\mathrm{Var}[X \pm Y] &= \mathrm{E}[(X - \mathrm{E}[X] \pm (Y - \mathrm{E}[Y]))^2] \\ &= \mathrm{Var}_X[X] \pm 2\mathrm{Cov}[X, Y] + \mathrm{Var}_Y[Y].\end{aligned}$$

Remarks B.34.

- *The subscript on the expectation symbol is often omitted but can be used in multivariate expectation to precisely specify which variable or variables are the arguments of the expectation operator and to avoid confusion.*
- *The integral notations are equivalent:*

$$\int_{x_1}^{x_2} dx \int_{y_1}^{y_2} dy f(x, y) = \int_{x_1}^{x_2} \int_{y_1}^{y_2} f(x, y) dy dx;$$

the former, having the element of integration following the integration sign, makes it easy to see the order of integration and which limits of integration go with what elements of integration.

Definitions B.35. ***Independently Distributed Random Variables.***

- *The **joint distribution of two independent random variables**, X and Y, is the product of the marginal distributions:*

  1. ***Two discrete random variables**, X and Y, are **independent** if their joint distribution is*

$$\pi_{X,Y}(x_i, y_j) = \pi_X(x_i) \cdot \pi_Y(y_j). \tag{B.75}$$

  2. ***Two continuous random variables**, X and Y, are **independent** if their joint distribution is*

$$\Phi_{X,Y}(x, y) = \Phi_X(x) \cdot \Phi_Y(y). \tag{B.76}$$

  3. ***Two mixed continuous discrete random variables**, X and Y, are **independent** if their joint distribution is*

$$\Phi_{X,Y}(x, y_j) = \Phi_X(x) \cdot \pi_Y(y_j). \tag{B.77}$$

- *The **joint density of two independent random variables**, X and Y, is the product of the marginal densities:*

1. ***Two discrete random variables**, X and Y, do not have a **joint density** in the usual way.*
2. ***Two continuous random variables**, X and Y, are **independent** if their joint distribution is*

$$\phi_{X,Y}(x, y) = \phi_X(x) \cdot \phi_Y(y). \tag{B.78}$$

3. ***Two mixed continuous and discrete random variables**, X and Y, are **independent** if their hybrid density distribution is*

$$\phi_{X,Y}(x, y_j) = \phi_X(x) \cdot \pi_Y(y_j), \tag{B.79}$$

assuming densities exist where relevant.

- *The **joint expectation of the product** $f(X) \cdot g(Y)$ in two independent random variables, X and Y, is the **product of the expectations**,*

$$\mathrm{E}_{X,Y}[f(X) \cdot g(Y)] = \mathrm{E}_X[f(X)] \cdot \mathrm{E}_Y[g(Y)], \tag{B.80}$$

covering all three cases.

- *The **covariance of two independent random variables**, X and Y, is **zero**,*

$$\mathrm{Cov}[X, Y] \equiv \mathrm{E}[(X - E[X])(Y - E[Y])] = 0, \tag{B.81}$$

since by the separability of the expectation in (B.80)*,*

$$\mathrm{Cov}[X, Y] = \mathrm{E}_X[(X - E[X])] \cdot \mathrm{E}_Y[(Y - E[Y])] = 0 \cdot 0 = 0.$$

Note that the converse is not true. If $\mathrm{Cov}[X, Y] = 0$*, then the random variables are not necessarily independent.*

B.3.1 Conditional Distributions and Expectations

Definitions B.36.

- *The **conditional probability** and **conditional distribution** of the random variable X, conditioned on the random variable Y, are defined as follows:*

1. ***If X and Y are both discrete random variables**,*

$$\pi_{X|Y}(x_i|y_j) \equiv \mathrm{Prob}[X = x_i \mid Y = y_j] = \frac{\mathrm{Prob}[X = x_i, Y = y_j]}{\mathrm{Prob}[Y = y_j]}, \tag{B.82}$$

provided the marginal distribution $\pi_Y(y_j) = \mathrm{Prob}[Y = y_j] \neq 0$ *from* (B.64)*.*

2. ***If X and Y are both continuous random variables**,*

$$\Phi_{X|Y}(x|y) \equiv \mathrm{Prob}[X \leq x \mid Y = y] = \frac{\int_{-\infty}^{x} d\xi \phi_{X,Y}(\xi, y)}{\phi_Y(y)}, \tag{B.83}$$

provided marginal density $\phi_Y(y) \neq 0$ *from* (B.67)*. See Karlin and Taylor* [162].

Remarks B.37.

- *Since we can write*

$$\text{Prob}[Y \in [y, y+dy]] \overset{dy}{=} \phi_Y(y)dy,$$

i.e., in precision-dy, the formula (B.83) *can be rewritten in probabilities,*

$$\text{Prob}[X \le x \mid Y = y] = \frac{\text{Prob}[X \le x, Y \in [y, y+dy]]}{\text{Prob}[Y \in [y, y+dy]]},$$

provided $\text{Prob}[Y \in [y, y+dy]] > 0$.

- *Regarding* (B.83)*, note that if Y is a continuous random variable, then* $\text{Prob}[Y = y] = 0$ *since a single point has no probability mass with*

$$\lim_{\delta \to 0} \int_y^{y+\delta} \phi_Y(\eta)d\eta = 0.$$

- *The reader can confirm the consistency of these conditional probability formulas when X and Y are independent random variables.*

3. ***If X is a continuous and Y is a discrete random variable****,*

$$\Phi_{X|Y}(x|y_j) \equiv \text{Prob}[X \le x \mid Y = y_j] = \frac{\text{Prob}[X \le x, Y = y_j]}{\text{Prob}[Y = y_j]} = \frac{\int_{-\infty}^{x} d\xi \phi_{X,Y}(\xi, y_j)}{\text{Prob}[Y = y_j]}, \tag{B.84}$$

provided marginal distribution $\pi_Y(y_j) = \text{Prob}[Y = y_j] \neq 0$ *from* (B.66b)*, where* $\phi_{X,Y}(\xi, y_j)$ *is the* ***hybrid density distribution*** *in* (B.62)*.*

- ***Iterated probability*** *uses the definitions of conditional probability in reverse to evaluate* ***joint probability*** *for the random variables X and Y:*

1. ***If X and Y are both discrete random variables****,*

$$\text{Prob}[X = x_i, Y = y_j] = \text{Prob}[X = x_i \mid Y = y_j] \cdot \text{Prob}[Y = y_j], \tag{B.85}$$

provided the conditional distribution $\text{Prob}[X = x_i \mid Y = y_j]$ *exists.*

2. ***If X and Y are both continuous random variables****,*

$$\text{Prob}[X \le x, Y \in [y, y+dy]] = \int_{-\infty}^{x} d\xi \phi_{X,Y}(\xi, y)dy = \text{Prob}[X \le x \mid Y = y] \cdot \phi_Y(y)dy, \tag{B.86}$$

provided the conditional distribution $\text{Prob}[X \le x \mid Y = y]$ *exists, but if not then* $\phi_Y(y) = 0$ *should cover the case.*

3. ***If X is a continuous and Y is a discrete random variable,***

$$\text{Prob}[X \le x, Y = y_j] = \text{Prob}[X \le x \mid Y = y_j] \cdot \text{Prob}[Y = y_j], \tag{B.87}$$

provided marginal distribution $\pi_Y(y_j) = \text{Prob}[Y = y_j] \neq 0$ from (B.66b)*, where $\phi_{X,Y}(\xi, y_j)$ is the hybrid density distribution in* (B.62)*.*

Remark B.38. *These forms are convenient for decomposing joint probability calculations into simpler parts.*

- *The **conditional density** is*

$$\phi_{X|Y}(x|y) = \frac{\partial \Phi_{X|Y}(x|y)}{\partial x} \tag{B.88}$$

provided X is a continuous random variable and Y is either continuous or discrete.

- *The **conditional expectation** of X given $Y = y$ is defined as*

$$\text{E}_X[X|Y = y] = \int_{-\infty}^{+\infty} x \phi_{X|Y}(x|y) dx \tag{B.89}$$

provided X is a continuous random variable and Y is either continuous or discrete; else

$$\text{E}_X[X|Y = y_j] = \sum_{i=1}^{\infty} x_i \pi_{X|Y}(x_i|y_j) \tag{B.90}$$

when both X and Y are discrete random variables with a similar form for $\text{E}_X[X|Y = y]$ if X is discrete but Y is continuous.

- *Similarly, the **expectation for a function $f(X, Y)$** given $Y = y$ is*

$$\text{E}_X[f(X, Y)|Y = y] = \int_{-\infty}^{+\infty} f(x, y) \phi_{X|Y}(x|y) dx$$

provided X is a continuous random variable and Y is either continuous or discrete; else

$$\text{E}_X[f(X, Y)|Y = y_j] = \sum_{i=1}^{\infty} f(x_i, y_j) \pi_{X|Y}(x_i|y_j)$$

when both X and Y are discrete random variables.

Properties B.39. ***Conditional Expectations.***

- $\text{E}[f(X)|X] = f(X)$ *for some function* f.
- $\text{E}_Y[\text{E}_{X|Y}[X|Y]] = \text{E}_{X,Y}[X]$, *but* $\text{E}_Y[\text{E}_{X|Y}[X|Y]] = E_X[X]$ *if X and Y are independent random variables.*

- $\mathrm{E}[c_1X_1 + c_2X_2|Y] = c_1\mathrm{E}[X_1|Y] + c_2\mathrm{E}[X_2|Y]$ *provided the conditional expectations exist for random variables* Y *and* X_i*, and constants* c_i*, for* $i = 1 : 2$*, i.e., the conditional expectation is a linear operation.*

- *If* X *and* Y *are random variables, then the* ***iterated expectation*** *is*

$$\mathrm{E}_{X,Y}[f(X,Y)] = \mathrm{E}_Y\left[\mathrm{E}_X[f(X,Y)|Y]\right] \tag{B.91}$$

provided the expectations exist, i.e., that $f(x, y)$ *is sufficiently integrable with respect to any density. This is also a general form of the law of total probability given in the next section.*

Proof. *In the case that* X *and* Y *are both continuous random variables, the justification is built upon the basic definition of the conditional distribution in* (B.83) *which leads to the conditional density according to* (B.88) *upon differentiation,*

$$\phi_{X|Y}(x|y) = \phi_{X,Y}(x, y)/\phi_Y(y),$$

assuming $\phi_Y(y) > 0$*. Further,* $\phi_Y(y) > 0$ *will be assumed on* $-R \le y \le R$ *for some* $R > 0$*, since* $\phi_Y(y) \to 0^+$ *as* $y \to +\infty$ *for conservation of probability through integrability at infinity. For convenience, the limit as* $R \to +\infty$ *will be ignored in the following formally justifying chain of equations:*

$$\begin{aligned}
\mathrm{E}_{X,Y}[f(X,Y)] &= \int_{-\infty}^{+\infty} dy \int_{-\infty}^{+\infty} dx \phi_{X,Y}(x, y) f(x, y) \\
&= \int_{-\infty}^{+\infty} dy \int_{-\infty}^{+\infty} dx \left(\phi_{X|Y}(x|y)\phi_Y(y)\right) f(x, y) \\
&= \int_{-\infty}^{+\infty} dy \phi_Y(y) \int_{-\infty}^{+\infty} dx \phi_{X|Y}(x|y) f(x, y) \\
&= \mathrm{E}_Y\left[\mathrm{E}_X[f(X,Y)|Y]\right].
\end{aligned}$$

The other random variable cases are similar with sums where discrete random variables are concerned. □

- *If* X *and* Y *are independent, then* $\mathrm{E}[X|Y] = \mathrm{E}[X]$ *and in general*

$$\mathrm{E}[f(X)g(Y)|Y] = \mathrm{E}[f(X)]g(Y),$$

provided the expectations exist.

See Mikosch [209] for more conditional expectation properties in a more abstract setting.

B.3.2 Law of Total Probability

Properties B.40. ***Law of Total Probability.***

- *When X is a* ***discrete random variable*** *and given a countable set of* ***mutually independent discrete random variables****, $\{Y_1, Y_2, \ldots, Y_i, \ldots\}$, and the conditional probabilities* $\text{Prob}[X|Y_i]$ *for $i = 1, 2, \ldots$, then the* ***law of total probability*** *(see Taylor and Karlin* [265]*) in this completely discrete case is*

$$\text{Prob}[X] = \sum_{i=1}^{\infty} \text{Prob}[X|Y_i]\text{Prob}[Y_i], \tag{B.92}$$

  i.e., an extension of the law of additive probabilities for disjoint events.

- *When X is a* ***continuous random variable****, the corresponding law of total probability for the* ***probability distribution*** *$\Phi_X(x)$ is*

$$\Phi_X(x) = \sum_{i=1}^{\infty} \Phi_{X|Y}(x|Y_i)\text{Prob}[Y_i]. \tag{B.93}$$

- *Providing the density exists in the continuous random variable case, the corresponding law of total probability for the* ***probability density*** *of $\phi_X(x)$ is*

$$\phi_X(x) = \sum_{i=1}^{\infty} \phi_{X|Y}(x|Y_i)\text{Prob}[Y_i]. \tag{B.94}$$

- *Finally, the* ***expectation*** *corresponding to the law of total probability is*

$$\text{E}[f(X)] = \sum_{i=1}^{\infty} \text{E}_X[f(X)|Y_i]\text{Prob}[Y_i] \tag{B.95}$$

  for either the discrete or continuous X case and assuming the expectations of $f(X)$ exist. This is a special case of the ***iterated expectations*** *given previously in* (B.91)*.*

Example B.41. *An interesting financial example of* (B.95) *derived from* [265] *is the set of statistics for the daily stock price return observed on a transaction by transaction basis. Let the transaction price return be $\xi_i = \Delta S_i = S_{i+1} - S_i$, where S_i is the price of the ith transaction, with S_0 the initial price such as that from the previous day's closing. Suppose the returns are independent identically distributed (IID) random variables with common mean $\text{E}_\xi[\xi_i] = \mu$ and variance $\text{Var}_\xi[\xi_i] = \sigma^2$. Assume the current total daily stock return after N transactions is*

$$X = \sum_{i=0}^{N} \xi_i,$$

where N is Poisson distributed, i.e., N is a counting process such that $\text{Prob}[N = n] = p_n(\Lambda)$ *with* Λ *being the Poisson parameter in* (B.50)*, so* $\text{E}_N[N] = \Lambda = \text{Var}_N[N]$*. Starting from the law of total probability, the expectation of the daily return is decomposed as*

$$\begin{aligned}
\text{E}_X[X] &= \sum_{n=0}^{\infty} \text{E}_{X|N}[X|N = n]p_n(\Lambda) = \sum_{n=0}^{\infty} \text{E}_{\xi|N}\left[\sum_{i=0}^{N} \xi_i \,\middle|\, N = n\right] p_n(\Lambda) \\
&= \sum_{n=0}^{\infty} \text{E}_{\xi}\left[\sum_{i=0}^{n} \xi_i\right] p_n(\Lambda) = \sum_{n=0}^{\infty}\sum_{i=0}^{n} \text{E}_{\xi}[\xi_i]p_n(\Lambda) \\
&= \sum_{n=0}^{\infty}\sum_{i=0}^{n} \mu p_n(\Lambda) = \mu \sum_{n=0}^{\infty} n p_n(\Lambda) = \mu\Lambda,
\end{aligned}$$

where the independence and identically distributed properties of the ξ_i random variables, as well as the mean properties of N, have been used.

The variance of X is more complicated but follows from similar techniques, except that terms are collected by ***completing the square*** *in the ith return deviation from the mean $(\xi_i - \mu)$ with several applications of the independence assumption,*

$$\begin{aligned}
\text{Var}_X[X] &= \text{E}_X[(X - \Lambda\mu)^2] = \sum_{n=0}^{\infty} \text{E}_{\xi|N}\left[\left(\sum_{i=0}^{N} \xi_i - \Lambda\mu\right)^2 \,\middle|\, N = n\right] p_n(\Lambda) \\
&= \sum_{n=0}^{\infty} \text{E}_{\xi}\left[\left(\sum_{i=0}^{n} (\xi_i - \mu) + (n - \Lambda)\mu\right)^2\right] p_n(\Lambda) \\
&= \sum_{n=0}^{\infty} \text{E}_{\xi}\left[\sum_{i=0}^{n}\sum_{j=0}^{n} (\xi_i - \mu)(\xi_j - \mu) + 2(n - \Lambda)\mu \sum_{i=0}^{n} (\xi_i - \mu) + (n - \Lambda)^2\mu^2\right] p_n(\Lambda) \\
&= \sum_{n=0}^{\infty} \text{E}_{\xi}\left[\sum_{i=0}^{n} (\xi_i - \mu)^2 + \sum_{i=0}^{n}\sum_{j\neq i}^{n} (\xi_i - \mu)(\xi_j - \mu) + (n - \Lambda)^2\mu^2\right] p_n(\Lambda) \\
&= \sum_{n=0}^{\infty}\left[\sum_{i=0}^{n} \text{E}_{\xi}[(\xi_i - \mu)^2] + (n - \Lambda)^2\mu^2\right] p_n(\Lambda) = \sum_{n=0}^{\infty}[n\sigma^2 + (n - \Lambda)^2\mu^2]p_n(\Lambda) \\
&= \Lambda\sigma^2 + \Lambda\mu^2 = \Lambda(\sigma^2 + \mu^2),
\end{aligned}$$

such that the ith return variance is augmented by the mean squared.

B.4 Probability Distribution of a Sum: Convolutions

Combinations of random variables play an important role in the analysis of stochastic processes, especially in the sum of two stochastic processes. Consider the following result.

Theorem B.42. ***Convolution for Sums of Random Variables.***
If X and Y are independent random variables with densities $\phi_X(x)$ and $\phi_Y(y)$, respectively, then the ***distribution of the sum*** *is*

$$\Phi_{X+Y}(z) \equiv \text{Prob}[X + Y \leq z] = \int_{-\infty}^{+\infty} \Phi_Y(z - x)\phi_X(x)dx, \tag{B.96}$$

provided the integral exists, where

$$\Phi_Y(y) = \int_{-\infty}^{y} \phi_Y(\eta)d\eta.$$

Proof. By the independence of the variables X and Y, the joint density is separable, $\phi_{X+Y}(x, y) = \phi_X(x)\phi_Y(y)$. Thus, using the properties of the Heaviside step function,

$$H(x) = \begin{Bmatrix} 0, & x < 0 \\ 1, & x \geq 0 \end{Bmatrix}, \tag{B.97}$$

then

$$\begin{aligned}
\text{Prob}[X + Y \leq z] &= \text{E}_{X+Y}[H(z - X - Y)] \\
&= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} H(z - x - y)\phi_X(x)\phi_Y(y)dydx \\
&= \int_{-\infty}^{+\infty} \int_{-\infty}^{z-x} \phi_Y(y)dy\phi_X(x)dx \\
&= \int_{-\infty}^{+\infty} \Phi_Y(z - x)\phi_X(x)dx \\
&= \text{E}_X[\Phi_Y(z - X)],
\end{aligned}$$

where iterated integrals have been freely interchanged by the theorem of Fubini, which asserts that if an integral exists as a two-dimensional integral, then the two iterative integrals can be interchanged, i.e., the order of integration does not matter. **Fubini's theorem** is often used in probability theory [85, 169]. □

Since it has been assumed that the densities exist, then differentiation of the sides of the equation in (B.96), but under the integral sign for those on the right, yields the formula for the probability density of a sum, as follows.

Corollary B.43.

$$\phi_{X+Y}(z) = \int_{-\infty}^{+\infty} \phi_Y(z - x)\phi_X(x)dx. \tag{B.98}$$

The particular functional product forms of (B.96), (B.98) are called convolutions [85].

Definition B.44. *Let the* ***convolution*** *of a distribution or density $f(y)$ and a density $\phi(x)$ be*

$$(f * \phi)(z) \equiv \int_{-\infty}^{+\infty} f(z-x)\phi(x)dx \tag{B.99}$$

provided the integral exists.

Consequently, we have the following properties, including the reformulation of the above sum rules.

Properties B.45. ***Convolutions.***

- *The* ***convolution of densities is symmetric*** $(f * \phi)(z) = (\phi * f)(z)$ *upon change of variables in the integrand.*
- $\phi_{X+Y}(z) = (\phi_Y * \phi_X)(z) = (\phi_X * \phi_Y)(z)$.
- $\Phi_{X+Y}(z) = (\Phi_X * \phi_Y)(z) = (\Phi_Y * \phi_X)(z)$.
- *The form for n mutually independent random variables, all with given densities, is*

$$\begin{aligned} \phi_{X_1+X_2+\cdots+X_n}(z) &= (\phi_{X_1} * \phi_{X_2} * \cdots * \phi_{X_n})(z) \\ &= \left\{ \begin{array}{l} ((\cdots((\phi_{X_1} * \phi_{X_2}) * \phi_{X_3}) \cdots * \phi_{X_{n-1}}) * \phi_{X_n})(z) \\ (\phi_{X_1} * (\phi_{X_2} * (\phi_{X_3} * \cdots (\phi_{X_{n-1}} * \phi_{X_n}) \cdots)))(z) \end{array} \right\}, \end{aligned} \tag{B.100}$$

  the latter forms depending on whether the convolution expansion is from the right or from the left, respectively.

Remark B.46. *The particular form depends on which particular inductive definition is used, i.e., the right and left convolution expansion forms, respectively, are*

$$\phi_{\sum_{i=1}^{n+1} X_i}(z) = \left\{ \begin{array}{l} \left(\phi_{\sum_{i=1}^{n} X_i} * \phi_{X_{n+1}}\right)(z) \\ \\ \left(\phi_{X_1} * \phi_{\sum_{i=2}^{n+1} X_i}\right)(z) \end{array} \right\},$$

as can be shown by mathematical induction.

Lemma B.47. ***Convolution of Normal Densities is Normal.***
If X and Y are normally distributed random variables, with probability densities $\phi_X(x) = \phi_n(x; \mu_x, \sigma_x^2)$ and $\phi_Y(y) = \phi_n(y; \mu_y, \sigma_y^2)$, respectively, then, letting $Z = X + Y$,

$$\begin{aligned} \phi_Z(z) &= (\phi_X * \phi_Y)(z) \\ &= \int_{-\infty}^{+\infty} \phi_X(z-y)\phi_Y(y)dy \\ &= \phi_n(z; \mu_x + \mu_y, \sigma_x^2 + \sigma_y^2). \end{aligned} \tag{B.101}$$

Maple Proof.

```
> phi:=(x,m,s)->exp(-(x-m)^2/(2*s^2))/sqrt(2*pi*s^2);
```

$$\phi := (x, m, s) \rightarrow \frac{e^{(-1/2\,\frac{(x-m)^2}{s^2})}}{\sqrt{2\,\pi\,s^2}}$$

```
> interface(showassumed=0); assume(sx>0); assume(sy>0);
> phi_Z:=simplify(int(phi(z-y,mx,sx)*phi(y,my,sy),
> y=-infinity..infinity));
```

$$phi\_Z := \frac{1}{2}\,\frac{e^{\left(-\frac{(z-mx-my)^2}{2\,(sy^2+sx^2)}\right)}\,\sqrt{2}\,\sqrt{\pi}}{\pi\,\sqrt{sy^2+sx^2}}$$

For more general results see Exercises 16, 17, and 18.

B.5 Characteristic Functions

Often it is convenient to transform distributions or densities so that moments can be generated more systematically, leading to a class of generating functions. Here, the emphasis will be on one class that is more useful for both positive and negative random variables, called *characteristic functions*.

Definition B.48. *The characteristic function of a random variable X is defined in general as*

$$\mathcal{C}_X(u) \equiv \mathrm{E}\left[e^{iuX}\right], \tag{B.102}$$

where $i = \sqrt{-1}$ is the imaginary unit constant, u is the characteristic function argument, assumed real here, the complex exponential is

$$e^{iux} = \cos(ux) + i\,\sin(ux)$$

by Euler's formula with complex conjugate $z^* = (x + iy)^* \equiv x - iy$ so

$$(\exp(iux))* = \exp(-iux)$$

and modulus (absolute value) $|z| \equiv \sqrt{(x^2 + y^2)}$ so

$$\left|e^{iux}\right| = \sqrt{cos^2(ux) + sin^2(ux)} = 1$$

according to Pythagoras' theorem (summarizing almost all the complex algebra that will be needed here). Only three main forms for $\mathcal{C}_X(u)$ are listed here:

- *If X is a continuous random variable with proper probability distribution function $\Phi_X(x)$, then*

$$\mathcal{C}_X(u) = \int_{-\infty}^{\infty} e^{iux} d\Phi_X(x), \tag{B.103}$$

which is called a Fourier–Stieltjes transform.

- *If X is a continuous random variable and there exists a density corresponding to $\Phi_X(x)$, then*

$$\mathcal{C}_X(u) = \int_{-\infty}^{\infty} e^{iux}\phi_X(x)dx, \tag{B.104}$$

which is just an ordinary Fourier transform.

- *If X is a discrete random variable with distribution function $\pi_k = \text{Prob}[X = x_k]$ for all nonnegative integers k, then*

$$\mathcal{C}_X(u) = \sum_{k=0}^{\infty} \pi_k e^{iux_k}, \tag{B.105}$$

which is called a Fourier exponential series.

Properties B.49. ***Characteristic Functions.***

- ***Moment properties**:*
  - $\mathcal{C}_X(0) = 1$ *by conservation of probability;*
  - $\mathcal{C}'_X(0) = \text{E}_X[X]$ *by differentiation of integrand;*
  - *By induction for $k = 0, 1, 2, \ldots,$*

$$\frac{d^k\mathcal{C}_X}{du^k}(0) = i^k\text{E}_X\left[X^k\right].$$

- ***Relationship to standard generating function**:*

$$G_X(z) \equiv \text{E}\left[z^X\right], \tag{B.106}$$

so letting $z^x = e^{iux}$, then $z = e^{iu}$, $u = -i\ln(z)$, $G_X(z) = \mathcal{C}_X(-i\ln(z))$, and $\mathcal{C}_X(u) = G_X(e^{iu})$.

- ***Complex properties**: By Euler's formula, the resolution into real and imaginary parts,*

$$\mathcal{C}_X(u) = C_X(u) + iS_X(u),$$

where the real part is the cosine transform

$$C_X(u) = \int_{-\infty}^{\infty} \cos(ux)\phi_X(x)dx$$

and the imaginary part is the sine transform

$$S_X(u) = \int_{-\infty}^{\infty} \sin(ux)\phi_X(x)dx,$$

so the complex conjugate is

$$\mathcal{C}^*_X(u) = C_X(u) - iS_X(u).$$

- ***Reality and symmetric densities**: The characteristic function $\mathcal{C}_X(u)$ is real if and only if the corresponding probability density is symmetric, i.e., $\phi_X(-x) = \phi_X(x)$. Note that $\mathcal{C}_X(u)$ is real if the imaginary part $S_X(u)$ is zero and $\mathcal{C}_X(-u) = \mathcal{C}_X^*(u) = C_X(u) - iS_X(u)$ ($\exp(-iux) = \cos(ux) - i\sin(ux)$), so*

$$\begin{aligned} iS_X(u) &= 0.5(\mathcal{C}_X(u) - \mathcal{C}_X(-u)) = 0.5\int_{-\infty}^{\infty}\left(e^{iux} - e^{-iux}\right)\phi_X(x)dx \\ &= 0.5\int_{-\infty}^{\infty} e^{iux}\left(\phi_X(x) - \phi_X(-x)\right)dx, \end{aligned}$$

then $\phi_X(x)$ symmetric implies $S_X(u) = 0$ and $S_X(u) = 0$ implies $\phi_X(x)$ symmetric.

- ***Upper bound**: $|\mathcal{C}_X(u)| \leq 1$, since by Euler's formula and trigonometric identities*

$$\begin{aligned} |\mathcal{C}_X(u)|^2 &= \left(\int_{-\infty}^{\infty}\cos(ux)\phi_X(x)dx\right)^2 + \left(\int_{-\infty}^{\infty}\sin(ux)\phi_X(x)dx\right)^2 \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(\cos(ux)\cos(uy) + \sin(ux)\sin(uy))\phi_X(x)\phi_X(y)dxdy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cos(u(x-y))\phi_X(x)\phi_X(y)dxdy \\ &\leq \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\phi_X(x)\phi_X(y)dxdy = 1. \end{aligned}$$

- ***Sums of random variables and convolutions**: Let $\{X_k; k = 1:N\}$ be a set of independent random variables; then $\mathcal{C}_{X_1+X_2}(u) = \mathcal{C}_{X_1}(u)\cdot\mathcal{C}_{X_1}(u)$ since by the convolution property* (B.98)

$$\begin{aligned} \mathcal{C}_{X_1+X_2}(u) &= \int_{-\infty}^{\infty} e^{iux}\phi_{X_1+X_2}(x)dx = \int_{-\infty}^{\infty} e^{iux}\left(\phi_{X_1} * \phi_{X_2}\right)(x)dx \\ &= \int_{-\infty}^{\infty} e^{iux}\int_{-\infty}^{\infty}\phi_{X_2}(x-y)\phi_{X_1}(y)dydx \\ &= \int_{-\infty}^{\infty} e^{iuy}\phi_{X_1}(y)\int_{-\infty}^{\infty} e^{iu(x-y)}\phi_{X_2}(x-y)dxdy \\ &= \mathcal{C}_{X_1}(u)\cdot\mathcal{C}_{X_1}(u), \end{aligned}$$

assuming integral interchange is permitted. Further, for a set of of N independent random variables,

$$\mathcal{C}_{\sum_{k=1}^N X_k}(u) = \prod_{k=1}^{N}\mathcal{C}_{X_k}(u).$$

- ***Uniqueness**: The characteristic function $\mathcal{C}_X(u)$ is uniquely related to its corresponding distribution $\Phi_X(x)$ and vice versa. (See Feller* [85] *for justification and more information on characteristic and other generating functions, as well as the inverse Fourier transform that is beyond the simple complex variables that are assumed here.)*

Examples B.50. ***Characteristic Functions for Common Distributions.***

- ***Normal distribution:***

$$\mathcal{C}_n(u;\mu,\sigma^2) = \int_{-\infty}^{\infty} e^{iux}\phi_n(x;\mu,\sigma^2)dx = e^{-0.5\sigma^2u^2+i\mu u}.$$

- ***Exponential distribution ($\mu > 0$):***

$$\mathcal{C}_e(u;\mu) = \int_0^{\infty} e^{iux}\phi_e(x;\mu)dx = \frac{1}{1-i\mu u} = \frac{1+i\mu u}{1+\mu^2u^2}.$$

- ***Uniform distribution ($a < b$):***

$$\mathcal{C}_u(u;a,b) = \frac{1}{b-a}\int_a^b e^{iux}dx = \frac{e^{iub}-e^{iua}}{i(b-a)u}.$$

- ***Double exponential (Laplace) distribution ($\mu > 0$):***

$$\mathcal{C}_{de}(u;a,\mu) = \frac{1}{2\mu}\int_0^{\infty} e^{iux}e^{-|x-a|/\mu}dx = \frac{e^{iau}}{1+\mu^2u^2}.$$

- ***Poisson distribution ($\Lambda > 0$, $x_k = k$):***

$$\mathcal{C}_p(u;\Lambda) = \sum_{k=0}^{\infty} e^{iuk}p_k(\Lambda) = \sum_{k=0}^{\infty} e^{iuk}e^{-\Lambda}\frac{\Lambda^k}{k!} = e^{-\Lambda}\sum_{k=0}^{\infty}\frac{\left(e^{iu}\Lambda\right)^k}{k!} = e^{\Lambda(e^{iu}-1)}.$$

Characteristic functions are also used to define Lévy processes, which are basically a generalization of jump-diffusion processes to include processes with infinite jump-rates. Thus, characteristic functions are essential for including such singular behavior. For references on Lévy processes see the cited sources on Lévy processes or jump-diffusion references that emphasize Lévy processes [12, 60, 223].

Another application is to financial option pricing for jump-diffusions with stochastic volatility (i.e., stochastic variance) where the characteristic function formulation and its inverse Fourier transform offer certain advantages for computation (see Carr et al. [47] or Yan and Hanson [289]).

B.6 Sample Mean and Variance: Sums of Independent, Identically Distributed (IID) Random Variables

Just as there is no such thing as a truly random variable in practice, although the theory of random variables is very useful, there is no such thing as a continuously sampled random variable in practice. Typically, we sample discretely from a theoretical continuous distribution and assume that the samples are independently sampled.

Definition B.51. ***IID Random Variables.***
A set of n random variables $\{X_k | k = 1 : n\}$ *is* ***independent and identically distributed (IID)*** *if the* X_k *have the same distribution, i.e.,*

$$\Phi_{X_k}(x) = \Phi_{X_j}(x),$$

for all $k, j = 1 : n$, *and* X_k *is independent of* X_j *when* $k \neq j$, *i.e., the joint distribution is*

$$\Phi_{X_k, X_j}(x_k, x_j) = \Phi_{X_k}(x_k) \cdot \Phi_{X_j}(x_j).$$

Definition B.52. ***Sample Mean and Variance.***
Let $\{X_k | k = 1 : n\}$ *be a sample of n random variables. Then the* ***sample mean*** *is defined as*

$$m_n = \frac{1}{n} \sum_{k=1}^{n} X_k, \tag{B.107}$$

and the ***sample variance*** *or population variance is*

$$s_n^2 = \frac{1}{n} \sum_{k=1}^{n} (X_k - m_n)^2, \tag{B.108}$$

but the ***unbiased estimate*** *of the* ***sample variance*** *is*

$$\widehat{s}_n^2 = \frac{1}{n-1} \sum_{k=1}^{n} (X_k - m_n)^2. \tag{B.109}$$

An estimate $\widehat{Y}$ *of a quantity y is called an* ***unbiased estimate*** *if*

$$\mathrm{E}\left[\widehat{Y}\right] = y.$$

Theorem B.53. ***IID Sample Mean and Variance.***
Let $\{X_k | k = 1 : n\}$ *be a set of IID random variables such that* $\mathrm{E}[X_k] = \mu$ *and* $\mathrm{Var}[X_k] = \sigma^2$ *for all k. Then*

$$\mathrm{E}[m_n] = \mu, \tag{B.110}$$

$$\mathrm{E}[s_n^2] = \frac{n-1}{n} \sigma^2, \tag{B.111}$$

$$\mathrm{E}[\widehat{s}_n^2] = \sigma^2, \tag{B.112}$$

$$\mathrm{Var}[m_n] = \frac{1}{n} \sigma^2. \tag{B.113}$$

Remarks B.54.

- *These sample moments and more are left as Exercises* 13, 15, *and* 14. *The first is trivial, but the other two rely heavily on the independence property, so it is very helpful to collect all terms as deviations from the mean forms like* $(X_k - \mu)$. *Also, split up multiple sums into a single sum for equal indices (say,* $j = k$*) and the product of an outer sum by an inner sum when the inner index is not equal to the outer index (say,* $j \neq k$*). Note that for large n, the difference between the regular and unbiased estimates of the variance will be small.*

- *Since m_n is a sum of random variables, then its distribution will be a nested convolution of the common distribution of the X_k variates. Convolutions are defined earlier in* (B.96) *of Section* B.4.
- *Later, the relevant limit theorems will be discussed. The law of large numbers* (B.114) *says that the sample mean will approach the distribution mean and the central limit theorem, Theorem* B.57*, discussed later, says that the sample distribution will approach the normal limiting distribution for large sample sizes.*
- *For properties of powers of partial sums of zero-mean IID random variables see Lemma* 5.15 *on p.* 147.

B.7 Law of Large Numbers

When applying probability to real applications, the user may need to compare the statistical properties of the practical sample with the ideal concepts of probability theory. For instance, when comparing the sample mean to an ideal distribution mean, some justification comes partly from the law of large numbers; a weak and a strong form are given here suitable for this appendix of preliminaries (see also Feller [84] or Karlin and Taylor [162]).

B.7.1 Weak Law of Large Numbers (WLLN)

Theorem B.55. ***Law of Large Numbers, Weak Form.***
Let $\{X_1, X_2, \ldots X_i, \ldots\}$ be a sequence of independent identically distributed random variables (i.e., IID random variables or mutually independent random variables with common distribution $\Phi(x)$) with common mean $\mu = \mathrm{E}[X_i]$ for all i. Let $S_n = \sum_{i=1}^{n} X_i$ be a sequence of partial sums such that S_n is the sum of n of these sample measurements, so that the sample mean is $m_n = S_n/n$. Then for every $\epsilon > 0$,

$$\mathrm{Prob}[|m_n - \mu| > \epsilon] \longrightarrow 0 \text{ as } n \to +\infty. \tag{B.114}$$

Thus, if the sample size is large enough, the sample mean will approximate the distribution mean.

B.7.2 Strong Law of Large Numbers (SLLN)

Theorem B.56. ***Law of Large Numbers, Strong Form.***
Let $\{X_1, X_2, \ldots X_i, \ldots\}$ be a sequence of independent identically distributed random variables (i.e., IID random variables or mutually independent random variables with common distribution $\Phi(x)$) with common mean $\mu = \mathrm{E}[X_i]$ for all i. Let $S_n = \sum_{i=1}^{n} X_i$ be a sequence of partial sums such that S_n is the sum of n of these sample measurements, so that the sample mean is $m_n = S_n/n$. Then

$$\begin{gathered}\mathrm{Prob}[\lim_{n\to\infty} m_n = \mu] = 1, \\ \text{i.e., } m_n \to \mu \text{ with probability one as } n \to +\infty.\end{gathered} \tag{B.115}$$

B.8 Central Limit Theorem

The central limit theorem is much more powerful than the law of large numbers. Again, a simple form is given for IID random variables [84].

Theorem B.57. ***Central Limit Theorem.***
Let $\{X_1, X_2, \ldots, X_i, \ldots\}$ *be a sequence of independent identically distributed random variables (i.e., IID random variables or mutually independent random variables with common distribution* $\Phi(x)$*) with common mean* $\mu = \mathrm{E}[X_i]$ *and variance* $\sigma^2 = \mathrm{Var}[X_i]$ *for all* i*. Let* $S_n = \sum_{i=1}^{n} X_i$ *be the sum of* n *of these sample measurements, so that the sample mean is* $m_n = S_n/n$*. Then for every fixed* ξ*,*

$$\mathrm{Prob}\left[\frac{m_n - \mu}{\sigma/\sqrt{n}} \leq \xi\right] \longrightarrow \Phi_n(\xi; 0, 1), \tag{B.116}$$

as $n \to +\infty$*, where* $\Phi_n(\xi; 0, 1)$ *is the standard normal distribution defined in* (B.1.4)*, when* $\mu = 0$ *and* $\sigma^2 = 1$*.*

Thus, if the sample size is large enough, the deviation of the sample mean from the distribution mean, scaled by $\sigma/\sqrt{n}$, will be asymptotically normally distributed with mean 0 and variance 1.

For stronger versions of the central limit theorem see the many probability references listed at the end of this appendix.

B.9 Matrix Algebra and Analysis

Many important distributions, stochastic processes, and control problems are multivariate, rather than scalar. Here matrix algebra and matrix analysis are summarized. Many of the given properties can be computed symbolically using Maple and Mathematica or numerically using MATLAB.

- **Vector notation:** $\mathbf{x} = [x_i]_{n\times 1}$, in boldface, denotes an n-vector, where the number x_i is the ith component. Let $\mathbf{y} = [y_i]_{n\times 1}$ be another n-vector. In this book vectors are column vectors, unless transposed. Numbers are also called scalars here.

- **Matrix or array notation:** $A = [a_{i,j}]_{n\times n}$ denotes an $n \times n$ square matrix (literally a table) or array, where the number $a_{i,j}$ is an element of the ith row and jth column. Sometimes we say that A is an order n matrix. Nonsquare matrices would be $Q = [q_{i,j}]_{m\times n}$ or $R = [r_{i,j}]_{n\times p}$. Matrix elements may also be functions.

- **Matrix equality:** $B = A$ means that all matrix elements are equal, $b_{i,j} = a_{i,j}$ for $i = 1 : n$ and $j = 1 : n$. The negation of the equality requires only one pair of unequal elements, $b_{k,\ell} \neq a_{k,\ell}$ for some (k, ℓ).

- **Matrix identity:**

$$I_n \equiv [\delta_{i,j}]_{n\times n}, \tag{B.117}$$

where $\delta_{i,j}$ is the Kronecker defined in (B.54) and has the sum property that $\sum_{j=1}^{n} a_j \delta_{i,j} = a_i$ provided i is in the range of j, $j = 1:n$.

- **Matrix transpose:**

$$Q^\top = [q_{j,i}]_{n\times m}, \tag{B.118}$$

i.e., transposing a real matrix is switching rows and columns. If there are complex elements, then the **Hermitian transpose** is used, $Q^H = [q^*_{j,i}]_{n\times m}$, where if $z = x + \widehat{i}\, y$ is a complex number, then the complex conjugate is $z^* = x - \widehat{i}\, y$ and $\widehat{i} = \sqrt{-1}$ is the imaginary unit such that $\widehat{i}^2 = -1$. Although this book is exclusively about real problems, there are important methods and even real problems that introduce complex numbers into the analysis.

- **Inner or dot or scalar product of two vectors:**

$$\mathbf{x}^\top \mathbf{y} = \mathbf{x} \bullet \mathbf{y} = \mathbf{x}^\top \mathbf{y} \equiv \sum_{i=1}^{n} x_i y_i, \tag{B.119}$$

provided $\mathbf{y}$ is also an n-vector. If there are complex vector elements or components, then the Hermitian inner product is used:

$$\mathbf{x}^H \mathbf{y} \equiv \sum_{i=1}^{n} x_i^* y_i.$$

- **Matrix trace:**

$$\text{Trace}[A] \equiv \sum_{i=1}^{n} a_{i,i}. \tag{B.120}$$

- **Matrix-vector product:**

$$Q\mathbf{x} \equiv \left[\sum_{j=1} q_{i,j} x_j \right]_{m\times 1}, \tag{B.121}$$

i.e., the ith component is $(Q\mathbf{x})_i = \sum_{j=1} q_{i,j} x_j$ (also, integer $m \geq 1$).

- **Matrix-matrix product:**

$$QR \equiv \left[\sum_{k=1} q_{i,k} r_{k,j} \right]_{m\times p}, \tag{B.122}$$

so for two matrices to be **commensurate** or consistent in multiplication the number of columns of the premultiplier Q must be the same as the number of rows of the postmultiplier R (also, integers $m \geq 1$ and $p \geq 1$).

- **Transpose of a matrix product:** $(QR)^\top = R^\top Q^\top$.
- **Matrix inverse:** For square matrices A, the inverse A^{-1} has the property

$$A^{-1}A = I_n = AA^{-1} \tag{B.123}$$

whenever A^{-1} exists and this property provides a set of algebraic equations for determining the elements of the inverse. See the MATLAB, Maple, and Mathematica packages.

- **Vector norm:**

$$||\mathbf{x}||_p \equiv \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} \tag{B.124}$$

is the pth norm with the properties that

1. $||\mathbf{x}||_p \geq 0$;
2. $||\mathbf{x}||_p = 0$ if and only if $\mathbf{x} = \mathbf{0}$;
3. $||s\mathbf{x}||_p = |s|||\mathbf{x}||_p$ if s is a scalar;
4. $||\mathbf{x} + \mathbf{y}||_p \leq ||\mathbf{x}||_p + ||\mathbf{y}||_p$, called the **triangular inequality**;
5. $||\mathbf{x}^\top \mathbf{y}||_p \leq ||\mathbf{x}||_p ||\mathbf{y}||_p$, called the **Cauchy inequality**.

Common norms are the

1. 1-norm, $||\mathbf{x}||_1 = \sum_{i=1}^{n} |x_i|$;
2. infinity-norm, $||\mathbf{x}||_\infty = \max_{i=1:n}[|x_i|]$;
3. 2-norm, $||\mathbf{x}||_2 = \sqrt{\sum_{i=1}^{n} x_i^2} = \sqrt{\mathbf{x}^\top \mathbf{x}}$ if $\mathbf{x}$ real, but $||\mathbf{x}||_2 = \sqrt{\mathbf{x}^H \mathbf{x}}$ if complex.

- **Matrix norm:** Matrix norms are defined on the more basic vector norms,

$$||A||_p \equiv \max_{||\mathbf{x}||_p \neq 0} \left[||A\mathbf{x}||_p \big/ ||\mathbf{x}||_p \right] = \max_{||\mathbf{u}||_p \neq 1} \left[||A\mathbf{u}||_p \right], \tag{B.125}$$

and they satisfy properties analogous to the vector norm properties above. Usual values are $p = 1, 2$, or ∞.

- **Matrix condition number:**

$$\text{cond}_{\,p}[A] \equiv ||A||_p ||A^{-1}||_p \tag{B.126}$$

is the pth condition number, bounded below by $\text{cond}_p[A] \geq 1$ and is scale-invariant since $\text{cond}_p[sA] = |s|\ \text{cond}_{\,p}[A]$ if s is a nonzero scalar. Implicit in the definition is that the inverse A^{-1} exists.

- **Matrix determinants:** If A is a square matrix, then the determinant $\text{Det}[A]$ has a scalar value that can be computed by recursion from smaller determinants, expanding by either a row or a column. For instance,

1. If $n = 1$, then $\mathrm{Det}[a_{1,1}] = a_{1,1}$.
2. If $n = 2$, then

$$\mathrm{Det}\begin{bmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{bmatrix} = a_{1,1}\mathrm{Det}[a_{2,2}] - a_{1,2}\mathrm{Det}[a_{2,1}].$$

3. If $n = 3$, then

$$\begin{aligned}\mathrm{Det}\begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \\ a_{3,1} & a_{3,2} & a_{3,3} \end{bmatrix} &= a_{1,1}\mathrm{Det}\begin{bmatrix} a_{2,2} & a_{2,3} \\ a_{3,2} & a_{3,3} \end{bmatrix} \\ &\quad - a_{1,2}\mathrm{Det}\begin{bmatrix} a_{2,1} & a_{2,3} \\ a_{3,1} & a_{3,3} \end{bmatrix} \\ &\quad + a_{1,3}\mathrm{Det}\begin{bmatrix} a_{2,1} & a_{2,2} \\ a_{3,1} & a_{3,2} \end{bmatrix}.\end{aligned}$$

4. And so forth.

Some useful properties are $\mathrm{Det}[A^\top] = \mathrm{Det}[A]$ since row and column expansions give the same result; the **Cauchy–Binet** formula states that

$$\mathrm{Det}[AB] = \mathrm{Det}[A]\mathrm{Det}[B] \tag{B.127}$$

provided A and B are commensurate, and $\mathrm{Det}[I_n] = 1$; a corollary is $\mathrm{Det}[A^{-1}] = 1/\mathrm{Det}[A]$ if $A^{-1}A = I_n$.

- **Systems of linear equations:**

$$A\mathbf{x} = \mathbf{b}, \tag{B.128}$$

where the coefficient matrix A and $\mathbf{b} = [b_i]_{n\times 1}$ are given and the object is to find the vector $\mathbf{x}$.

1. In theory, a unique solution exists if $\mathrm{Det}[A] \neq 0$; else if $\mathrm{Det}[A] = 0$, then A is called singular.
2. In numerical practice, a nearly singular A usually has serious problems and the condition number $\mathrm{cond}[A]$ due to its scale-invariance is a better measure of difficulties. If $\mathrm{cond}[A]$ is of moderate size (not much bigger than O(1), say), then the problem is called **well-conditioned**, but if $\mathrm{cond}[A]$ is very large, then the problem is called **ill-conditioned**. In Gaussian elimination with back substitution, row pivoting with row scaling or full pivoting can reduce the conditioning problems and produce more reliable approximate solutions. The MATLAB, Maple, and Mathematica systems provide either numerical or symbolic functions to solve $A\mathbf{x} = \mathbf{b}$.

- **Matrix eigenvalue problems:**

$$A\mathbf{x} = \lambda\mathbf{x} \tag{B.129}$$

is the eigenvalue problem statement, where the object is to find a set of characteristic values or eigenvalues λ_k and associated eigenvectors $\mathbf{x}_k$ that characterize the matrix A.

1. Since the algebraic problem $(A - \lambda_k I_n)\mathbf{x}_k = \mathbf{0}$ is equivalent to the original (B.129),
$$\text{Det}[A - \lambda I_n] = 0$$
is called the **characteristic** or **eigen equation**.
2. $(A - \lambda_k I_n)$ is an nth polynomial in λ_k,
$$P_n(\lambda) = \sum_{i=0}^{n} c_i \lambda^i,$$
where $c_0 = \text{Det}[A]$, $c_1 = -\text{Trace}[A], \ldots, c_n = (-1)^n$.
3. The characteristic equation is the condition for finding a nontrivial eigenvalue, $\mathbf{x}_k[x_{i,k}]_{n\times 1} \neq \mathbf{0}$.
4. Solving $\text{Det}[A - \lambda I_n] = 0$ yields n eigenvalues $[\lambda_i]_{n\times 1}$.
5. The eigenvectors can be found from a subset of the original problem but are not unique.
6. If $\mathbf{x}_k$ is an eigenvector, then so is $\mathbf{y} = s*\mathbf{x}$, where s is an arbitrary, nonzero scalar.
7. A unit or normalized eigenvector is of the form $||\mathbf{u}_k||_p = 1$.
8. If A is real and symmetric, then the eigenvectors are **orthogonal** if $\mathbf{x}_j^\top \mathbf{x}_k = ||\mathbf{x}_k||_2^2 \delta_{j,k}$ or **orthonormal** if $||\mathbf{x}_k||_2 = 1$ in addition.
9. If A is not real and nonsymmetric, then the left or adjoint eigen problem
$$\mathbf{y}_j^H A = \mu_j^* \mathbf{y}_j^H \quad \text{or} \quad A^H \mathbf{y}_j = \mu_j \mathbf{y}_j$$
would be needed for orthogonality conditions since $0 = (\lambda_k - \mu_j^*)\mathbf{y}_j^H \mathbf{x}_k$, so if $\mu_j^* \neq \lambda_k$, then $\mathbf{y}_j^H \mathbf{x}_k = 0$.

- **Gradient of a scalar valued function of a vector-argument:**

$$\nabla_{\mathbf{x}}[F](\mathbf{x}) = \frac{\partial F}{\partial \mathbf{x}}(\mathbf{x}) = F_{\mathbf{x}}(\mathbf{x}) \equiv \left[\frac{\partial F}{\partial x_i}(\mathbf{x})\right]_{n\times 1}, \tag{B.130}$$

so the gradient is a column vector with the same shape as $\mathbf{x}$ here. In some texts [44], the gradient may be a row vector, so matrix-vector products will be different there.

- **Gradient of a matrix-vector product transpose:**

$$\begin{aligned}\nabla_{\mathbf{x}}\left[(A\mathbf{x})^\top\right] &= \left[\frac{\partial}{\partial x_i}\sum_{k=1}^{n} a_{j,k}x_k\right]_{n\times n} = \left[\sum_{k=1}^{n} a_{j,k}\delta_{i,k}\right]_{n\times n} \\ &= \left[a_{j,i}\right]_{n\times n} = A^\top,\end{aligned} \tag{B.131}$$

so the gradient just peels off the premultiplied $\mathbf{x}^\top$ since $(A\mathbf{x})^\top = \mathbf{x}^\top A^\top$ (i.e., the **gradient peel theorem**).

- **Quadratic forms:**

$$\mathcal{Q} = \mathbf{x}^\top A\mathbf{x} = \sum_{i=1}^{n}\sum_{j=1}^{n} x_i a_{i,j} x_j, \tag{B.132}$$

which is a scalar, and since $\mathcal{Q}$ is a scalar and the transpose has no effect on scalars, then

$$\mathcal{Q} = \mathcal{Q}^\top = \mathbf{x}^\top A^\top \mathbf{x} = \frac{1}{2}\left(\mathcal{Q} + \mathcal{Q}^\top\right) = \mathbf{x}^\top A^S \mathbf{x}, \tag{B.133}$$

where $A^S \equiv \frac{1}{2}(A + A^\top)$ is the symmetric part of A. Thus, for quadratic forms, the user might as well assume A to be symmetric or that $A^\top = A$.

- **Positive definite matrices:** The matrix A is **positive definite** if for every nonzero vector $\mathbf{x}$ ($\mathbf{x} \neq \mathbf{0}$) the quadratic form

$$\mathbf{x}^\top A\mathbf{x} > 0, \tag{B.134}$$

sometimes abbreviated as $A > 0$. Similarly, A is **positive semidefinite** if for all $\mathbf{x} \neq \mathbf{0}$,

$$\mathbf{x}^\top A\mathbf{x} \geq 0, \tag{B.135}$$

or if so, then we say $A \geq 0$. Further, A is positive definite if and only if all its eigen-values are positive [68], so then A is invertible, i.e., A^{-1} exists.

- **Gradient of a quadratic form:**

$$\nabla_{\mathbf{x}}\left[\mathbf{x}^\top A\mathbf{x}\right] = 2A\mathbf{x}, \tag{B.136}$$

assuming A is symmetric, by two applications of the peel theorem, one on the left and another on the right by transposing first.

- **Hessian matrix of a scalar valued function:**

$$\nabla_{\mathbf{x}}\left[\nabla_{\mathbf{x}}^\top [F]\right](\mathbf{x}) = \left[\frac{\partial^2 F}{\partial x_i \partial x_j}(\mathbf{x})\right]_{n\times n}, \tag{B.137}$$

so the matrix of second derivatives is a square $n \times n$ matrix.

- **Hessian matrix of a quadratic form:**

$$\nabla_{\mathbf{x}}\left[\nabla_{\mathbf{x}}^\top [\mathbf{x}^\top A\mathbf{x}]\right] = \nabla_{\mathbf{x}}\left[2(A\mathbf{x})^\top\right] = 2\nabla_{\mathbf{x}}\left[\mathbf{x}^\top A\right] = 2A \tag{B.138}$$

by the peel theorem, assuming that A is symmetric.

B.10 Some Multivariate Distributions

The probability distributions, such as normal, exponential, and Poisson, previously considered have been functions of a single real sample variable representing a single random variate. However, some applications require multidimensional distributions representing jointly distributed multivariate random variables. The continuous multivariate normal (multinormal) distribution and the discrete multinomial distribution will serve as examples.

B.10.1 Multivariate Normal Distribution

Definition B.58. *The **multivariate normal distribution** for the real m-dimensional vector random variate $\mathbf{X} = [X_i]_{m\times 1} \in \mathbb{R}^m$ is defined by the density in matrix-vector notation as*

$$\phi_n(\mathbf{x};\, \boldsymbol{\mu},\, \boldsymbol{\Sigma}) \equiv \frac{1}{(2\pi)^{m/2}\sqrt{\mathrm{Det}[\boldsymbol{\Sigma}]}} \exp\left(-0.5(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right), \tag{B.139}$$

*where $\boldsymbol{\mu} = [\mu_i]_{m\times 1} = \mathrm{E}[\mathbf{X}]$ is the **vector mean**,*

$$\boldsymbol{\Sigma} = [\sigma_{i,j}]_{m\times m} = \mathrm{E}\left[\left[(X_i-\mu_i)(X_j-\mu_j)\right]_{m\times m}\right]$$

*is the positive definite **variance-covariance matrix**, i.e., $\sigma_{i,i} \equiv \sigma_i^2 = \mathrm{Var}[X_i]$ for $i = 1:m$, while $\sigma_{i,j} \equiv \mathrm{Cov}[X_i, X_j]$ if $j \neq i$ for $i, j = 1:m$, and $\mathrm{Det}[\boldsymbol{\Sigma}]$ is the determinant of $\boldsymbol{\Sigma}$. The **correlation coefficient** is the normalized covariance,*

$$\rho_{i,j} \equiv \frac{\mathrm{Cov}[X_i, X_j]}{\sqrt{\mathrm{Var}[X_i]\mathrm{Var}[X_j]}} = \frac{\sigma_{i,j}}{\sigma_i \sigma_j}, \tag{B.140}$$

provided $\sigma_i,\ \sigma_j \neq 0$, and $i,\ j \neq 0$.

Total probability is conserved since

$$\int_{\mathbb{R}^m} \phi_n(\mathbf{x};\, \boldsymbol{\mu},\, \boldsymbol{\Sigma}) d\mathbf{x} = 1.$$

Theorem B.59. ***Correlation coefficient bounds.***
Let X_1 and X_2 be two random variables. Then

$$|\rho(X_1, X_2)| \leq 1, \tag{B.141}$$

provided $\sigma_1 > 0$ and $\sigma_2 > 0$, but if $\rho(X_1, X_2) = \pm 1$, then

$$X_2/\sigma_2 = \pm X_1/\sigma_1 + C \tag{B.142}$$

for some constant C.

Proof. The proof is modeled after Feller's proof [84, p. 236]. Let $\rho = \rho(X_1, X_2)$, and using (B.74)

$$\begin{aligned}\mathrm{Var}[X_1/\sigma_1 \pm X_2/\sigma_2] &= \mathrm{Var}[X_1/\sigma_1] \pm 2\mathrm{Cov}[X_1/\sigma_1, X_2/\sigma_2] + \mathrm{Var}[X_2/\sigma_2] \\ &= 2(1 \pm \rho) \geq 0,\end{aligned}$$

since $\mathrm{Var}[X] \geq 0$, so $|\rho| \leq 1$.

If $\rho = 1$, then let $\pm 1 = -1$ and thus $X_1/\sigma_1 - X_2/\sigma_2 = C_1$, where C_1 is a constant, but if $\rho = -1$, then let $\pm 1 = +1$ and thus $X_1/\sigma_1 + X_2/\sigma_2 = C_2$, where C_2 is a constant. Combining these two cases leads to the form (B.142). □

Example B.60. *The **bivariate normal distribution**, i.e., the two-dimensional case, needs several conditions to keep the density well-defined: $\sigma_i > 0$ for $i = 1 : 2$, $\sigma_{1,2} = \rho\sigma_1\sigma_2$ where $\rho = \rho_{1,2}$ is the correlation coefficient between states 1 and 2 such that $-1 < \rho < +1$. Thus,*

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}, \tag{B.143a}$$

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{1-\rho^2} \begin{bmatrix} 1/\sigma_1^2 & -\rho/(\sigma_1\sigma_2) \\ -\rho/(\sigma_1\sigma_2) & 1/\sigma_2^2 \end{bmatrix}. \tag{B.143b}$$

The $\boldsymbol{\Sigma}^{-1}$ follows upon calculating the two-dimensional inverse of $\boldsymbol{\Sigma}$, while substituting for $\boldsymbol{\Sigma}^{-1}$ and $\mathrm{Det}[\boldsymbol{\Sigma}] = (1-\rho^2)\sigma_1^2\sigma_2^2$ yields the more explicit density form:

$$\phi_n\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}; \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left(-\frac{0.5}{1-\rho^2}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]\right). \tag{B.144}$$

Some of the first few moments are tabulated (results from the Maple symbolic computation system) in Table B.1*:*

Table B.1. *Some expected moments of bivariate normal distribution.*

Some Binormal Expectations
$\mathrm{E}[1] = 1$
$\mathrm{E}[x_i] = \mu_i,\ i = 1:2$
$\mathrm{Var}[x_i] = \sigma_i^2,\ i = 1:2$
$\mathrm{Cov}[x_1, x_2] = \rho\sigma_1\sigma_2$
$\mathrm{E}[(x_i - \mu_i)^3] = 0,\ i = 1:2$
$\mathrm{E}[(x_i - \mu_i)^4] = 3\sigma_i^4,\ i = 1:2$
$\mathrm{E}[(x_1 - \mu_1)^2(x_2 - \mu_2)^2] = (1 + 2\rho^2)\sigma_1^2\sigma_2^2$

Remark B.61. *The bivariate normal density becomes singular when $\sigma_1 \to 0^+$ or $\sigma_2 \to 0^+$ or $\rho^2 \to 1^-$ and the density becomes degenerate. If $\rho > 0$, then X_1 and X_2 are **positively correlated**, while if $\rho < 0$, then X_1 and X_2 are **negatively correlated**.*

B.10.2 Multinomial Distribution

The multinomial distribution may be useful for studying discrete collections of samples from continuous distributions, such as the bin frequencies of histograms and many other applications [84, 129].

Definition B.62. *Using m bins where π_k ($0 < \pi_k < 1$) is the theoretical probability associated with the kth bin as well as a parameter of the distribution for $k = 1 : m$ bins such that*

$$\sum_{k=1}^{m} \pi_k = 1 \tag{B.145}$$

and f_k is the observed frequency (integer outcome count, $f_k \geq 0$) for the kth bin for a sample of N observations such that

$$\sum_{k=1}^{m} f_k = N, \tag{B.146}$$

*the **multinomial distribution** is given by the joint probability mass function*

$$p(\mathbf{f}; \boldsymbol{\pi}) = \text{Prob}\left[\mathbf{F} = \mathbf{f} \,\middle|\, \mathbf{1}^T\boldsymbol{\pi} = 1, \mathbf{1}^T\mathbf{f} = N\right] = N! \prod_{k=1}^{m} \frac{\pi_k^{f_k}}{f_k!}, \tag{B.147}$$

where $\mathbf{f} = [f_i]_{m\times 1}$ is the frequency value vector, $\mathbf{F} = [F_i]_{m\times 1}$ is the random frequency vector, and $\mathbf{1} = [1]_{m\times 1}$ is the ones or summing vector.

Example B.63. *When $m = 2$, the multinomial distribution is called the **binomial distribution** and has probability function*

$$p(f_1, f_2; \pi_1, \pi_2) = \frac{N!\pi_1^{f_1}\pi_2^{f_2}}{f_1!f_2!} = \begin{pmatrix} N \\ f_1 \end{pmatrix} \pi_1^{f_1}(1-\pi_1)^{N-f_1}, \tag{B.148}$$

*where the **binomial coefficient***

$$\begin{pmatrix} n \\ k \end{pmatrix} \equiv \frac{n!}{k!(n-k)!} \tag{B.149}$$

*with the constraints $f_2 = N - f_1$ and $\pi_2 = 1 - \pi_1$ used on the far right-hand side. The binomial distribution is applicable to trials with just two outcomes, called **Bernoulli trials** (Feller [84]). Often these two outcomes or bins are identified as either a **success**, with probability π_1, or **failure**, for example, with probability $\pi_2 = 1 - \pi_1$. Feller [84] calls the binomial distribution, the normal distribution, and the Poisson distribution the three principal distributions throughout probability theory.*

*The **binomial theorem** gives the **binomial expansion***

$$(\pi_1 + \pi_2)^N = \sum_{f_1=0}^{N} \begin{pmatrix} N \\ f_1 \end{pmatrix} \pi_1^{f_1}\pi_2^{N-f_1}, \tag{B.150}$$

but the coefficients are precisely the binomial probability functions

$$(\pi_1 + \pi_2)^N = \sum_{f_1=0}^{N} p(f_1, N - f_1; \pi_1, \pi_2), \tag{B.151}$$

which is why the distribution in (B.148) *is called the Binomial distribution for binomial frequencies* f_1 *for* $f_1 = 0 : N$ *(Feller* [84]*).*

Consequently, the binomial expectation for some function g *is given by*

$$\mathrm{E}[g(F_1)] = \sum_{f_1=0}^{N} g(f_1)\, p(f_1, N - f_1; \pi_1, 1 - \pi_1).$$

Using parametric differentiation of the sums, with F_k *being the* k*th random variable and* f_k *being the* k*th given conditioned variable, it can be shown that*

- $\mathrm{E}[1] = 1$ *when* $g(f_k) = 1$ *(actually* (B.150) *or* (B.151) *with* $\pi_2 = \pi_1$*),*
- $\mathrm{E}[F_k] = N\pi_k$ *when* $g(f_k) = f_k$*,*
- $\mathrm{Var}[F_k] = N\pi_k(1 - \pi_k)$ *when* $g(f_k) = (f_k - N\pi_k)^2$*,*
- $\mathrm{Cov}[F_1, F_2] = -N\pi_1\pi_2 = -N\pi_k(1 - \pi_k) = -\mathrm{Var}[F_1]$ *when* $g(f_1) = (f_1 - N\pi_1)((N - f_1) - N(1 - \pi_1)) = -N(f_1 - N\pi_1)^2$.

As an illustration of an application of parametric differentiation to sum a finite number of terms, consider the first moment:

$$\begin{aligned}
\mathrm{E}[F_1] &= \sum_{f_1=0}^{N} f_1 \left(\begin{array}{c} N \\ f_1 \end{array} \right) \pi_1^{f_1} (1 - \pi_1)^N - f_1 \\
&= \pi_1 \frac{d}{d\pi_1} \left[\sum_{f_1=0}^{N} \left(\begin{array}{c} N \\ f_1 \end{array} \right) \pi_1^{f_1} (\pi_2)^N - f_1 \right] \Bigg|_{\pi_2 = 1 - \pi_1} \\
&= \pi_1 \frac{d}{d\pi_1} \left[(\pi_1 + \pi_2)^N \right] \big|_{\pi_2 = 1 - \pi_1} = \pi_1 N \left[(\pi_1 + \pi_2)^{N-1} \right] \big|_{\pi_2 = 1 - \pi_1} = N\pi_1.
\end{aligned}$$

Similarly, forms with powers of $\{\pi_1, d/d\pi_1\}$ *can be used for higher moments.*

Figure B.6 *illustrates the binomial distributions as a function of the binomial frequency* f_1 *when the total count is* $N = 10$ *for three values of the binomial probability parameter,* $\pi_1 = 0.25, 0.5,$ *and* 0.75*. See Online Appendix* C*, Section* C.6*, for the MATLAB figure code. These binomial distributions roughly resemble a discretized version of the normal distribution, except that they are skewed for* $\pi_1 = 0.25$ *and* 0.75 *while the distribution for* $\pi_1 = 0.50$ *is symmetric. Feller* [84] *states that when* $N\pi_1(1 - \pi_1)$ *is large, the binomial distribution can be approximated by the normal distribution with mean* $N\pi_1$ *and variance* $N\pi_1(1 - \pi_1)$*, but when* N *is large and* π_1 *is the same order as* $1/N$*, then the binomial distribution can be approximated by the Poisson distribution with* $\Lambda = N\pi_1$ *order one. Since the Poisson can also be approximated by the normal approximation, there is some overlap of the two approximations, but only the Poisson approximation is good when* $\Lambda = N\pi_1$ *is small.*

The multinomial distribution has the same basic moments as the binomial, but the constraints on the π_k and f_k also constrain the expectation summations. The multinomial

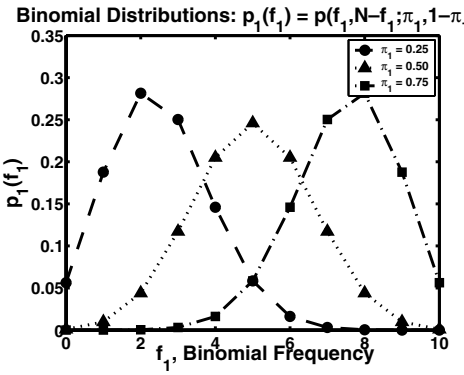


Figure B.6. *Binomial distributions with respect to the binomial frequency* f_1 *with* $N = 10$ *for values of the probability parameter,* $\pi_1 = 0.25$*,* 0.5*, and* 0.75*. These represent discrete distributions, but discrete values are connected by dashed, dotted, and dash-dotted lines only to help visualize the distribution form for each parameter value.*

distribution in (B.147) is in fact the terms in the multinomial expansion theorem,

$$\left(\sum_{k=1}^{m} \pi_k\right)^N = N! \prod_{i=1}^{m-1} \left[\sum_{f_i=0}^{(N-\mathcal{F}_{i-1})} \frac{\pi_i^{f_i}}{f_i!} \right] \frac{\pi_m^{N-\mathcal{F}_{m-1}}}{(N-\mathcal{F}_{m-1})!},$$
$$= \prod_{i=1}^{m-1} \left[\sum_{f_i=0}^{(N-\mathcal{F}_{i-1})} \right] p(\mathbf{f};\boldsymbol{\pi}) \Bigg|_{f_m=(N-\mathcal{F}_{m-1})}, \tag{B.152}$$

which can be obtained from $(m-1)$ successive applications of the binomial expansion. It can be shown by induction upon replacing π_m by $(\pi_m + \pi_{m+1})$ in the induction hypothesis above and using an additional application of the binomial expansion with the power $(N - \mathcal{F}_{m-1})$. Here, $\mathcal{F}_k \equiv \sum_{j=1}^{k} f_j$ is the partial sum of the first k frequencies, such that $\mathcal{F}_0 \equiv 0$. For application to the multinomial distribution, the constraints lead to the elimination formula $f_m = N - \mathcal{F}_{m-1}$ for the mth terms, so that the final fraction in (B.152) depends on the first $m-1$ sample frequencies f_k. In the case of the multinomial distribution, also the mth theoretical probability $\pi_m = 1 - \sum_{j=1}^{m-1} \pi_j$ can be eliminated by conservation of probability.

B.11 Basic Asymptotic Notation and Results

Definitions and Results B.64. *For purposes of a refined study of limits and asymptotic behaviors found in many stochastic problems, basic* ***asymptotic concepts*** *can be defined as follows.*

- ***Equals big Oh*** *or* ***is the order of*** *symbol is such that* $\boldsymbol{f(x) = O(g(x))}$ *as* $x \to x_0$ *if* $f(x)/g(x)$ *is bounded as* $x \to x_0$ *provided* $g(x) \neq 0$ *in a deleted neighborhood of* $x = x_0$*.*

For example, $8\sin(\epsilon/7) = \mathrm{O}(\epsilon)$ *as* $\epsilon \to 0$ *or* $(2N^2 + 3N + 100)/(3N + 5) = \mathrm{O}(N)$ *as* $N \to \infty$ *or* $\exp(-0.5\Delta t) = 1 - 0.5\Delta t + \mathrm{O}((\Delta t)^2)$ *as* $\Delta t \to 0$. *Also,* $\mathrm{O}(100\Delta t) = \mathrm{O}(\Delta t)$ *as* $\Delta t \to 0$, *since constants need not be considered. As alternate notation,* $\mathrm{O}((\Delta t)^2) = \mathrm{O}^2(\Delta t)$ *as* $\Delta t \to 0$.

- ***Equals little oh*** *or* ***is smaller order than*** *is such that* $\boldsymbol{f(x) = \mathrm{o}(g(x))}$ *as* $x \to x_0$ *if* $f(x)/g(x) \to 0$ *as* $x \to x_0$ *provided* $g(x) \neq 0$ *in a deleted neighborhood of* $x = x_0$. *Also the notation* $f(x) \ll g(x)$ *is equivalent to* $f(x) = \mathrm{o}(g(x))$.

  For example, $\exp(-0.5\Delta t) = 1 - 0.5\Delta t + \mathrm{o}(\Delta t)$ *as* $\Delta t \to 0$ *or* $\int_t^{t+\Delta t} f(\tau)d\tau = f(t)\Delta t + \mathrm{o}(\Delta t)$ *as* $\Delta t \to 0$ *provided* $f(t)$ *is continuous. Note* $\mathrm{O}(\Delta t) + \mathrm{o}(\Delta t) = \mathrm{O}(\Delta t)$ *as* $\Delta t \to 0$.

- ***Equals*** **ord** *or* ***is the same order as*** *is such that* $\boldsymbol{f(x) = \mathrm{ord}(g(x))}$ *as* $x \to x_0$ *if* $f(x) = \mathrm{O}(g(x))$ *but that* $f(x) \neq \mathrm{o}(g(x))$. *The relation* $f(x) \leq \mathrm{ord}(g(x))$ *is equivalent to* $f(x) = \mathrm{O}(g(x))$ *and* $f(x) < \mathrm{ord}(g(x))$ *is equivalent to* $f(x) = \mathrm{o}(g(x))$.

  For example, $(\Delta t)^2 < \mathrm{ord}(\Delta t)$ *as* $\Delta t \to 0$ *but* $\Delta t > \mathrm{ord}((\Delta t)^2)$ *as* $\Delta t \to 0$.

- *The* ***symbol*** $\boldsymbol{\sim}$ *or* ***is asymptotic to*** *is such that* $\boldsymbol{f(x) \sim g(x)}$ *as* $x \to x_0$ *if* $f(x)/g(x) \to 1$ *as* $x \to x_0$ *provided* $g(x) \neq 0$ *in a deleted neighborhood of* $x = x_0$.

  For example, $(1 - \exp(-0.425\Delta t))/\Delta t \sim 0.425$ *as* $\Delta t \to 0$.

  Remark B.65. *The symbol* $\sim$ *is commutative since if* $f(\epsilon) \sim g(\epsilon)$, *then* $g(\epsilon) \sim f(\epsilon)$ *as* $\epsilon \to 0$ *provided both* $f(\epsilon)$ *and* $g(\epsilon)$ *are not equal to zero in a neighborhood of* $\epsilon = 0$. *Also, one should* ***never*** *say that* $f(\epsilon) \sim 0$ *(bad asymptotics and mathematics) since according to our definition that would be dividing by zero.*

- *A sequence* $\{\phi_n(x)\}$ *for* $n = 0 : \infty$ *is an* ***asymptotic sequence*** *if* $\phi_{n+1}(x) < \mathrm{ord}(\phi_n(x))$ *as* $x \to x_0$.

  For example, $\phi_n(x) = (x - x_0)^n$ *as* $x \to x_0$ *or* $\phi_n(\Delta t) = (\Delta t)^{n/2}$ *as* $\Delta t \to 0^+$ *for* $n = 0 : \infty$.

- *An expansion* $\sum_{n=0}^{\infty} a_n\phi_n(x)$, *where* a_n *are coefficients constant in* x *and* $\phi_n(x)$ *are elements of an asymptotic sequence, is an* ***asymptotic expansion*** *which is asymptotic to a function* $f(x)$ *if*

$$f(x) - \sum_{n=0}^{N} a_n\phi_n(x) < \mathrm{ord}(\phi_N(x))$$

  as $x \to x_0$ *for all* N, *and if so, then*

$$f(x) \sim \sum_{n=0}^{\infty} a_n\phi_n(x)$$

as $x \to x_0$. As a corollary, the inductive algorithm for the coefficients follows starting with $a_0 = \lim_{x\to x_0} f(x)/\phi_0(x)$ and

$$a_{N+1} = \lim_{x\to x_0} \frac{f(x) - \sum_{n=0}^{N} a_n \phi_n(x)}{\phi_{N+1}(x)}$$

for $N = 0 : +\infty$, assuming that all limits exist.

For example, most convergent Taylor series, when considered under limiting conditions, are asymptotic expansions, or asymptotic power series in particular,

$$f(x) \sim \sum_{n=0}^{\infty} f^{(n)}(x_0)(x - x_0)^n/n!$$

as $x \to x_0$, but some asymptotic expansions can be divergent and still be useful if a finite number of terms are used, such as the expansion of the famous Stieltjes integral divergent asymptotic expansion example [28]

$$\int_0^{\infty} \frac{e^{-t}dt}{(1 + xt)} \sim \sum_{n=0}^{\infty} (-1)^n n! x^n$$

as $x \to 0$, which clearly diverges. For asymptotic applications, we are usually interested in only a few terms, whether the expansion is convergent or divergent, so the first few terms of a divergent expansion can be useful. Limits play a different role in asymptotic expansions than they do for Taylor series, in that limits of the independent variable (here, x) are used in asymptotics, while limits of the index (here, n) are used to test the convergence or divergence of Taylor series for a fixed value of the independent variable.

- *For integrals dominated by an exponential whose exponent, say, $\phi(x)/\epsilon$, has a maximum at x^* within the interior of the range of integration (a, b) such that $\phi'(x^*) = 0$ and $\phi''(x^*) < 0$, i.e., $\phi(x) \sim \phi(x^*) + 0.5\phi''(x^*)(x - x^*)^2$, while $f(x) \sim f(x^*)$ is continuous and subdominant, as $x \to x^*$ and $0 < \epsilon << 1$, **Laplace's method** for **asymptotic evaluation of integrals** [28] leads to the asymptotic approximation*

$$\int_a^b e^{\phi(x)/\epsilon} f(x)dx \sim \sqrt{\frac{2\pi\epsilon}{-\phi''(x^*)}} e^{\phi(x^*)/\epsilon} f(x^*) \tag{B.153}$$

as $\epsilon \to 0^+$. If $x^* = a$ or $x^* = b$, i.e., an end point maximum, then the integral is asymptotic to one half the above approximation.

*For example, the general **factorial function** or **gamma function** [2] for real x with $x > -1$,*

$$\begin{aligned} x! = \Gamma(x + 1) &= \int_0^{\infty} e^{-t} t^x dt = x^{x+1} \int_0^{\infty} e^{x(-y+\ln(y))} dy \\ &\sim \sqrt{2\pi x} e^{-x} x^x \end{aligned} \tag{B.154}$$

as $x \to \infty$, after transforming the original integral to the Laplace form using $t = xy$ with $\phi(y) = -y + \ln(y)$ and $\epsilon = 1/x$, since the fast exponentially decaying coefficient function $\exp(-t)$ *does not satisfy the subdominant requirement for Laplace's method. (Often, some transformation is necessary to fit a method.) The result is a correction to Stirling's (asymptotic) formula* $\ln(x!) \sim x \ln(x)$*, which is only the leading term of the exponent expansion of* $x!$ *as* $x \to \infty$*. Some authors refer to the leading term* (B.154) *of the full integral as Stirling's formula, e.g., Feller* [84].

Remark B.66. ***Laplace and Probability.***
Since Laplace was associated with the early foundational work in the analytical theory of probability with his treatise Théorie Analytique des Probabilités*, it is likely that Laplace's method was developed for probability integrals, in particular normal probability integrals, which were not restricted to infinite or zero limits of integration and the integrals can be found exactly.*

B.12 Generalized Functions: Combined Continuous and Discrete Processes

In stochastic problems, especially in extreme limits and distributions, representations beyond ordinary functions, such as generalized functions, are useful for the complete description of stochastic problems, such as combined continuous and discrete processes. While there are alternative abstract representations, generalized functions are very helpful in motivating stochastic models and solutions for associated stochastic problems as they are for the study of differential equations. Many generalized functions are defined only under integration but can be constructed as the limit of a sequence of ordinary functions.

Definitions B.67.

- *The* ***Heaviside step function****,* $H(x)$*, is a generalized function with the property that*

$$\int_{-\infty}^{+\infty} f(x)H(x - x_0)dx = \int_{x_0}^{+\infty} f(x)dx \tag{B.155}$$

for some integrable function $f(x)$ *on* $(-\infty, +\infty)$*.*

- ***Heaviside step function****:*
One pointwise definition of the ***Heaviside step function*** *is*

$$H(x) = \begin{Bmatrix} 0, & x < 0 \\ 1, & x \geq 0 \end{Bmatrix}, \tag{B.156}$$

which is right-continuous, but another version takes the average (a) *value at zero so that it has better numerical properties,*

$$H_a(x) = \begin{Bmatrix} 0, & x < 0 \\ 1/2, & x = 0 \\ 1, & x > 0 \end{Bmatrix}, \tag{B.157}$$

although the Heaviside function is often left undefined at $x = 0$ since a single isolated point does not contribute to an ordinary or Riemann integral. For generalized functions, the averaged one, $H_a(x)$, is better for underlying numerical approximations.

- *For intervals on the real line, the right-continuous Heaviside step function is related to the* ***indicator function*** *for some set A,*

$$\mathbf{1}_{x\in A} \equiv \begin{Bmatrix} 1, & x \in A \\ 0, & x \notin A \end{Bmatrix}, \tag{B.158}$$

so that

$$\mathbf{1}_{x\in(0,+\infty)} = H(x)$$

using the above Heaviside step function definition.

For example, the probability distribution can be written

$$\Phi_X(\xi) = \mathrm{E}_X[H(\xi - X)] = \mathrm{E}_X[\mathbf{1}_{X\in(-\infty,\xi]}], \tag{B.159}$$

provided the density is sufficiently continuous. Note that $\mathbf{1}_{(y-x)\in[a,b)} = \mathbf{1}_{x\in(y-b,y-a]}$, by definition, is a technique which becomes more useful in calculating multivariate probability distributions.

Definition B.68. ***Dirac Delta Function.***
The ***Dirac delta function****, $\delta(x)$, is a generalized function with the property that*

$$\int_{-\infty}^{+\infty} f(x)\delta(x - x_0)dx = f(x_0) \tag{B.160}$$

for any continuous function $f(x)$ defined for x on $\mathbb{R}$ and some point x_0 on $\mathbb{R}$ (see Friedman [89]*).*

Remark B.69. *The generalized function $\delta(x - x_0)$ is not a regular function and it has meaning only in the integrand of an integral. Since $\delta(x - x_0)$ picks out a single value of the function $f(x)$, it must be concentrated at a point, i.e., for any $\epsilon > 0$,*

$$\int_{x_0-\epsilon}^{x_0+\epsilon} f(x)\delta(x - x_0)dx = f(x_0).$$

Hence, for $\epsilon \to 0^+$, this integral will give the same answer $f(x_0)$, whereas for an ordinary integral of calculus and $f(x)$ continuous the answer would be $\mathrm{O}(\epsilon)$ as $\epsilon \to 0^+$ and thus zero in the limit. Consequently, the integral with $\delta(x - x_0)$ can be ignored away from the point of concentration x_0. The delta function, $\delta(x - x_0)$, is also called an ***impulse function*** *when it is used to impart an impulse to drive a differential equation.*

A simple constructive approximation that in the limit leads to the delta function $\delta(x)$ is the simple triangular approximation

$$d_\epsilon(x) \equiv \frac{1}{\epsilon}\begin{Bmatrix} (1 - |x|/\epsilon), & 0 \le |x| \le \epsilon \\ 0, & \epsilon \le |x| \end{Bmatrix}. \tag{B.161}$$

Now consider an arbitrary test function $f(x)$ that is continuous and continuously differentiable. Then using the definition (B.161),

$$\begin{aligned}\int_{-\infty}^{+\infty} d_\epsilon(x) f(x) dx &= \frac{1}{\epsilon}\int_{-\epsilon}^{+\epsilon} (1-|x|/\epsilon) f(x) dx \\ &= \int_{-1}^{+1} (1-|y|) f(\epsilon y) dy \\ &= \int_{-1}^{+1} (1-|y|)[f(0)+\mathrm{O}(\epsilon)] dy \\ &= f(0)+\mathrm{O}(\epsilon) \to f(0)\end{aligned}$$

as $\epsilon \to 0^+$. Since $d_\epsilon(x)$ has the same effect at $\delta(x)$ in the limit, it can be said that

$$\delta_{+0}(x) = \lim_{\epsilon\to 0^+} \overset{\text{gen}}{=} \delta(x),$$

where the symbol of generalized equality is $\overset{\text{gen}}{=}$ defined as follows.

Definition B.70. ***Generalized Equality.***
Let

$$g(x) \overset{\text{gen}}{=} h(x)$$

if for a sufficient class of test functions, $f(x)$ (sufficiently smooth, bounded with exponential decay as $x \to \infty$, depending on the application) both $g(x)$ and $h(x)$ have the same effect in integration,

$$\int_{-\infty}^{+\infty} f(x)g(x)dx = \int_{-\infty}^{+\infty} f(x)h(x)dx.$$

Using the Wiener process density $\phi_{W(t)}(w)$ (B.23), it can also be shown that in the generalized sense,

$$\phi_{W(0^+)}(w) \overset{\text{gen}}{=} \delta(w). \tag{B.162}$$

The generalized result (B.162) is obtained by examining the asymptotic limit as $t \to 0^+$,

$$\mathrm{E}[f(W(t))] = \int_{-\infty}^{+\infty} f(w)\phi_n(w;0,t)dw \to f(0),$$

for a continuous, exponentially bounded test function $|f(w)| < K\exp(aw)$ for some $K > 0$ and $a < a_0$ for some a_0 is sufficient, since the negative quadratic exponent of the density dominates any simple exponential at infinity. One need only consider the finite interval $[-R, R]$ for some sufficiently large R, $R/\sqrt{t} \gg 1$ when $t \ll 1$ will suffice, so that the tail portion of the integral on $(-\infty, +\infty)$ is negligible.

Remarks B.71.

- *The technique suggested above is Laplace's method for integrals given in* (B.153)*; see also* [61, 28]*, for instance, or Exercise* 23.

- *Since we are interested here in limits of the normal distribution and its density, and the density has a delta function limit such that* $\phi_{W(0^+)}(w) \overset{\text{gen}}{=} \delta(w)$ *according to* (B.162)*, then the use of the* $H(x)$ *step function form* (B.156) *in the relation* $\Phi_X(\xi) = \mathrm{E}_X[H(\xi - X)]$ (B.159) *is inconsistent. This is because* $\Phi_{\Delta W(t)}(0) = 1/2$ *for all positive values of* Δt*, so*

$$\Phi_{W(0^+)}(w) = \int_{-\infty}^{w} \delta(v)dv = \begin{Bmatrix} 0, & w < 0 \\ 1/2, & w = 0 \\ 1, & w > 0 \end{Bmatrix} = H_a(w)$$

or (B.157)*, since the averaged value at zero is needed. However, using the expectation form of the distribution* (B.159) *(normally, products of delta functions cannot be made), then*

$$\mathrm{E}[H(w - W(0^+))] = \int_{-\infty}^{+\infty} H(w - v)\delta(v)dv = H(w),$$

which is incorrect if $w = 0$ *when using the generalized limits for the normal density.*

Examples B.72. ***Generalized Function.***

- $\delta(ax + b) \overset{\text{gen}}{=} (1/a)\delta(x + b/a)$ *for constant* $a > 0$ *and* b *by changing variables* $\xi = ax$ *in the integral definition* (B.160)*.*

- $\delta(-x) \overset{\text{gen}}{=} \delta(x)$*, i.e.,* $\delta(x)$ *behaves as an even function, since* $f(0^-) = f(0)$ *if the function* f *is continuous.*

- $x\delta(x) \overset{\text{gen}}{=} 0$*, since by* (B.160) *with any* $f(x) = xF(x)$*,* $F(x)$ *continuous and* $x_0 = 0$*,*

$$\int_{-\infty}^{+\infty} F(x)x\delta(x)dx = 0 \cdot F(0) = 0.$$

- *Let* $f(x)$ *be any continuously differentiable function on* $\mathbb{R}$*. Then the derivative of the Dirac delta function* $\delta'(x)$ *is defined by*

$$\int_{-\infty}^{+\infty} f(x)\delta'(x)dx = -f'(0). \tag{B.163}$$

The motivation for this definition is the integration by parts calculus tool that

$$\int_{-\infty}^{+\infty} f(x)\delta'(x)dx = \left[f(x)\delta(x) - \int f'(x)\delta(x)dx \right]\Bigg|_{-\infty}^{+\infty} = -f'(0),$$

where the fact that $\delta(x)$ *is concentrated at* $x = 0$ *means the* $f(x)\delta(x)$ *vanishes at infinity since* $\delta(x)$ *dominates by vanishing faster than any* $f(x)$ *can grow. An alternate motivation is to use the original definition of* $\delta(x - x_0)$ *in* (B.160) *and*

assume that $\delta(x - x_0)$ is differentiable under the integral, i.e., it has been generated by a continuously differential approximation satisfying uniformity conditions. Then

$$\frac{d}{dx_0}\int_{-\infty}^{+\infty} f(x)\delta(x - x_0)dx = -\int_{-\infty}^{+\infty} f(x)\delta'(x - x_0)dx = f'(x_0)\,, \qquad \text{(B.164)}$$

the minus sign arising from differentiating $(x - x_0)$ with respect to x_0 as a simple application of the chain rule.

- *Similarly, $\delta''(x)$ for a twice continuously differentiable function f is defined in the generalized sense by*

$$\int_{-\infty}^{+\infty} f(x)\delta''(x)dx = +f''(0), \qquad \text{(B.165)}$$

derivable by two integrations by parts and using the concentration at $x = 0$. The same result also follows by differentiating the integral definition of $\delta(x - x_0)$ in (B.160) *twice.*

- *$H'(x) \overset{\text{gen}}{=} \delta(x)$ with respect to continuous function $f(x)$ for which $f(x)$ and its derivative vanish as $|x| \to \infty$, since by integration by parts,*

$$\int_{-\infty}^{+\infty} f(x)H'(x)dx = \left[f(x)H(x) - \int f'(x)H(x)dx\right]\Bigg|_{-\infty}^{+\infty}$$

$$= -\int_{0}^{+\infty} f'(x)dx = f(0).$$

An alternate motivation for this result, is to start with the original definition of the Heaviside step function,

$$\frac{d}{dx_0}\int_{-\infty}^{+\infty} f(x)H(x - x_0)dx = -\int_{-\infty}^{+\infty} f(x)H'(x - x_0)dx$$

$$= -f(x_0)dx, \qquad \text{(B.166)}$$

so ignoring the two minus signs, we have $H'(x - x_0) \overset{\text{gen}}{=} \delta(x - x_0)$.

- *A discrete distribution can be transformed into a continuous distribution by using a sequence of delta functions such that the density for the discrete random variable X with $(m + 1)$ possible discrete values $\{x_k | k = 0 : m\}$ each with probability π_k, such that the generalized density is given by*

$$\phi_X^{(\text{gen})}(x) \overset{\text{gen}}{=} \sum_{k=0}^{m} \pi_k \delta(x - x_k).$$

Hence, the expectation of some function $f(x)$ *is*

$$\begin{aligned}\mathrm{E}_X^{(\mathrm{gen})}[f(X)] &= \int_{-\infty}^{+\infty} f(x)\phi_X(x)dx = \sum_{k=0}^{m}\pi_k \int_{-\infty}^{+\infty} f(x)\delta(x-x_k)dx \\ &= \sum_{k=0}^{m}\pi_k f(x_k),\end{aligned}$$

which is the same formula as given in (B.48) *previously. Also, conservation of probability is confirmed by*

$$\mathrm{E}_x^{(\mathrm{gen})}[1] = 1$$

using the discrete probability property (B.46). *However, the implied probability distribution* $\Phi_X^{(\mathrm{gen})}(x)$ *is problematic since neither definition,* $H(x-x_k)$ *or* $H_a(x-x_k)$*, of the step function is suitable at* $x = x_k$*, but see the appropriate right-continuous step function* $H_R(x)$ *ahead in* (B.169).

Since it is an aim of the text to treat continuous and discrete distributions together, a unified applied treatment is needed. For this treatment, generalized functions [185, 89], primarily step and delta functions, will be used for discrete distributions in a manner similar to the way they are used in differential equations, but more suited to stochastic processes. Thus, the *continued* discrete distribution will be illustrated and defined for the Poisson process since the probabilities are already ordered by integer values.

Lemma B.73.

- *The* ***Poisson distribution made right-continuous (RC)*** *is*

$$\Phi_{P(t)}(X) = \mathrm{Prob}[X \le x] = \begin{Bmatrix} \sum_{j=0}^{\lfloor x \rfloor} p_j(\lambda t)\,, & x \ge 0 \\ 0\,, & x < 0 \end{Bmatrix}, \tag{B.167}$$

  which readily follows, and where $\lfloor x \rfloor$ *is the integer* ***floor function*** *such that* $x - 1 < \lfloor x \rfloor \le x$.

- *In terms of the* ***generalized RC step-function*** $H_R(x)$ *this Poisson distribution can be generalized to*

$$\Phi_{P(t)}(X) = \sum_{k=0}^{\infty} p_k(\lambda t)H_R(x-k) \tag{B.168}$$

  such that

$$H_R(x) = \begin{Bmatrix} 0\,, & x < 0 \\ 1\,, & x \ge 0 \end{Bmatrix}, \tag{B.169}$$

  where the property $H_R(0) = H_R(0^+)$ *and* $H_R(0^-) = 0$ *embodies the required right-continuity property. Clearly,* $\Phi_{P(t)}(X)$ *is right-continuous, rather than purely continuous.*

Proof. The distribution form (B.167) follows directly from the definition of the continuous distribution using the discrete Poisson distribution $\text{Prob}[P(t) = k] = p_k(\lambda t)$ for $k = 0 : \infty$. Thus,

$$\text{Prob}[P(t) \le x] = \sum_{j=0}^{k} p_j(\lambda t), \ k \le x < k+1$$

for $k = 0 : \infty$, since it takes k jumps for x to exceed k, i.e., $k = \lfloor x \rfloor$, so $k \le x < k+1$ is equivalent to $x - 1 < \lfloor x \rfloor \le x$, and any more will require the $(k+1)$st jump. Thus, the kth probability $p_k(\lambda t)$ is included in the sums if $x \ge k$, i.e., $p_k(\lambda t)$ is included in the form

$$p_k(\lambda t) H_R(x - k),$$

leading to (B.168). □

Definition B.74. *The **Poisson process density** corresponding to this continuous distribution is denoted by*

$$\phi_{P(t)}(X) = \sum_{k=0}^{\infty} p_k(\lambda t) \delta_R(x - k), \tag{B.170}$$

*where $\delta_R(x)$ is the **right-continuous (RC) delta function** such that*

$$H_R(x) = \int_{-\infty}^{x} \delta_R(y) dy \tag{B.171}$$

having the desired property that $H_R(0) = 1$ and the integral property

$$\int_{-\infty}^{\infty} f(y) \delta_R(y) dy = f(0^-). \tag{B.172}$$

These generalized functions and their properties will be encountered in more detail later in this text. The generalized $H_R(x)$ function is somewhat different from the concretely defined $H(x)$ in (B.156). Also, if the function f is continuous at $x = 0$ in B.172, then $f(0^-)$ can be replaced by $f(0)$.

The relationship between the exponential distribution and the Poisson distribution follows from the time of the arrival of the first jump T_1 under the standard assumption that the Poisson processes $P(t)$ starts at $t = T_0 \equiv 0$ and that the distribution for the first jump is the same as the probability that the Poisson jump-counter exceeded one, i.e.,

$$\begin{aligned} \Phi_{T_1}(t; \lambda) \equiv \text{Prob}[T_1 \le t] = \text{Prob}[P(t) \ge 1] &= \sum_{k=1}^{\infty} p_k(\lambda t) \\ = \sum_{k=1}^{\infty} e^{-\lambda t} \frac{(\lambda t)^k}{k!} &= e^{-\lambda t} \left(e^{\lambda t} - 1 \right) = 1 - e^{-\lambda t}, \end{aligned} \tag{B.173}$$

which is the same result as (B.40). The same result holds for the interarrival time, $T_{k+1} - T_k$, between successive Poisson jumps, except that the more general result depends on the property of stationarity of the Poisson process that is introduced in Chapter 1.

Summarizing distribution properties for combinations of continuous random variables and right-continuous jump processes, we have the following.

Properties B.75. ***Right-Continuous Distribution Functions*** $\boldsymbol{\Phi(x)}$**.**

- Φ *is* ***nondecreasing****, since probabilities must be nonnegative.*
- Φ *is* ***right-continuous****, by properties of integrals with nonnegative integrands including integrands with right-continuous delta functions or probability masses.*
- $\Phi(-\infty) = +0$ *by properties of integrals and* $X > -\infty$*.*
- $\Phi(+\infty) = 1$ *if* Φ *is a* ***proper distribution****.*

B.13 Fundamental Properties of Stochastic and Markov Processes

B.13.1 Basic Classification of Stochastic Processes

The classification of stochastic processes is important since the classification leads to the appropriate method of treatment of the stochastic process applications.

A **stochastic process** or **random process** is a random function of time $\xi = X(t; \omega)$, where $X(t; \omega)$ is a random variable depending on time t and some underlying random variable ω on the sample space Ω. (Again the ω dependence will often be suppressed unless it is needed to describe some stochastic process attribute.)

If the time domain is continuous on some interval $[0, T]$, then it is said to be a **stochastic processes** in **continuous time** whether the domain is bounded or unbounded. However, if the time domain is discrete, $\xi = X_i$ in discrete time units $i = 1 : \infty$ called **stages**, then it is a **stochastic process** in **discrete time** or random sequence. If $\xi = X(t)$ is not a random variable, then $X(t)$ would be called a **deterministic process**.

Stochastic processes are also generally classified according to the properties of the range of the random variable $\xi = X(t)$, called the **state space** of the process. This state space can be continuous, in which case it is still referred to as a stochastic process, but if the state space is discrete with a finite or infinite number of states, then the stochastic process is called a chain. The Gaussian process is an example of a process with a continuous state space, while the simple Poisson process with unit jumps is an example of a process with a discrete state space. A mixture of Gaussian and Poisson processes, called a jump-diffusion, is an example of a **hybrid stochastic system**.

B.13.2 Markov Processes and Markov Chains

An important class of stochastic processes is the Markov process $X(t)$ in which the future state depends on only some current state but not on a past state. This Markov property offers many advantages in the analysis of the behavior of these processes.

Definitions B.76.

- *A stochastic process $X(t)$ for $t \geq 0$ in continuous time (ct) is a **Markov process** on a continuous state (cs) space $\mathcal{S}_{\text{csct}}$ if for any $t \geq 0$, $\Delta t \geq 0$, and $x \in \mathcal{S}_{\text{csct}}$,*

$$\text{Prob}[X(t+\Delta t) \leq x | X(s), s \leq t] = \text{Prob}[X(t+\Delta t) \leq x | X(t)]. \quad \text{(B.174)}$$

- *A stochastic process X_i for $i = 0:\infty$ in discrete time (dt) is a **Markov process** on a continuous state space $\mathcal{S}_{\text{csdt}}$ if for any $n = 0:\infty$, $i = 0:\infty$, and $x_n \in \mathcal{S}_{\text{csdt}}$,*

$$\begin{aligned}\text{Prob}[X_{n+1} \leq x_{n+1} | X_i = x_i,\ i = 0:n] \\ = \text{Prob}[X_{n+1} \leq x_{n+1} | X_n = x_n].\end{aligned} \quad \text{(B.175)}$$

- *A stochastic process $X(t)$ for $t \geq 0$ in continuous time is a **Markov chain** on a discrete state (ds) space $\mathcal{S}_{\text{dsct}} = \{0, 1, 2, \dots\}$ if for any $t \geq 0$, $\Delta t \geq 0$, and $j(t) \in \mathcal{S}_{\text{dsct}}$,*

$$\begin{aligned}\text{Prob}[X(t+\Delta t) = j(t+\Delta t) | X(s) = j(s), s \leq t] \\ = \text{Prob}[X(t+\Delta t) = j(t+\Delta t) | X(t) = j(t)].\end{aligned} \quad \text{(B.176)}$$

- *A stochastic process X_i for $i = 0:\infty$ in discrete time is a **Markov chain** on a discrete state space $\mathcal{S}_{\text{dsdt}} = \{0, 1, 2, \dots\}$ if for any $n = 0:\infty$, $i = 0:\infty$, and $j_i \in \mathcal{S}_{\text{dsdt}}$,*

$$\begin{aligned}\text{Prob}[X_{n+1} = j_{n+1} | X_i = j_i,\ i = 0:n] \\ = \text{Prob}[X_{n+1} = j_{n+1} | X_n = j_n].\end{aligned} \quad \text{(B.177)}$$

  *The conditional probability $\text{Prob}[X_{n+1} = j_{n+1} | X_n = j_n] = P_{n,n+1}(j_n, j_{n+1})$ is called the **transition probability** for the step from stage n to stage $n+1$.*

Thus, the Markov process can be called **memoryless** or **without after-effects** since, for example, in the continuous time case, the **future state** $X(t+\Delta t)$ depends only on the **current state** $X(t)$, but not on the **past states** $\{x(s), s < t\}$. This memoryless property of Markov processes leads immediately to the **independent increments** property of Markov processes:

Lemma B.77. *If $X(t)$ is a Markov process in continuous time, then the state **increment** $\Delta X(t) \equiv X(t+\Delta t) - X(t)$ is **independent** of $\Delta X(s) \equiv X(s+\Delta s) - X(s)$, $s, t, \Delta s, \Delta t \geq 0$, if the time intervals are disjoint except for trivial overlap, i.e., either $s+\Delta s \leq t$ or $t+\Delta t \leq s$, such that*

$$\begin{aligned}\Phi_{\Delta X(t),\Delta X(s)}(\Delta x, \Delta y) &\equiv \text{Prob}[\Delta X(t) \leq \Delta x, \Delta X(s) \leq \Delta y] \\ &= \text{Prob}[\Delta X(t) \leq \Delta x]\text{Prob}[\Delta X(s) \leq \Delta y].\end{aligned}$$

Note that the Markov property definition can be reformulated as

$$\text{Prob}[X(t+\Delta t) \leq x + \Delta x | X(s), s < t; X(t) = x] = \text{Prob}[\Delta X(t) \leq \Delta x | X(t) = x]$$

and thus is independent of any increments in the past.

B.13.3 Stationary Markov Processes and Markov Chains

Definition B.78. *A Markov process is called* ***stationary*** *or* ***time-homogeneous*** *if the probability distribution depends only on the time difference, i.e.,*

- *If* $\text{Prob}[X(t+\Delta t) - X(t) \leq y] = \text{Prob}[\Delta X(t) \leq y]$ *depends on* $\Delta t \geq 0$ *and is independent of* $t \geq 0$ *in the continuous time case given* y *in the state space, continuous or discrete, or*
- *If* $\text{Prob}[X_{i+k} - X_i \leq y]$ *depends on* $k \geq 0$ *and is independent of* $i \geq 0$ *in the discrete time case given* y *in the state space, continuous or discrete. (It is also said that the* ***transition probabilities*** *are stationary.)*

The stationary Markov chain in discrete time is fully characterized by the **transition probability matrix** $[P_{i-1,j-1}]_{N\times N}$, where $P_{i,j} = \text{Prob}[X_{n+1} = j | X_n = i]$ for all stages $n = 0 : N-1$, where N may be finite or infinite [265]. Although the main focus here is on Markov processes in continuous time, Markov chains serve as numerical approximation for Markov processes, such as in the Markov chain approximation methods of Kushner and coworkers [175, 176, 179].

B.14 Continuity, Jump Discontinuity, and Nonsmoothness Approximations

In the standard calculus, much of the emphasis is on functions that are continuous, differentiable, continuously differentiable, or have similar nice properties. However, many of the models for Markov processes do not always have such nice analytical properties, since Poisson processes are discontinuous and Gaussian processes are not smooth. Thus, the standard calculus will be reviewed and revised to include the not-so-nice but essential properties.

B.14.1 Beyond Continuity Properties

If $X(t)$ is a process, i.e., function of time, whether stochastic or deterministic, the basic differences are summarized below.

Definitions B.79.

- *Let the* ***increment*** *for the process* $X(t)$ *be* $\Delta X(t) \equiv X(t+\Delta t) - X(t)$*, where* Δt *is the time increment.*
- *Let the* ***differential*** *for the process* $X(t)$ *be* $dX(t) \equiv X(t+dt) - X(t)$ *with respect to the time* t*, where* dt *is the infinitesimal time differential.*
- *The increment and differential are precisely related by the integral*

$$\Delta X(t) = \int_t^{t+\Delta t} dX(s).$$

While much of the regular calculus is usually cast in a more abstract form, much of applied stochastic calculus is based on differentials and increments, so the following will be formulated with increments or differentials, ready to use.

Definitions B.80.

- *The process $X(t)$ is a **continuous process** at the point t_0 if*

$$\lim_{\Delta t \to 0} X(t_0 + \Delta t) = X(t_0)$$

provided the limit exists.

- *Else the process $X(t)$ is **discontinuous** at t_0.*
- *The process $X(t)$ is **continuous** on the interval (t_1, t_2) if it is continuous at each point of the interval.*
- *The process $X(t)$ has a **jump discontinuity** at t_0 if*

$$\lim_{\substack{\Delta t \to 0 \\ |\Delta t| > 0}} X(t_0 + \Delta t) \neq X(t_0)$$

provided both the limit exists, i.e., the limit from the left

$$X(t_0^-) = \lim_{\Delta t \to 0+} X(t_0 - \Delta t)$$

and does not agree with the limit from the right

$$X(t_0^+) = \lim_{\Delta t \to 0+} X(t_0 + \Delta t),$$

where $\Delta t \to 0+$ means $\{\Delta t \to 0, \Delta t > 0\}$. In other words, if

$$X(t_0^+) \neq X(t_0^-),$$

*then $X(t)$ has a **jump** at $t = t_0$ [169]. The corresponding **jump** at the **jump discontinuity** (discontinuity of the first kind) is defined as*

$$[X](t_0) \equiv X(t_0^+) - X(t_0^-) = \lim_{\Delta t \to 0^+} X(t_0 + \Delta t) - \lim_{\Delta t' \to 0^+} X(t_0 - \Delta t'). \tag{B.178}$$

- *The process $X(t)$ is **right-continuous** at t_0 if*

$$\lim_{\substack{\Delta t \to 0 \\ \Delta t > 0}} X(t_0 + \Delta t) = X(t_0)$$

*such that the **jump of** X **at** t is defined as*

$$[X](t_0) \equiv X(t_0) - X(t_0^-), \tag{B.179}$$

*since $X(t_0^+) = X(t_0)$. **Left-continuous processes** are similarly defined.*

Remark B.81. *The jump definition is consistent with the definition of the increment and consequently the differential, since if there is a jump at time t_1, then $dX(t_1^-) = X(t_1^- + dt) - X(t_1^-) = X(t_1^+) - X(t_1^-) = [X](t_1)$, accepting the convention that dt is both positive and infinitesimal so that $X(t_1^- + dt) = X(t_1^+)$. Similarly, for the increment $\Delta X(t_1^-) \to [X](t_1)$ as $\Delta t \to 0^+$.*

Definitions B.82.

- *The process $X(t)$ is **smooth** at t_0 if*

$$\lim_{\Delta t \to 0} \Delta X(t_0)/\Delta t$$

  exists, i.e., $X(t)$ is differentiable at t_0.

- *Else the process $X(t)$ is **nonsmooth**.*

Remark B.83. *For example, if $\Delta X(t_1) \sim C\sqrt{\Delta t}$ for some nontrivial constant C, then $\Delta X(t_1) \to 0$ and $\Delta X(t_1)/\Delta t \sim C/\sqrt{\Delta t} \to \infty$ as $\Delta t \to 0^+$, so $X(t)$ is continuous but not smooth at t_1.*

B.14.2 Taylor Approximations of Composite Functions

Construction of application models often relies on **Taylor's formula** with remainder (Lagrange form) for small perturbations about some given point, given here in the following form.

Theorem B.84. ***Taylor Approximation for a Scalar-Valued Function of a Scalar Argument, $f(x)$.***
Let the function $f(x)$ be defined, continuous, and $(n+1)$ times continuously differentiable for $|\Delta x| < R$, then

$$f(x + \Delta x) = \sum_{m=0}^{n} \frac{f^{(m)}(x)}{m!}(\Delta x)^m + \frac{f^{(n+1)}(x + \theta \Delta x)}{(n+1)!}(\Delta x)^{n+1}, \qquad \text{(B.180)}$$

*where $f^{(m)}(x)$ is the mth order derivative of f at x, $\theta \in (0, 1)$ is the relative location of the mean value point $x + \theta \Delta x$ in the **remainder term**, and R is the **convergence radius**.*

Further, if the highest derivative $f^{(n+1)}$ is bounded on the interval of convergence, $|\Delta x| < R$, then the remainder

$$S_n(x; \Delta x) - f(x + \Delta x) = O((\Delta x)^{n+1}),$$

as $\Delta x \to 0$, where

$$S_n(x; \Delta x) \equiv \sum_{m=0}^{n} \frac{f^{(m)}(x)}{m!}(\Delta x)^m$$

is the **partial sum** of the first $(n+1)$ terms for $m = 0 : n$.

For most applications, only a few terms are needed, while for stochastic applications in continuous time this form will be applied when the variable x is a process like $X(t)$. More generally, the interest is in functions that depend explicitly on time t and implicitly on time through the process $X(t)$, like $F(X(t), t)$. This is illustrated for a deterministic process increment in function $F(X(t), t)$, three times continuously differentiable in both t and x. First, the increment is split up to partially separate out the first argument $X(t)$-process and second t-argument explicit time changes so that the one-dimensional Taylor approximation (B.180) can be separately applied to the component parts. Using partial derivatives, we have the next theorem.

Theorem B.85. ***Taylor Approximation for a Scalar-Valued Function of a Scalar-Argument $X(t)$ and Time t, $f(X(t), t)$.***
Let $f(x, t)$ be three times differentiable in both x and t, let the process $X(t)$ be continuous, and let $\Delta X(t) = X(t + \Delta t) - X(t)$ so $X(t + \Delta t) = X(t) + \Delta X(t)$. Then

$$\begin{aligned}
\Delta f(X(t), t) &\equiv f(X(t) + \Delta X(t), t + \Delta t) - f(X(t), t) \\
&= (f(X(t) + \Delta X(t), t + \Delta t) - f(X(t) + \Delta X(t), t)) \\
&\quad + (f(X(t) + \Delta X(t), t) - f(X(t), t)) \\
&= \frac{\partial f}{\partial t}(X(t), t)\Delta t + \frac{\partial f}{\partial x}(X(t), t)\Delta X(t) \qquad \text{(B.181)} \\
&\quad + \frac{1}{2}\frac{\partial^2 f}{\partial t^2}(X(t), t)(\Delta t)^2 + \frac{\partial^2 f}{\partial t \partial x}(X(t), t)\Delta t \Delta X(t) \\
&\quad + \frac{1}{2}\frac{\partial^2 f}{\partial x^2}(X(t), t)(\Delta X)^2(t) \\
&\quad + \mathrm{O}((\Delta t)^3) + \mathrm{O}((\Delta t)^2 \Delta X) + \mathrm{O}(\Delta t (\Delta X)^2) + \mathrm{O}((\Delta X)^3)
\end{aligned}$$

as $\Delta t \to 0$ and $\Delta X(t) \to 0$.

Remarks B.86.

- *Keeping the second order partial derivative terms written out explicitly is in anticipation that, although the process may be continuous, the process may not be smooth as in the case of the Gaussian process.*
- *The above expansion can be extended to vector processes $\mathbf{X}(t) = [X_i(t)]_{nx\times 1}$ and is best expanded by components.*
- *Another difference with the stochastic cases is that X will also be a function of the underlying probability space variable ω, so $X = X(t; \omega)$ and $\Delta X = \Delta X(t; \omega) \to 0$ in probability (only) as $\Delta t \to 0^+$. Since $\Delta X(t; \omega)$ may have an unbounded range, e.g., in the case that $\Delta X(t; \omega)$ is normally distributed as $\Delta t \to 0^+$, but $\Delta t > 0$, the boundedness part of the order symbol definition* O *would be invalid if, for instance, the ΔX in $\mathrm{O}^3(\Delta X)$ were replaced by $\Delta X(t; \omega)$. However, something like $\mathrm{O}(\mathrm{E}[\Delta X^3(t; \omega)])$ would be valid. Nevertheless, formula* (B.181) *will be useful as a preliminary or formal expansion calculation, prior to applying an expectation and neglecting very small terms.*

In the case where the space process is a vector function of time, then performing the Taylor expansion by components facilitates the calculation of the Taylor approximation.

Theorem B.87. ***Taylor Approximation for a Scalar-Valued Function of a Vector-Argument*** $\mathbf{X}(t)$ ***and Time*** t, $f(\mathbf{X}(t), t)$.
Let $f(\mathbf{x}, t)$ *be three times differentiable in both* $\mathbf{x}$ *and* t, *let the column vector process* $\mathbf{X}(t) = [X_i]_{nx\times 1}$ *be continuous, i.e., by component, and let* $\Delta\mathbf{X}(t) = \mathbf{X}(t + \Delta t) - \mathbf{X}(t)$ *so* $\mathbf{X}(t + \Delta t) = \mathbf{X}(t + \Delta t) + \Delta\mathbf{X}(t)$. *Then*

$$
\begin{aligned}
\Delta f(\mathbf{X}(t), t) &\equiv f(\mathbf{X}(t) + \Delta\mathbf{X}(t), t + \Delta t) - f(\mathbf{X}(t), t) \\
&= (f(\mathbf{X}(t) + \Delta\mathbf{X}(t), t + \Delta t) - f(\mathbf{X}(t) + \Delta\mathbf{X}(t), t)) \\
&\quad + (f(\mathbf{X}(t) + \Delta\mathbf{X}(t), t) - f(\mathbf{X}(t), t)) \\
&= \frac{\partial f}{\partial t}(\mathbf{X}(t) + \Delta\mathbf{X}(t), t)\Delta t \\
&\quad + \frac{1}{2}\frac{\partial^2 f}{\partial t^2}(\mathbf{X}(t) + \Delta\mathbf{X}(t), t)(\Delta t)^2 + \mathrm{O}((\Delta t)^3) \\
&\quad + \sum_{i=1}^{nx} \frac{\partial f}{\partial x_i}(\mathbf{X}(t), t)\Delta X_i(t) \\
&\quad + \sum_{i=1}^{nx}\sum_{j=1}^{nx} \frac{1}{2}\frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{X}(t), t)\Delta X_i(t)\Delta X_j(t) + \mathrm{O}(||\Delta\mathbf{X}||^3) \\
&= \frac{\partial f}{\partial t}(\mathbf{X}(t), t)\Delta t + \nabla_{\mathbf{x}}^{\top}[f](\mathbf{X}(t), t)\Delta\mathbf{X}(t) \qquad \text{(B.182)} \\
&\quad + \frac{1}{2}\frac{\partial^2 f}{\partial t^2}(\mathbf{X}(t), t)(\Delta t)^2 + \frac{1}{2}\Delta\mathbf{X}^{\top}(t)\nabla_{\mathbf{x}}\left[\nabla_{\mathbf{x}}^{\top}[f]\right](\mathbf{X}(t), t)\Delta\mathbf{X}(t) \\
&\quad + \nabla_{\mathbf{x}}\left[\frac{\partial f}{\partial t}\right](\mathbf{X}(t), t)\Delta\mathbf{X}(t)\Delta t \\
&\quad + \mathrm{O}((\Delta t)^3) + \mathrm{O}((\Delta t)^2||\Delta\mathbf{X}||) + \mathrm{O}(\Delta t||\Delta\mathbf{X}||^2) + \mathrm{O}(||\Delta\mathbf{X}||^3)
\end{aligned}
$$

as $\Delta t \to 0$ *and* $\Delta\mathbf{X}(t) \to \mathbf{0}$, *where the gradient of* f *is the vector*

$$\nabla_{\mathbf{x}}[f](\mathbf{X}(t), t) \equiv \left[\frac{\partial f}{\partial x_i}(\mathbf{X}(t), t)\right]_{nx\times 1},$$

the transpose vector is the row vector $\Delta\mathbf{x}^{\top} = [\Delta x_j]_{1\times nx}$, *and* $||\Delta\mathbf{x}||$ *is some norm, e.g., the infinite norm* $||\Delta\mathbf{x}||_{\infty} = \max_i[|Dx_i|]$.

In the case where there is a vector-valued function $\mathbf{f}$ depending on time t and a space process $\mathbf{X}(t)$ that is a vector function of time, then systematically performing the Taylor expansion by both $\mathbf{f}$ and $\mathbf{X}$ components as well as by the t argument of $\mathbf{f}$ and finally reassembling the results into matrix-vector form facilitates the calculation of the Taylor approximation.

Theorem B.88. ***Taylor Approximation for a Vector-Valued Function of a Vector-Argument*** $\mathbf{X}(t)$ ***and Time*** t, $\mathbf{f}(\mathbf{X}(t), t)$.
Let $\mathbf{f}(\mathbf{x}, t) = [f_i(\mathbf{x}, t)]_{nx\times 1}$ *be three times differentiable in both* $\mathbf{x}$ *and* t, *let the column vector process* $\mathbf{X}(t) = [X_i(t)]_{nx\times 1}$ *be continuous, i.e., continuous by component, and let* $\Delta\mathbf{X}(t) = \mathbf{X}(t+\Delta t) - \mathbf{X}(t)$ *so* $\mathbf{X}(t+\Delta t) = \mathbf{X}(t+\Delta t) + \Delta\mathbf{X}(t)$. *Then*

$$
\begin{aligned}
\Delta\mathbf{f}(\mathbf{X}(t), t) &\equiv \mathbf{f}(\mathbf{X}(t) + \Delta\mathbf{X}(t), t+\Delta t) - \mathbf{f}(\mathbf{X}(t), t) \\
&= \mathbf{f}(\mathbf{X}(t) + \Delta\mathbf{X}(t), t+\Delta t) - \mathbf{f}(\mathbf{X}(t), t) \\
&= [f_i(\mathbf{X}(t) + \Delta\mathbf{X}(t), t+\Delta t) - f_i(\mathbf{X}(t), t)]_{nx\times 1} \\
&= \Bigg[\frac{\partial f_i}{\partial t}(\mathbf{X}(t), t)\Delta t + \sum_{j=1}^{nx} \frac{\partial f_i}{\partial x_j}(\mathbf{X}(t), t)\Delta X_j(t) \\
&\quad + \frac{1}{2}\frac{\partial^2 f_i}{\partial t^2}(\mathbf{X}(t) + \Delta\mathbf{X}(t), t)(\Delta t)^2 + \sum_{j=1}^{nx} \frac{\partial^2 f_i}{\Delta t \partial x_j}(\mathbf{X}(t), t)\Delta X_j(t)\Delta t \\
&\quad + \frac{1}{2}\sum_{k=1}^{nx}\sum_{j=1}^{nx} \frac{\partial^2 f_i}{\partial x_k \partial x_j}(\mathbf{X}(t), t)\Delta X_j(t)\Delta X_k(t) \\
&\quad + \mathrm{O}((\Delta t)^3) + \mathrm{O}((\Delta t)^2||\Delta\mathbf{X}||) + \mathrm{O}(\Delta t||\Delta\mathbf{X}||^2) + \mathrm{O}(||\Delta\mathbf{X}||^3) \Bigg]_{nx\times 1} \\
&= \frac{\partial\mathbf{f}}{\partial t}(\mathbf{X}(t), t)\Delta t + \left(\Delta\mathbf{X}^\top(t)\nabla_{\mathbf{x}}\right)[\mathbf{f}](\mathbf{X}(t), t) \qquad \text{(B.183)} \\
&\quad + \frac{1}{2}\frac{\partial^2\mathbf{f}}{\partial t^2}(\mathbf{X}(t), t)(\Delta t)^2 + \left(\Delta\mathbf{X}^\top(t)\nabla_{\mathbf{x}}\right)\left[\frac{\partial\mathbf{f}}{\partial t}\right](\mathbf{X}(t), t)\Delta t \\
&\quad + \frac{1}{2}\left(\Delta\mathbf{X}(t)\Delta\mathbf{X}^\top(t)\right) : \left(\nabla_{\mathbf{x}}\nabla_{\mathbf{x}}^\top\right)[\mathbf{f}](\mathbf{X}(t), t) \\
&\quad + \mathrm{O}((\Delta t)^3) + \mathrm{O}((\Delta t)^2||\Delta\mathbf{X}||) + \mathrm{O}(\Delta t||\Delta\mathbf{X}||^2) + \mathrm{O}(||\Delta\mathbf{X}||^3)
\end{aligned}
$$

as $\Delta t \to 0$ *and* $\Delta\mathbf{X}(t) \to \mathbf{0}$, *where the gradient of* $\mathbf{f}$ *is premultiplied by the transpose of* $\Delta\mathbf{X}(t)$ *so that dimension of* $\mathbf{f}$ *is obtained,*

$$
(\Delta\mathbf{X}^\top(t)\nabla_{\mathbf{x}})[f](\mathbf{X}(t), t) \equiv \left[\sum_{j=1}^{nx} \Delta X_j(t)\frac{\partial f_i}{\partial x_j}(\mathbf{X}(t), t)\right]_{nx\times 1},
$$

the second order derivative Hessian is similarly arranged as a scalar-valued operator double dot product,

$$
\begin{aligned}
\left(\Delta\mathbf{X}(t)\Delta\mathbf{X}^\top(t)\right) : \left(\nabla_{\mathbf{x}}\nabla_{\mathbf{x}}^\top\right)[\mathbf{f}](\mathbf{X}(t), t) \equiv \Bigg[&\sum_{j=1}^{nx}\sum_{k=1}^{nx} \Delta X_j(t)\Delta X_k(t) \\
&\cdot \frac{\partial^2 f_i}{\partial x_k \partial x_j}(\mathbf{X}(t), t) \Bigg]_{nx\times 1},
\end{aligned}
$$

the transpose vector is the row vector $\Delta \mathbf{x}^\top = [\Delta x_j]_{1\times nx}$*, and* $||\Delta \mathbf{x}||$ *is some norm, e.g., the infinite norm* $||\Delta \mathbf{x}||_\infty = \max_i[|Dx_i|]$.

In general the double dot product is related to the trace of a matrix (B.120).

Definition B.89. ***Double Dot Product of Two Square Matrices.***

$$A : B \equiv \text{Trace}[AB] = \sum_{j=1}^{n}\sum_{k=1}^{n} A_{j,k}Bk, j \tag{B.184}$$

for square matrices A and B.

However, if the **process is discontinuous**, as it will be for the jumps of the Poisson process, then (B.181) is no longer valid since the assumption on $X(t)$ is not valid at the jump. Thus, if $X(t)$ has a jump discontinuity at $t = t_1$, then the most basic form for change in f, the jump, must be used.

Theorem B.90. ***"Zero Order Taylor Approximation" or Jump Function Limit for a Scalar-Valued Function of a Discontinuous Vector Process Argument*** $\mathbf{X}(t)$ ***and Time*** t***,*** $f(\mathbf{X}(t), t)$**.**

$$\Delta f(\mathbf{X}(t_1^-), t_1^-) \to [f](\mathbf{X}(t_1), t_1) \equiv f(\mathbf{X}(t_1^+), t_1^+) - f(\mathbf{X}(t_1^-), t_1^-) \tag{B.185}$$

as $\Delta t \to 0^+$.

This result extends the jump function definition (B.178). For right-continuous jumps t_1^+ can be replaced by t_1 (B.185) as in (B.179). The most fundamental changes in processes are the large jumps, such as crashes or rallies in financial markets or disasters and bonanzas in nature or machine failure and repair in manufacturing production. It is important to be able to handle jumps, even though the analysis may be much more complicated than for continuous processes.

B.15 Extremal Principles

Finding extremal properties, maxima and minima, through optimization is another area where nice function properties may be overemphasized, but for many optimal control applications, results are needed for more general functions, whether deterministic or random functions.

Definitions B.91. ***Extrema.***
Let $f(\mathbf{x})$ *be defined on some connected domain* $\mathcal{D}$ *in* $\mathbb{R}^m$.

- *Then* $f(\mathbf{x})$ *has a* ***global maximum*** *at* $\mathbf{x}^*$ *in* $\mathcal{D}$ *if* $f(\mathbf{x}) \leq f(\mathbf{x}^*)$ *for all* $\mathbf{x}$ *on* $\mathcal{D}$.
- *Similarly,* $f(\mathbf{x})$ *has a* ***global minimum*** *at some point* $\mathbf{x}^*$ *on* $\mathcal{D}$ *if* $f(\mathbf{x}) \geq f(\mathbf{x}^*)$ *for all* $\mathbf{x}$ *on* $\mathcal{D}$.

- *Often, such* ***global extrema*** *are called* ***absolute extrema****.*

- *Then* $f(\mathbf{x})$ *has a* ***local maximum*** *or* ***relative maximum*** *at* $\mathbf{x}^*$ *on* $\mathcal{D}$ *if there is a neighborhood,* $\mathcal{N}(\mathbf{x}^*)$ *of* $\mathbf{x}^*$ *on* $\mathcal{D}$*, such that* $f(\mathbf{x}^* + \Delta\mathbf{x}) \leq f(\mathbf{x}^*)$ *for sufficiently small* $|\Delta\mathbf{x}|$*.*

- *Similarly,* $f(\mathbf{x})$ *has a* ***local minimum*** *or* ***relative minimum*** *at* $\mathbf{x}^*$ *on* $\mathcal{D}$ *if there is a neighborhood,* $\mathcal{N}(\mathbf{x}^*)$ *of* $\mathbf{x}^*$ *on* $\mathcal{D}$*, such that* $f(\mathbf{x}^* + \Delta\mathbf{x}) \geq f(\mathbf{x}^*)$ *for sufficiently small* $|\Delta\mathbf{x}|$*.*

- *Often, such* ***local extrema*** *are called* ***relative extrema****.*

Remarks B.92.

- *The standard definition of* ***global extrema****, i.e., global maxima and global minima, covers all of the most extreme values, the biggest and the smallest, regardless of the analytic properties of the target function. The definition of global extrema is the most basic definition, the one we need to turn to when derivative methods fail. On the other hand, finding global extrema is very difficult in general and is by no means a closed problem.*

- *However, the standard definition of* ***local extrema*** *is as strictly interior extrema, due to the restriction that the neighbor be in the domain of interest, which would exclude* ***boundary extrema*** *which may include the extreme value being sought.*

- *The* ***general recipe for global extrema*** *is often given by the following:*

  1. *Find* ***local extrema****, usually restricted to where the target function is well behaved.*
  2. *Find* ***boundary extrema****, perhaps also restricted to points where the function is well behaved.*
  3. *Find the* ***function values at all points where the function is not well behaved****, i.e., discontinuous, nonsmooth, etc.*
  4. *Find the* ***most extreme values of all of the above*** *for the global extreme values.*

Theorem B.93. ***First Order Necessary Conditions for a Local Minimum (Maximum).*** *Let* $f(\mathbf{x})$ *be continuously differentiable in an open neighborhood* $\mathcal{N}(\mathbf{x}^*)$ *of* $\mathbf{x}^*$*. If* $\mathbf{x}^*$ *is a local minimum (maximum), then* $\nabla[f](\mathbf{x}^*) = \mathbf{0}$*.*

If $\nabla[f](\mathbf{x}^*) = \mathbf{0}$, then $\mathbf{x}^*$ is also called a **stationary point** or **interior critical point** of f. For proof see any good calculus or analysis text, else see Nocedal and Wright [221] for a proof using Taylor's approximation and for the following theorem.

Theorem B.94. ***Second Order Necessary and Sufficient Conditions for a Local Minimum (Maximum).***
Let $\nabla^2[f](\mathbf{x})$ *be continuous in an open neighborhood* $\mathcal{N}(\mathbf{x}^*)$ *of* $\mathbf{x}^*$*.*

- *If* $\mathbf{x}^*$ *is a local minimum (maximum) of* f*, then* $\nabla[f](\mathbf{x}^*) = \mathbf{0}$ *and* $\nabla^2[f](\mathbf{x})$ *is positive (negative) definite.*
- *If* $\nabla[f](\mathbf{x}^*) = \mathbf{0}$ *and* $\nabla^2[f](\mathbf{x})$ *is positive (negative) definite, then* $\mathbf{x}^*$ *is a minimum (maximum) of* f *.*

B.16 Exercises

Many of these exercises, depending on the instructor, can be done by using MATLAB, Maple, or Mathematica, but if theoretical, the MATLAB Symbolic Toolbox will be needed.

1. Prove the **variance-expectation identity** for any random variable X:

$$\mathrm{Var}[X] = \mathrm{E}[X^2] - \mathrm{E}^2[X]. \tag{B.186}$$

(Note that $\mathrm{E}^2[X] = (\mathrm{E}[X])^2$ *here, since squaring the operator also squares the value.)*

2. Prove the following identity for the **variance of the sum of two random variables** X and Y:

$$\mathrm{Var}[X+Y] = \mathrm{Var}[X] + 2\mathrm{Cov}[X,Y] + \mathrm{Var}[Y]. \tag{B.187}$$

3. Prove the following identity for the **variance of the product of two random variables** X and Y,

$$\begin{aligned}\mathrm{Var}[XY] = {}& \overline{X}^2\mathrm{Var}[Y] + 2\overline{XY}\mathrm{Cov}[X,Y] + \overline{Y}^2\mathrm{Var}[X] - \mathrm{Cov}^2[X,Y] \\ & + 2\overline{X}\mathrm{E}[\delta X(\delta Y)^2] + 2\overline{X}\mathrm{E}[(\delta X)^2\delta Y] + \mathrm{E}[(\delta X)^2(\delta Y)^2],\end{aligned}$$

where $\overline{X} = \mathrm{E}[X]$ and $\overline{Y} = \mathrm{E}[Y]$ are means, while $\delta X = X - \overline{X}$ and $\delta Y = Y - \overline{Y}$ are deviations from the mean. Further, in the case that X and Y are independent random variables, show that

$$\mathrm{Var}[XY] = \overline{X}^2\mathrm{Var}[Y] + \overline{Y}^2\mathrm{Var}[X] + \mathrm{Var}[X]\mathrm{Var}[Y]. \tag{B.188}$$

4. Prove the **Chebyshev inequality**,

$$\mathrm{Prob}[|X| \geq \epsilon] \leq \mathrm{E}[|X|^2]/\epsilon^2, \tag{B.189}$$

where $\epsilon > 0$.
(Hint: It is sufficient to assume that a probability density $\phi(x)$ *exists. Subtract the left-hand side from the right-hand side of the inequality, convert the expectation and probability to integrals, and then show that the sum is nonnegative.*

5. Prove the **Schwarz inequality** (**Cauchy–Schwarz inequality**) in terms of expectations,

$$\mathrm{E}[|XY|] \leq \sqrt{\mathrm{E}[X^2]\cdot\mathrm{E}[Y^2]}. \tag{B.190}$$

(Hint (big): Use the fact that $(u - v)^2 \geq 0$ *and let* $u = X/\sqrt{E[X^2]}$ *and* $v = Y/\sqrt{E[Y^2]}$*, assuming that X and Y have finite, positive variances. Alternatively, explore the characteristic roots of* $E[(\lambda X + Y)^2] \geq 0$ *and consider that if there are only real roots* λ_i *at the minimum, then the discriminant (square root argument) must be positive in the quadratic formula.)*

6. Prove **Jensen's inequality**: If f is a **convex function**, i.e., f is real and

$$f(\theta x + (1 - \theta)y) \leq \theta f(x) + (1 - \theta) f(y) \tag{B.191}$$

for all x, y and $0 < \theta < 1$, then

$$\mathrm{E}[f(X)] \geq f(E[X]). \tag{B.192}$$

7. (a) Derive this simple form of **Bayes' rule** for two related random variables X and Y:

$$\mathrm{Prob}[X = x|Y = y] = \frac{\mathrm{Prob}[Y = y,\ X = x]}{\mathrm{Prob}[Y = y]} \tag{B.193}$$

provided $\mathrm{Prob}[Y = y] > 0$.

(Hint: You need only to use the conditional probability definition (B.83)*.)*

(b) Derive, using an expansion of (B.193) and also the law of total probability (B.92), the multiple random variables or events form of **Bayes' rule** for the case of the random event Y that occurs in conjunction with a member of the exhaustive (complete) and countable set of disjoint (mutually exclusive) events, $\{X_i, i = 1 : n\}$, i.e., the total law of probability if applicable,

$$\mathrm{Prob}[X_i = x_i|Y = y] = \frac{\mathrm{Prob}[Y = y,\ X_i = x_i]}{\sum_{j=1} \mathrm{Prob}[Y = y,\ X_j = x_j] \cdot \mathrm{Prob}[X_j = x_j]}.$$

8. For the **uniform distribution**, confirm the formulas for the mean, variance, coefficient of skewness, and coefficient of kurtosis.

9. Derive the following identity between the standard normal and the general **normal distributions**:

$$\Phi_n((\ln(x) - \mu)/\sigma; 0, 1) = \Phi_n(\ln(x); \mu, \sigma^2).$$

10. Show, for the **lognormal density** with random variable $X_{ln}(t)$, that the **maximum location**, the **mode** of the distribution, or the **most likely value** is given by

$$x^* = \mathrm{Mode}\left[X_{ln}(t)\right] = \exp\left(\mu - \sigma^2\right).$$

Also, compare the **mean** or **expected** value to the **mode** for the lognormal distribution by calculating the ratio

$$\mathrm{E}\left[X_{ln}(t)\right] / \mathrm{Mode}\left[X_{ln}(t)\right];$$

then compare this lognormal ratio to that for the normal variates,

$$\mathrm{E}\left[X_n(t)\right] / \mathrm{Mode}\left[X_n(t)\right].$$

11. For the **exponential distribution**, confirm the formulas for the mean, variance, coefficient of skewness, and coefficient of kurtosis.

12. Show the following equivalence between the **exponential distribution** expectation and the **uniform distribution** expectation:

$$\mathrm{E}_e[f(X_e)] = \mathrm{E}_u[f(-\mu \ln(X_u))]$$

for any integrable function f.

13. Show the sample moment formulas for a set of IID random variables X_k with $\mathrm{E}[X_k] = \mu$ and $\mathrm{Var}[X_k] = \sigma^2$ for $k = 1 : n$ of Subsection B.6 are correct, i.e.,

    (a) $\mathrm{E}[m_n] = \mu$ for sample mean m_n (B.107);
    (b) $\mathrm{E}[s_n^2] = (n-1)\sigma^2/n$ for sample variance s_n^2 (B.108);
    (c) $\mathrm{E}[\widehat{s}_n^2] = \sigma^2$ for sample variance unbiased estimate $\widehat{s}_n^2$ (B.109);
    (d) $\mathrm{Var}[m_n] = \sigma^2/n$ for sample mean m_n .

(Hint: See Remarks B.54 *on p.* B37*.)*

14. Show that for a set of IID random variables, the covariance of the sample mean m_n and the sample variance s_n^2 satisfy

$$\mathrm{Cov}[m_n, s_n^2] = \mu_3/n,$$

where the third central moment is $\mu_3 = \mathrm{E}[(X_k - \mu)^3]$. Discuss what probability property relating m_n and s_n^2 is implied by the result if the IID distribution is even like the normal distribution and what property is implied asymptotically as $n \to +\infty$. See Subsection B.6.

15. Let $\mathcal{S} = \sum_{k=1}^{n} X_k$ be the **partial sum** of n IID random variables $\{X_k\}$ each with mean $\mathrm{E}[X_k] = \mu$ and variance $\mathrm{Var}[X_k] = \sigma^2$. Further, let the mth central moment be defined as $\mu^{(m)} = \mathrm{E}[(X_k - \mu)^m]$, so that $\mu^{(1)} = 0$ and $\mu^{(2)} = \sigma^2$. Show that

    (a) $\mathrm{E}[\mathcal{S}] = n\mu$;
    (b) $\mathrm{Var}[\mathcal{S}] = n\sigma^2$;
    (c) $\mathrm{E}[(\mathcal{S} - \mathrm{E}[\mathcal{S}])^3] = n\mu^{(3)}$ so is zero if the distribution of X_k has no **skew** (B.11);
    (d) $\mathrm{E}[(\mathcal{S} - \mathrm{E}[\mathcal{S}])^4] = n\mu^{(4)} + 3n(n-1)\sigma^2$, where the first term is related to the coefficient of **kurtosis** (B.12).

(Hint: Use the binomial theorem, $\mathcal{S} - \mathrm{E}[\mathcal{S}] = \sum_{k=1}^{n}(X_k - \mu)$ *and the fact* $\mu^{(1)} = 0$*.)*

16. Show that the **product of two normal densities** is proportional to a normal density, i.e.,

$$\phi_n(x; \mu_1, \sigma_1^2) \cdot \phi_n(x; \mu_2, \sigma_2^2) = \phi_n\left(x; \frac{\mu_1\sigma_2^2 + \mu_2\sigma_1^2}{\sigma_1^2 + \sigma_2^2}, \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2}\right) \cdot \frac{1}{\sqrt{2\pi(\sigma_1^2 + \sigma_2^2)}} \exp\left(-\frac{(\mu_1 - \mu_2)^2}{2(\sigma_1^2 + \sigma_2^2)}\right). \tag{B.194}$$

(Hint: Apply the completing the square technique to combine the two densities.)

17. Let X_i be independent normal random variables with density $\phi_{X_i}(x)$, mean μ_i, and variance σ_i^2 for $i = 1$ to K.

    (a) Show that the product of two normal densities is a normal density whose mean is the sum of the means and whose variances is the sum of the variance, using (B.194),

$$\begin{aligned} \mathcal{I}_2(x) \equiv \left(\phi_{X_1} * \phi_{X_2}\right)(x) &= \int_{-\infty}^{+\infty} \phi_{X_1}(x-y)\phi_{X_2}(y)dy \\ &= \phi_n(x; \mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2). \end{aligned} \tag{B.195}$$

    (b) Using (B.195) for $K = 2$ as the induction initial condition, show the general result by induction that

$$\mathcal{I}_K(x) \equiv \left(\left(\prod_{i=1}^{K-1} \phi_{X_i} *\right) \phi_{X_K}\right)(x) = \phi_n\left(x; \sum_{i=1}^{K} \mu_i, \sum_{i=1}^{K} \sigma_i^2\right). \tag{B.196}$$

18. Show that the distribution of the sum of two IID random variables, U_1 and U_2, uniformly distributed on $[a, b]$, is a **triangular distribution** on $[2a, 2b]$, i.e., show in terms of densities that

$$\begin{aligned} \phi_{U_1+U_2}(x) &= \int_{-\infty}^{+\infty} \phi_{U_1}(x-y)\phi_{U_2}(y)dy \\ &= \frac{1}{(b-a)^2} \begin{Bmatrix} (x-2a), & 2a \le x < b+a \\ (2b-x), & b+a \le x \le 2b \\ 0, & \text{otherwise} \end{Bmatrix}. \end{aligned} \tag{B.197}$$

    Confirm that the resulting density conserves probability on $(-\infty, +\infty)$.
    (Hint: It may be helpful to sketch the paths for nonzero integration in y on the xy-plane, paying attention to the limits of integration for each fixed x.)

    Remark B.95. *Different from the normal distribution results in Exercise* 17*, the convolution of two uniform random variables does not conserve the uniformity of the distribution.*

19. Show that the distribution of the sum of three IID random variables, U_i, for $i = 1 : 3$ uniformly distributed on $[a, b]$, is a piecewise **quadratic distribution** on $[3a, 3b]$, i.e., show in terms of densities that

$$\begin{aligned} \phi_{\sum_{i=1}^{3} U_i}(x) &= \int_{-\infty}^{+\infty} \phi_{U_1+U_2}(x-y)\phi_{U_3}(y)dy \\ &= \frac{1}{2(b-a)^3} \begin{Bmatrix} +(x-3a)^2, & 3a \le x < 2a+b \\ \begin{Bmatrix} -(x-(b+2a))^2 \\ +2(b-a)^2 \\ -(2b+a-x)^2 \end{Bmatrix}, & 2a+b \le x < a+2b \\ +(3b-x)^2, & a+2b \le x \le 3b \\ 0, & \text{otherwise} \end{Bmatrix} \end{aligned} \tag{B.198}$$

    using the result of the previous exercise for $\phi_{U_1+U_2}(x)$.

(Hint: With this and the previous exercise, symbolic computation may be more desirable, e.g,, Maple or Mathematica.)

20. For the **bivariate normal distribution**, verify the inverse of $\mathbf{\Sigma}$ in (B.143) and the explicit form for the density (B.144). Also, confirm by iterated integration that $\mathrm{E}[X_1] = \mu_1$, $\mathrm{Var}[X_1] = \sigma_1^2$, and $\mathrm{Cov}[X_1, X_2] = \rho\sigma_1\sigma_2$.
*(Hint: Only techniques such as **completing the square** and transformations to the generic integral*

$$\int_{-\infty}^{+\infty} \exp(-x^2/2)[c_0 + c_1 x + c_2 x^2]dx = \sqrt{2\pi}[c_0 \cdot 1 + c_2 \cdot 1]$$

for any constants $\{c_0, c_1, c_2\}$ are needed.)

21. For the **binomial distribution** in (B.148) verify that the given basic moments are correct, i.e., $\mathrm{E}[F_k] = N\pi_k$ and $\mathrm{Var}[F_k] = N\pi_k(1 - \pi_k)$ for $k = 1 : 2$.

22. Show that $W(0^+) = 0$ **with probability one** by showing that $\phi_{W(0^+)}(w) \overset{\text{gen}}{=} \delta(w)$, i.e., in the generalized sense, which means that

$$\mathrm{E}[f(W(t))] = \int_{\infty}^{+\infty} \phi_{W(t)}(w) f(w) dw \to f(0^+)$$

as $t \to 0^+$ for continuous, continuously differentiable, and sufficiently bounded functions $f(w)$ which vanish at infinity.
(Hint: For formal justification, scale t out of the density by a change of variables in the integral and expand f for small t, assuming that the exponential convergence property of the normal density allows termwise integration of the expansion. Note that $X(t)$ is in the set $\mathcal{S}$ with probability one simply means that $\mathrm{Prob}[X(t) \in \mathcal{S}] = 1$. *If more rigor is desired, use the asymptotic techniques, such as Laplace's method for integrals* (B.153) *on p.* B51 *from the text and Exercise* 23.*)*

23. **Asymptotic analysis, generalized function problem**:
Show that the following sequences for the approximate right-continuous step function $H_R(x)$ in (B.169) and the right-continuous delta function $\delta_R(x)$ in (B.171),

$$H_{R,n}(x) = \int_{-\infty}^{x} \delta_{R,n}(y)dy,$$
$$\delta_{R,n}(x) \equiv e^{-(y+\mu_n)^2/(2\epsilon_n)}/\sqrt{2\pi\epsilon_n},$$

are valid where $\epsilon_n > 0$, $\mu_n > 0$, $\sqrt{\epsilon_n} \ll \mu_n \ll 1$ when $n \gg 1$. That is, show for $n \gg 1$ that $H_{R,n}(0) = H_{R,n}(0^+) \sim 1$, $H_{R,n}(0^-) \to 0^+$, and

$$\int_{-\infty}^{+\infty} f(y)\delta_{R,n}(y - x)dy \sim f(x^-)$$

for any continuous function $f(x)$ that is exponentially bounded, $|f(x)| \leq Ke^{-a|x|}$ on $(-\infty, +\infty)$ with $a > 0$ and $K > 0$, justifying the use of $H_{R,n}(x) \to H_R(x)$ and $\delta_{R,n}(x) \to \delta_R(x)$ as $n \to \infty$ for the generalized representation of Poisson processes. *(Hint: When using the Laplace asymptotic approximation of integrals technique* [61, 28]*, change variables to* $\xi = y - x + \mu_n$*, select the integral tail-cutoff* $(-\rho_n, \rho_n)$ *in* ξ *about the argument of the maximum of* $\delta_{R,n}(\xi - \mu_n)$ *at* $\xi = 0$ *with* $\epsilon_n \ll \rho_n^2 \ll \mu_n \ll 1$ *so that the tails are exponentially negligible being dominated by the factor* $\exp(-\rho_n^2/(2\epsilon_n))$*, approximate* $f(x - \mu_n + \xi) \sim f(x - \mu_n)$ *using continuity, and then change variables to* $\eta = \xi/\sqrt{\epsilon_n}$ *so that the limits of integration can be expanded to* $\pm\infty$*. The order in which these approximations are performed is critical.)*

Suggested References for Further Reading

- Bartlett, 1978 [19]
- Bender and Orszag, 1978 [28]
- Çinlar, 1975 [56]
- Copson, 1965 [61]
- Cox and Miller, 1968 [63]
- Doob, 1953 [70]
- Feller, 1968 [84]
- Feller, 1971 [85]
- Friedman, 1956 [89]
- Glasserman, 2003 [97]
- Higham and Higham, 2000 [143]
- Karlin and Taylor, 1975 [162]
- Karlin and Taylor, 1981 [163]
- Lighthill, 1964 [185]
- Moler et al., 2000 [210]
- Neftci, 2000 [217]
- Nocedal and Wright, 1999 [221]
- Pliska, 1997 [225]
- Parzen, 1962 [224]
- Ross, 1983 [237]

- Ross, 2000 [238]
- Taylor and Karlin, 1998 [265]
- Taylor and Mann, 1972 [263]
- Tuckwell, 1995 [270]