Bruce K. Driver

Math 180B Stochastic Processes I, Spring 2020

March 31, 2020 *File:180B˙Notes.tex*

Contents

Part I

Background Material

1

Introduction

Definition 1.1 (Stochastic Process via Wikipedia). *..., a* ***stochastic process****, or often* ***random process****, is a collection of random variables representing the evolution of some system of random values over time. This is the probabilistic counterpart to a deterministic process (or deterministic system). Instead of describing a process which can only evolve in one way (as in the case, for example, of solutions of an ordinary differential equation), in a stochastic, or random process, there is some indeterminacy: even if the initial condition (or starting point) is known, there are several (often infinitely many) directions in which the process may evolve.*

1.1 Deterministic Modeling

In deterministic modeling one often has a dynamical system on a state space S. The dynamical system often takes on one of the two forms;

1. there exists $f : S \to S$ and a state x_n then evolves according to the rule $x_{n+1} = f(x_n)$. [More generally one might allow $x_{n+1} = f_n(x_0, \dots, x_n)$ where $f_n : S^{n+1} \to S$ is a given function for each n.
2. There exists a vector field f on S (where now $S = \mathbb{R}^d$ or a manifold) such that $\dot{x}(t) = f(x(t))$. [More generally, we might allow for $\dot{x}(t) = f\left(t; x|_{[0,t]}\right)$, a functional differential equation.]

Goals: the goals in this case then have to do with deriving the properties of the trajectories given the properties of the driving dynamics incorporated in f. For example, think of a golfer trying to make a put or a hot-air balloonist trying to find a path from point A to point B.

1.2 Stochastic Modeling

Much of our time in this course will be to explore the above two situations where some extra randomness is added at each state of the game. The point being that in many situations the exact nature of the dynamics is not known or is rapidly changing. What is known are statistical properties of the dynamics – i.e. likely hoods that the dynamics will be of a certain form. This amounts to replacing f above by some sort of random f and then resolving the problems. However, now rather than trying to find the properties of a given trajectory we instead try to find properties of the statistics of the now random trajectories. Typically when comparing theory to experiment one has to now average experimental results (hoping to use the law of large numbers) to make contact with the mathematical theory. Here is a little more detail on the typical sort of scenarios that we will consider in this course.

1. We may now have that $X_{n+1} \in S$ is random and evolves according to
$$X_{n+1} = f(X_n, \xi_n)$$
where $\{\xi_n\}_{n=0}^{\infty}$ is a sequence of i.i.d. random variables. Alternatively put, we might simply let $f_n := f(\cdot, \xi_n)$ so that $f_n : S \to S$ is a sequence of i.i.d. random functions from S to S. Then $\{X_n\}_{n=0}^{\infty}$ is defined recursively by
$$X_{n+1} = f_n(X_n) \text{ for } n = 0, 1, 2, \dots. \tag{1.1}$$
This is the typical example of a time-homogeneous Markov chain. We assume that $X_0 \in S$ is given with an initial condition which is either deterministic or is independent of the $\{f_n\}_{n=0}^{\infty}$.
2. Later in the course we will study the continuous time analogue,
$$\text{“}\dot{X}_t = f_t(X_t)\text{”}$$

where $\{f_t\}_{t\geq 0}$ are again i.i.d. random vector-fields. The continuous time case will require substantially more technical care. For example, one often considers the controlled differential equation,

$$\dot{X}_t = f(X_t)\,\dot{B}_t \tag{1.2}$$

where $\{B_t\}_{t\geq 0}$ is Brownian motion or equivalently $\dot{B}_t$ is "white noise" or B_t is a Poisson process. The Poisson noise is often used to model arrival times in networks or in queues (i.e. service lines) or appear in electrical circuits due to "thermal fluctuations" to name a few. See for example, Johnson Noise and Shot Noise by Dennis V. Perepelitsa, November 27, 2006.) Here are two quotes from this article.

"The thermal agitation of the charge carriers in any circuit causes a small, yet detectable, current to flow. J.B. Johnson was the first to present a quantitative analysis of this phenomenon, which is unaffected by the geometry and material of the circuit."

"The quantization of charge carried by electrons in a circuit also contributes to a small amount of noise. Consider a photoelectric circuit in which current caused by the photoexcitation of electrons flow to the anode."

3. We will also consider a class of processes known as (Sub/Super) martingales which encode information about fair (or not so fair) games of chance amongst many other applications.

2

(Discrete) Distributions Review

Notation 2.1 *We typically let Ω be a **sample space.** An element $\omega \in \Omega$ is a **sample point,** a subset $A \subset \Omega$ and **event,** and a function $X : \Omega \to \mathbb{R}$ a **random variable.** More generally a function, $X : \Omega \to \mathbb{R}^d$, is called a **random vector.** Even more generally we consider functions, $f : \Omega \to S$, where S is another set that we will often refer to as **state space.***

We are interested in probability functions, $\mathbb{P}$, on Ω and in particular in computing, $\mathbb{P}(A)$, the probability of an event A and, $\mathbb{E}X$, the expectation of a random variable X. We take for granted may of the basic notions of probability but, see Appendix A below for a refresher on some of the basic notions. We so however pause to discuss the notion of a distribution and give a number of example.

2.1 Discrete Distributions

Definition 2.2 (Discrete distributions). *If S is a discrete set, i.e. finite or countable and $X : \Omega \to S$ we let*

$$\rho_X(s) := \mathbb{P}(X = s).$$

If $Y : \Omega \to T$ is a another random function and again T is a discrete set, then $(X, Y) : \Omega \to S \times T$ is again a random function and we let

$$\rho_{X,Y}(s,t) := \mathbb{P}((X,Y) = (s,t)) = \mathbb{P}(X = s \text{ and } Y = t)$$

be the corresponding **joint distribution of** (X, Y). In this setting we say

$$\rho_X(s) = \mathbb{P}(X = s) = \sum_{t \in T} \mathbb{P}(X = s \text{ and } Y = t) = \sum_{t \in T} \rho_{X,Y}(s,t)$$

as the X**-marginal** of $\rho_{X,Y}$ and similarly

$$\rho_Y(t) = \mathbb{P}(Y = t) = \sum_{s \in S} \mathbb{P}(X = s \text{ and } Y = t) = \sum_{s \in S} \rho_{X,Y}(s,t)$$

is that Y**-marginal** of $\rho_{X,Y}$.

Remark 2.3. Given a function, $f : S \to \mathbb{R}$, we have

$$\mathbb{E}[f(X)] = \sum_{s \in S} f(s)\, \mathbb{P}(X = s) = \sum_{s \in S} f(s)\, \rho_X(s).$$

Theorem 2.4 (Independence). *If $X : \Omega \to S$ and $Y : \Omega \to T$ are as above, then the following statements are equivalent;*

1. *X and Y are **independent,** i.e.*

$$\mathbb{P}(X = s \text{ and } Y = t) = \mathbb{P}(X = s)\, \mathbb{P}(Y = t) \text{ for all } (s,t) \in S \times T.$$

2. *$\rho_{X,Y}(s,t) = \rho_X(s) \cdot \rho_Y(t)$ for all $(s,t) \in S \times T$.*
3. *$\mathbb{E}[f(X)\, g(Y)] = \mathbb{E}[f(X)]\, \mathbb{E}[g(Y)]$ for all bounded (or non-negative) functions $f : S \to \mathbb{R}$ and $g : T \to \mathbb{R}$.*

In order to give some examples of discrete random variables, let us take $S = \mathbb{N}_0 = \{0, 1, 2, \dots\}$ for the rest of this section.

Definition 2.5 (Generating Function). *Suppose that $N : \Omega \to \mathbb{N}_0$ is an integer valued random variable on a probability space, $(\Omega, \mathcal{B}, \mathbb{P})$. The generating function associated to N is defined by*

$$G_N(z) := \mathbb{E}\left[z^N\right] = \sum_{n=0}^{\infty} \mathbb{P}(N = n)\, z^n \text{ for } |z| \le 1. \tag{2.1}$$

The distribution of N may be recovered from G_N (by standard power series considerations) using;

$$\mathbb{P}(N=n)=\frac{1}{n!}G_N^{(n)}(0) \text{ for } n\in\mathbb{N}_0.$$

[The convention here is, as usual, that $0!=1$.]

Proposition 2.6 (Generating Functions). *The generating function satisfies,*

$$G_N^{(k)}(z)=\mathbb{E}\left[N(N-1)\dots(N-k+1)z^{N-k}\right] \text{ for } |z|<1$$

and

$$G^{(k)}(1)=\lim_{z\uparrow 1}G^{(k)}(z)=\mathbb{E}\left[N(N-1)\dots(N-k+1)\right],$$

where it is possible that one and hence both sides of this equation are infinite. In particular,

$$G'(1):=\lim_{z\uparrow 1}G'(z)=\mathbb{E}N \tag{2.2}$$

and if $\mathbb{E}N^2<\infty$,

$$\operatorname{Var}(N)=G''(1)+G'(1)-\left[G'(1)\right]^2. \tag{2.3}$$

Proof. By standard power series considerations, for $|z|<1$,

$$\begin{aligned}G_N^{(k)}(z)&=\sum_{n=0}^{\infty}\mathbb{P}(N=n)\cdot n(n-1)\dots(n-k+1)z^{n-k}\\&=\mathbb{E}\left[N(N-1)\dots(N-k+1)z^{N-k}\right].\end{aligned} \tag{2.4}$$

Since, for $z\in(0,1)$,

$$0\le N(N-1)\dots(N-k+1)z^{N-k}\uparrow N(N-1)\dots(N-k+1) \text{ as } z\uparrow 1,$$

we may apply the "Monotone Convergence Theorem" (MCT) to pass to the limit as $z\uparrow 1$ in Eq. (2.4) to find,

$$G^{(k)}(1)=\lim_{z\uparrow 1}G^{(k)}(z)=\mathbb{E}\left[N(N-1)\dots(N-k+1)\right].$$

■

Exercise 2.1 (Some Discrete Distributions). Let $p\in(0,1]$ and $\lambda>0$. In the four parts below, the distribution of N will be described. You should work out the generating function,

$$G_N(z):=\mathbb{E}\left[z^N\right]=\sum_{n=0}^{\infty}\mathbb{P}(N=n)z^n \text{ for } |z|\le 1$$

and then use it (along with Eqs. (2.2) and (2.3)) to verify the given formulas for $\mathbb{E}N$ and $\operatorname{Var}(N)$.

1. Bernoulli(p) : $\mathbb{P}(N=1)=p$ and $\mathbb{P}(N=0)=1-p$. You should find
$$\mathbb{E}N=p \text{ and } \operatorname{Var}(N)=p-p^2.$$
2. Binomial(n,p) = Bin(n,p) :
$$\mathbb{P}(N=k)=\binom{n}{k}p^k(1-p)^{n-k} \text{ for } k=0,1,\dots,n,$$
i.e. $\mathbb{P}(N=k)$ is the probability of k successes in a sequence of n independent yes/no experiments with probability of success being p.) You should find
$$\mathbb{E}N=np \text{ and } \operatorname{Var}(N)=n\left(p-p^2\right).$$
3. Geo(p) : $\mathbb{P}(N=k)=p(1-p)^{k-1}$ for $k\in\mathbb{N}$ is the **geometric distribution.** [$\mathbb{P}(N=k)$ is the probability that the k^{th} – trial is the first time of success out a sequence of independent trials with probability of success being p.] You should find
$$\mathbb{E}N=\frac{1}{p} \text{ and } \operatorname{Var}(N)=\frac{1-p}{p^2}.$$
4. Poisson(λ) : $\mathbb{P}(N=k)=\frac{\lambda^k}{k!}e^{-\lambda}$ for all $k\in\mathbb{N}_0=\{0,1,2,\dots\}$, see Proposition 2.11 below for some context. You should find
$$\mathbb{E}N=\lambda=\operatorname{Var}(N).$$

Definition 2.7 (Negative Binomial). *Let* $\{Z_i\}_{i=1}^{\infty}$ *be i.i.d. Bernoulli random variables with* $\mathbb{P}(Z_i=1)=p\in(0,1]$. *For* $r\in\mathbb{N}$ *let* W_r *be the number of* $Z_i=0$ *before the first time that* $Z_i=1$ *occurs for the* r^{th}*-time. In more detail, if* $W_r=k$, *then* $Z_{k+r}=1$ *(otherwise we would have* $W_r<k$*) and we then have* $\binom{r+k-1}{r-1}$ *ways to choose the other* $r-1$ *locations and each such configuration occurs with probability* $p^r(1-p)^k$. *Therefore the probability mass function for* W_r *is given by*

$$\mathbb{P}(W_r=k)=\binom{r+k-1}{k}p^r(1-p)^k=\binom{r+k-1}{r-1}p^rq^k \tag{2.5}$$

where $q=1-p$ *as usual.*

Remark 2.8. The binomial coefficient in Eq. (2.5) may be rewritten as

$$\begin{aligned}\binom{r+k-1}{k}&=\frac{(k+r-1)\cdots(r)}{(k)!}\\&=(-1)^k\frac{(-r)(-r-1)(-r-2)\cdots(-r-k+1)}{(k)!}=(-1)^k\binom{-r}{k}\end{aligned}$$

and hence we may also write,

$$\mathbb{P}(W_r = k) = (-1)^k \binom{-r}{k} p^r (1-p)^k = (-1)^k \binom{-r}{k} p q^k.$$

This explains the negative Binomial distribution terminology.

Let us further note that when $r = -1$,

$$\binom{-r}{k} = (-1)^k \frac{(-1)(-1-1)(-1-2)\cdots(-1-k+1)}{(k)!} = 1$$

and so

$$\mathbb{P}(W_1 = k) = p(1-p)^k = pq^k = \mathbb{P}(\text{Geo}(p) = k+1) = \mathbb{P}(\text{Geo}(p) - 1 = k)$$

so that $W_1 \overset{d}{=} \text{Geo}(p) - 1$.

Lemma 2.9 (Generating Function). *If $0 < p \leq 1$, $q = 1-p$, and W_r is as in Definition 2.7, then its generating function is given by*

$$G_{W_r}(z) = p^r (1-qz)^{-r} = \left(\frac{p}{1-qz}\right)^r \text{ for } |z| < 1/q \tag{2.6}$$

and moreover,

$$\mathbb{E}W_r = r \cdot \frac{q}{p} \text{ and } \text{Var}(W_r) = r\frac{q}{p^2}. \tag{2.7}$$

Proof. By definition of the generating function,

$$G(z) := G_{W_r}(z) = \mathbb{E}\left[z^{W_r}\right] = \sum_{k=0}^{\infty} z^k \mathbb{P}(W_r = k) = p^r \sum_{k=0}^{\infty} (-1)^k \binom{-r}{k} (qz)^k$$
$$= p^r (1-qz)^{-r} = \left(\frac{p}{1-qz}\right)^r$$

wherein we have used the Binomial Theorem 2.17 to evaluate the sum. We then compute the derivatives,

$$G'(z) = qrp^r (1-qz)^{-r-1} \text{ and } G''(z) = q^2 r(r+1) p^r (1-qz)^{-r-2}$$

and so

$$G'(1) = qrp^r (1-q)^{-r-1} = qrp^r p^{-r-1} = r\frac{q}{p} \text{ and}$$
$$G''(1) = q^2 r(r+1) p^r (1-q)^{-r-2} = r(r+1)\frac{q^2}{p^2}.$$

Therefore by Eqs. (2.2) and (2.3),

$$\mathbb{E}W_r = qrp^r (1-q)^{-r-1} = r\frac{q}{p} \text{ and}$$
$$\text{Var}(W_r) = r(r+1)\frac{q^2}{p^2} + r\frac{q}{p} - \left(r\frac{q}{p}\right)^2 = r\left[\frac{q^2}{p^2} + \frac{q}{p}\right]$$
$$= r\frac{q}{p^2}[q+p] = r\frac{q}{p^2}.$$

■

Corollary 2.10. *If $\{X_i\}_{i=1}^r$ are i.i.d. with $X_i \overset{d}{=} W_1 \overset{d}{=} [\text{Geo}(p) - 1]$, then*[1]

$$W_r \overset{d}{=} X_1 + \cdots + X_r. \tag{2.8}$$

[This can be used to give another proof of the identities in Eq. (2.7).]

Proof. From Lemma 2.9 we know that

$$G_{X_i}(z) = G_{W_1}(z) = \frac{p}{1-qz}$$

and therefore, using the assumed independence,

$$G_{X_1+\cdots+X_r}(z) = \mathbb{E}\left[z^{X_1+\cdots+X_r}\right] = \mathbb{E}\left[z^{X_1} z^{X_2} \cdots z^{X_r}\right]$$
$$= \prod_{j=1}^{r} \mathbb{E}\left[z^{X_j}\right] = \left(\frac{p}{1-qz}\right)^r = G_{W_r}(z).$$

Since the generating function completely determines the probability mass function, it follows that $W_r \overset{d}{=} X_1 + \cdots + X_r$.

Using Eq. (2.8) this fact along with Remark 2.8 and Exercise 2.1, it then follows that

$$\mathbb{E}W_r = \mathbb{E}W_1 = r(\mathbb{E}[\text{Geo}(p) - 1]) = r\left(\frac{1}{p} - 1\right) = r \cdot \frac{1-p}{p}$$

and

$$\text{Var}(W_r) = r\,\text{Var}(X_1) = r\,\text{Var}(\text{Geo}(p)) = r\frac{1-p}{p^2}.$$

■

[1] This result is fairly intuitive since $X_1 + \cdots + X_r$ represents a string $\{Z_j\}_{j=1}^{X_1+\cdots+X_r}$ where

$$\#\{1 \leq j \leq X_1 + \cdots + X_r : Z_j = 1 \text{ with the last being } 1\}.$$

Exercise 2.2. Let $\{\lambda_n\}_{n=1}^{\infty}$ be a sequence of real numbers such that $\lambda = \lim_{n\to\infty}\lambda_n$ exists in $\mathbb{R}$. Show for each $k \in \mathbb{N}_0$ that

$$\lim_{n\to\infty}(1-\lambda_n/n)^{n-k} = e^{-\lambda}.$$

Hint: you might use $\ln(1+x) = x + O(x^2)$ for $|x|$ small.]

Proposition 2.11 (The Law of Rare Events I). *Let $S_{n,p} \overset{d}{=} \text{Bin}(n,p)$, $k \in \mathbb{N}$, $p_n = \lambda_n/n$ where $\lambda_n \to \lambda > 0$ as $n \to \infty$. Show $\text{Bin}(n,\lambda_n/n) \implies \text{Poi}(\lambda)$ as $n \to \infty$,[2] i.e. show*

$$\lim_{n\to\infty}\mathbb{P}(S_{n,p_n}=k) = \frac{\lambda^k}{k!}e^{-\lambda} = \mathbb{P}(\text{Poi}(\lambda)=k) \text{ for } k\in\mathbb{N}_0. \tag{2.9}$$

(We will come back to the Poisson distribution and the related Poisson process later on.)

Proof. We have,

$$\begin{aligned}\mathbb{P}(S_{n,p_n}=k) &= \binom{n}{k}(\lambda_n/n)^k(1-\lambda_n/n)^{n-k}\\ &= \frac{\lambda_n^k}{k!}\frac{n(n-1)\dots(n-k+1)}{n^k}(1-\lambda_n/n)^{n-k}.\end{aligned}$$

The result now follows from Exercise 2.2 and the observation that (for each fixed $k \in \mathbb{N}_0$),

$$\lim_{n\to\infty}\frac{n(n-1)\dots(n-k+1)}{n^k} = 1.$$

■

Remark 2.12. Slightly informally, Eq. (2.9) states; if $n \in \mathbb{N}$ is large and $p = O(1/n)$[3], then

$$\mathbb{P}(\text{Bin}(n,p)=k) \cong \mathbb{P}(\text{Poi}(pn)=k) = \frac{(pn)^k}{k!}e^{-pn} \text{ for } k \lll n. \tag{2.10}$$

See the next two figures where $(p,n) = (5/100, 100)$ and $(p,n) = (5/1000, 1000)$ so that $\lambda = pn = 5$ in each case.

[2] The probability of success, p_n, is going to zero as $n \to \infty$. Thus as $n \to \infty$ we are doing lots of trials of an experiment with very low probability of success and hence the name, the Law of rare events.

[3] Writing $p = O(1/n)$ is being used informally here to mean the pn is "much smaller" than n.

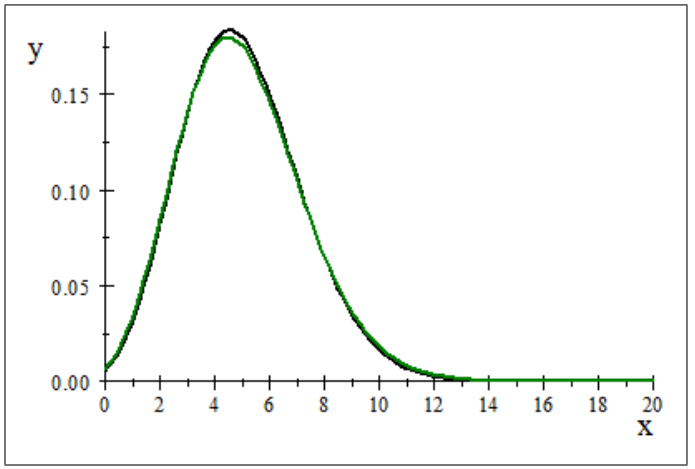


Fig. 2.1. Plot of the probability functions for $Bern\left(\frac{5}{100},100\right)$ in black and $Poi(5)$ in green.

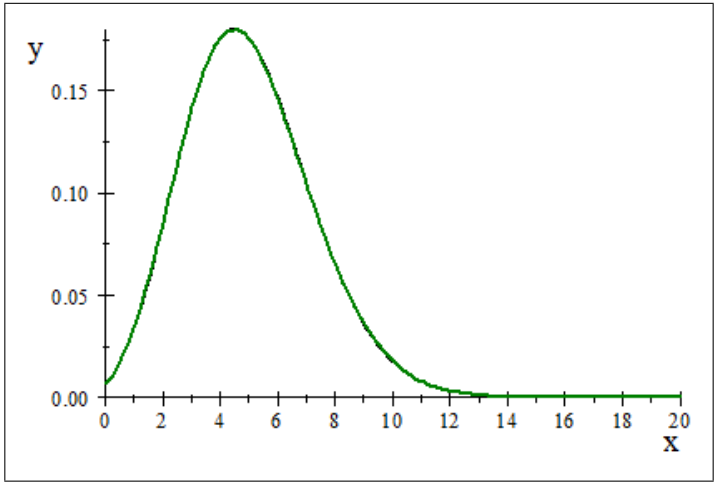


Fig. 2.2. Plot of the probability functions for $Bern\left(\frac{5}{1000},1000\right)$ in black and $Poi(5)$ in green.

2.2 Appendix: Taylor's Theorem

Proposition 2.13 (Taylor's Theorem I). *Suppose that $h:(-\varepsilon,1+\varepsilon)\to\mathbb{C}$ is a C^N – differentiable function. Then*

$$h(1)=\sum_{l=0}^{N-1}\frac{h^{(l)}(0)}{l!}+\int_0^1\frac{(1-t)^{N-1}}{(N-1)!}h^{(N)}(t)dt. \tag{2.11}$$

Proof. We prove this formula by induction on N using integration by parts. For $N=1$ the formula states,

$$h(1)=h(0)+\int_0^1 h'(t)\,dt$$

which is true by the fundamental theorem of calculus. Moreover if f is C^{N+1} – differentiable, then by integration by parts

$$\begin{aligned}\int_0^1\frac{(1-t)^{(N-1)}}{(N-1)!}h^{(N)}(t)dt&=-\int_0^1 h^{(N)}(t)\left[\frac{d}{dt}\frac{(1-t)^N}{N!}\right]dt\\&=\int_0^1 h^{(N+1)}(t)\frac{(1-t)^N}{N!}dt-h^{(N)}(t)\frac{(1-t)^N}{N!}\Big|_0^1\\&=\int_0^1 h^{(N+1)}(t)\frac{(1-t)^N}{N!}dt+\frac{1}{N!}h^{(N)}(0).\end{aligned}$$

Hence if Eq. (2.11) holds then

$$\begin{aligned}h(1)&=\sum_{l=0}^{N-1}\frac{h^{(l)}(0)}{l!}+\frac{1}{N!}h^{(N)}(0)+\int_0^1 h^{(N+1)}(t)\frac{(1-t)^N}{N!}dt\\&=\sum_{l=0}^{N}\frac{h^{(l)}(0)}{l!}+\int_0^1 h^{(N+1)}(t)\frac{(1-t)^N}{N!}dt\end{aligned}$$

which completes the induction argument. ∎

Theorem 2.14 (Taylor's Theorem with Integral Remainder). *Suppose that $f:(a,b)\to\mathbb{C}$ is C^N – differentiable and $x_0\in(a,b)$. Then for all $x\in(a,b)$,*

$$f(x)=\sum_{l=0}^{N-1}\frac{f^{(k)}(x_0)}{l!}(x-x_0)^k+\frac{(x-x_0)^N}{N!}R_N(x)$$

where $R_N(x)=R_N(f,x_0;x)$ is given by

$$\begin{aligned}R_N(x)&=\int_0^1 f^{(N)}(x_0+t(x-x_0))\,N(1-t)^{N-1}dt\\&=\int_0^1 f^{(N)}((1-t)x_0+tx)\,N(1-t)^{N-1}dt.\end{aligned}$$

[Observe that $\int_0^1 N(1-t)^{N-1}dt=-(1-t)^N|_0^1=1$.]

Proof. We apply Proposition 2.13 with $h(t):=f(x_0+t(x-x_0))$ using, $h^{(k)}(t)=f^{(k)}(x_0+t(x-x_0))(x-x_0)^k$. Therefore,

$$\begin{aligned}f(x)=&h(1)=\sum_{l=0}^{N-1}\frac{h^{(l)}(0)}{l!}+\int_0^1\frac{(1-t)^{N-1}}{(N-1)!}h^{(N)}(t)dt\\&=\sum_{l=0}^{N-1}\frac{f^{(k)}(x_0)}{l!}(x-x_0)^k\\&\qquad+\frac{(x-x_0)^N}{N!}\int_0^1 f^{(N)}(x_0+t(x-x_0))\,N(1-t)^{N-1}dt.\end{aligned}$$

∎

Definition 2.15. *For $\beta\in\mathbb{R}$ and $k\in\mathbb{N}_0$ we define the Binomial coefficient by*

$$\binom{\beta}{k}:=\frac{\beta(\beta-1)\dots(\beta-k+1)}{k!}$$

with the convention that $\binom{\beta}{0}=1$ for all $\beta\in\mathbb{R}$.

Remark 2.16. If $\beta\in\mathbb{N}$ then

$$\binom{\beta}{k}=\frac{1}{k!}\beta(\beta-1)\dots(\beta-k+1)=\frac{1}{k!}\frac{\beta!}{(\beta-k)!}=\frac{\beta!}{k!\cdot(\beta-k)!}.$$

Theorem 2.17 (Binomial Series). *If $\beta\in\mathbb{R}$ and $|x|<1$, then*

$$(1-x)^\beta=\sum_{k=0}^{\infty}(-1)^k\binom{\beta}{k}x^k.$$

Proof. By repeated differentiation you may show,

$$f^{(k)}(x)=(-1)^k\beta(\beta-1)\dots(\beta-k+1)(1-x)^{\beta-k}=(-1)k!\cdot\binom{\beta}{k}(1-x)^{\beta-k}$$

So by Taylor's theorem (Eq. (2.11) with $x=0$ and $y=x$)

$$(1-x)^\beta = 1 + \sum_{k=1}^{N-1} \frac{1}{k!}(-1)^k \beta(\beta-1)\dots(\beta-k+1)x^k + R_N(x) \tag{2.12}$$

where

$$\begin{aligned} R_N(x) &= \frac{x^N}{N!}\int_0^1 (-1)^N \beta(\beta-1)\dots(\beta-N+1)(1-sx)^{\beta-N} d\nu_N(s) \\ &= \frac{x^N}{N!}(-1)^N \beta(\beta-1)\dots(\beta-N+1)\int_0^1 \frac{N(1-s)^{N-1}}{(1-sx)^{N-\beta}} ds. \end{aligned}$$

Now for $x \in (-1,1)$ and $N > \beta$,

$$0 \le \int_0^1 \frac{N(1-s)^{N-1}}{(1-sx)^{N-\beta}} ds \le \int_0^1 \frac{N(1-s)^{N-1}}{(1-s)^{N-\beta}} ds = \int_0^1 N(1-s)^{\beta-1} ds = \frac{N}{\beta}$$

and therefore,

$$|R_N(x)| \le \frac{|x|^N}{(N-1)!} |(\beta-1)\dots(\beta-N+1)| =: \rho_N.$$

Since

$$\limsup_{N\to\infty} \frac{\rho_{N+1}}{\rho_N} = |x| \cdot \limsup_{N\to\infty} \frac{N-\beta}{N} = |x| < 1,$$

the Ratio test implies that $|R_N(x)| \le \rho_N \to 0$ (exponentially fast) as $N \to \infty$. Therefore by passing to the limit in Eq. (2.12) we have proved

$$(1-x)^\beta = 1 + \sum_{k=1}^{\infty} \frac{(-1)^k}{k!} \beta(\beta-1)\dots(\beta-k+1)x^k \tag{2.13}$$

which is valid for $|x| < 1$ and $\beta \in \mathbb{R}$.

■

Example 2.18. An important special cases is $\beta = -1$ in which case, Eq. (2.13) becomes the standard geometric series formula;

$$\frac{1}{1-x} = \sum_{k=0}^{\infty} x^k.$$

Another another useful special case is $\beta = 1/2$ in which case Eq. (2.13) becomes

$$\begin{aligned} \sqrt{1-x} &= 1 + \sum_{k=1}^{\infty} \frac{(-1)^k}{k!} \frac{1}{2}(\frac{1}{2}-1)\dots(\frac{1}{2}-k+1)x^k \\ &= 1 - \sum_{k=1}^{\infty} \frac{(2k-3)!!}{2^k k!} x^k \text{ for all } |x| < 1. \end{aligned} \tag{2.14}$$

3

Conditional Expectation (Discrete Case)

3.1 Conditional and Joint Distributions

Let us now suppose that S and T are finite or at most countable sets and $X:\Omega\to S$ and $Y:\Omega\to T$ are (random) functions.

Definition 3.1 ((conditional) probability mass functions). *Let $p_X:S\to[0,1]$, $p_Y:T\to[0,1]$, $p_{X,Y}:S\times T\to[0,1]$, and $p_{X|Y}:S\times T\to[0,1]$, be the functions defined by*

$$\begin{aligned}
p_X(x) &= \mathbb{P}[X=x] \text{ for all } x\in S\\
p_Y(y) &= \mathbb{P}[Y=y] \text{ for all } y\in T\\
p_{X,Y}(x,y) &= \mathbb{P}[X=x \text{ and } Y=y], \text{ for } (x,y)\in S\times T, \text{ and}\\
p_{X|Y}(x|y) &= \mathbb{P}[X=x|Y=y] = \frac{\mathbb{P}[X=x \text{ and } Y=y]}{\mathbb{P}[Y=y]}\\
&= \frac{p_{X,Y}(x,y)}{p_Y(y)} \text{ when } p_Y(y)>0.
\end{aligned}$$

*If $p_Y(y)=0$ we define $p_{X,Y}(x|y)\in[0,1]$ as function of x in any way we like such that $\sum_{y\in T}p_{X|Y}(x|y)=1$. We call $p_{X|Y}(x|y)$ the **conditional probability mass function** of X given Y, $p_{X,Y}$ is the **joint distribution** of (X,Y) and p_X and p_Y are the **probability mass functions** for X and Y respectively.*

Let us note the following identities;

$$p_{X,Y}(x,y)=p_{X|Y}(x|y)\,p_Y(y),$$

$$\begin{aligned}
p_X(x) &= \sum_{y\in T}p_{X,Y}(x,y)=\sum_{y\in T}p_{X|Y}(x|y)\,p_Y(y), \text{ and}\\
p_Y(y) &= \sum_{x\in S}p_{X,Y}(x,y)=\sum_{x\in S}p_{Y|X}(y|x)\,p_X(x).
\end{aligned}$$

In this setting one refers to p_X and p_Y as the X-**marginal** and Y-**marginal** of $p_{(X,Y)}$ respectively.

Example 3.2. Let $Y\in[6]=\{1,2,3,4,5,6\}$ be the result of a fair die toss and then let X be the number of heads resulting from tossing a fair coin Y-times. Find $p_{(X,Y)}$ and p_X, i.e. find the joint law of (X,Y), its X-marginal, and then compute $\mathbb{E}X$.

Solution. We are given $p_Y(y)=\frac{1}{6}$ and

$$\begin{aligned}
p_{X|Y}(x|y) &= \mathbb{P}(X=x|Y=y)=\binom{y}{x}\left(\frac{1}{2}\right)^x\left(\frac{1}{2}\right)^{y-x}\\
&= \binom{y}{x}\left(\frac{1}{2}\right)^y \text{ for } 0\le x\le y.
\end{aligned}$$

Thus we find, for $0\le x\le y\le 6$ and $y\ge 1$, that

$$\begin{aligned}
p_{(X,Y)}(x,y) &= p_{X|Y}(x|y)\,p_Y(y)=\frac{1}{6}\binom{y}{x}\left(\frac{1}{2}\right)^y\\
&= \frac{1}{6\cdot 2^6}\binom{y}{x}2^{6-y}.
\end{aligned}$$

Here is the table of values of $p_{(X,Y)}(x,y)$;

$$p_{(X,Y)}(x,y)=\frac{1}{6\cdot 2^6}\left[\begin{array}{c|cccccc}
x\backslash y & 1 & 2 & 3 & 4 & 5 & 6\\ \hline
0 & 32 & 16 & 8 & 4 & 2 & 1\\
1 & 32 & 32 & 24 & 16 & 10 & 6\\
2 & 0 & 16 & 24 & 24 & 20 & 15\\
3 & 0 & 0 & 8 & 16 & 20 & 20\\
4 & 0 & 0 & 0 & 4 & 10 & 15\\
5 & 0 & 0 & 0 & 0 & 2 & 6\\
6 & 0 & 0 & 0 & 0 & 0 & 1
\end{array}\right] \tag{3.1}$$

where for example the, $x=3$ and $y=4$ entry is given by

$$\binom{y}{x}2^{6-y}|_{x=3,y=4}=\binom{4}{3}2^{6-4}=4\cdot 2^2=16.$$

To compute the X-marginal, we need to add the rows of the matrix in Eq. (3.1) to find

x	$6\cdot 2^6\cdot p_X(x)$
0	63
1	120
2	99
3	64
4	29
5	8
6	1

.

For example the top entry is found using, $32+16+8+4+2+1=63$. As a check let us note that

$$63+120+99+64+29+8+1=384=6\cdot 2^6.$$

We may now compute $\mathbb{E}X$ as

$$\begin{aligned}\mathbb{E}X&=\frac{1}{6\cdot 2^6}\left(0\cdot 63+1\cdot 120+2\cdot 99+3\cdot 64+4\cdot 29+5\cdot 8+6\cdot 1\right)\\&=\frac{1}{6\cdot 2^6}672=\frac{7}{4}.\end{aligned}$$

The next two examples are also discussed on pages 47-49 of P.K.

Example 3.3 ($\text{Bin}(p,\text{Bin}(q,M))\overset{d}{=}\text{Bin}(pq,M)$*).* If $X\overset{d}{=}\text{Bin}(p,N)$ where $N\overset{d}{=}\text{Bin}(q,M)$, then $X\overset{d}{=}\text{Bin}(pq,M)$.

Solution. By assumption,

$$\mathbb{P}[N=n]=\binom{M}{n}q^n(1-q)^{N-n}$$

$$\mathbb{P}[X=k|N=n]=\binom{n}{k}p^k(1-p)^{n-k}\text{ and}$$

and therefore the joint distribution of (X,N) is given by;

$$\begin{aligned}\mathbb{P}[X=k,N=n]&=\binom{n}{k}p^k(1-p)^{n-k}\binom{M}{n}q^n(1-q)^{M-n}1_{0\le k\le n\le M}\\&=\frac{M!}{k!(n-k)!(M-n)!}p^k(1-p)^{n-k}q^n(1-q)^{M-n}1_{0\le k\le n\le M}.\end{aligned}$$

Summing this identity on n (while making the change of variables, let $n=k+\ell$, in the second line) shows

$$\begin{aligned}\mathbb{P}[X=k]&=\sum_{n=0}^{M}\mathbb{P}[X=k,N=n]\\&=\sum_{n=0}^{M}\frac{M!}{k!(n-k)!(M-n)!}p^k(1-p)^{n-k}q^n(1-q)^{M-n}1_{0\le k\le n\le M}\\&=\sum_{\ell=0}^{M-k}\frac{M!}{k!(\ell)!(M-k-\ell)!}p^k(1-p)^{\ell}q^{k+\ell}(1-q)^{M-k-\ell}\\&=\frac{M!}{k!(M-k-\ell)!}p^kq^k\sum_{\ell=0}^{M-k}\frac{(M-k)!}{(\ell)!(M-k-\ell)!}(1-p)^{\ell}q^{\ell}(1-q)^{M-k-\ell}\\&=\frac{M!}{k!(M-k-\ell)!}p^kq^k\left((1-p)q+(1-q)\right)^{M-k}\\&=\frac{M!}{k!(M-k-\ell)!}(pq)^k(1-pq)^{M-k},\end{aligned}$$

i.e. $X\overset{d}{=}\text{Bin}(pq,M)$.

Remark 3.4. Let $q=\lambda/M$ and then making use of Proposition 2.11 suggests that

$$\begin{array}{ccc}\text{Bin}(p,\text{Bin}(\lambda/M,M)) & \overset{d}{=} & \text{Bin}(p\lambda/M,M)\\ \Downarrow & \text{as } M\to\infty & \Downarrow\\ \text{Bin}(p,\text{Poi}(\lambda)) & \overset{?}{=} & \text{Poi}(p\lambda)\end{array}$$

which suggest that $\text{Bin}(p,\text{Poi}(\lambda))\overset{d}{=}\text{Poi}(p\lambda)$. We will verify this conclusion directly in the next proposition.

Proposition 3.5. *Suppose that $\lambda>0$, $p\in(0,1)$, $q=1-p$, $N\overset{d}{=}\text{Poi}(\lambda)$ and given $N=n$, suppose that $X\overset{d}{=}\text{Bin}(n,p)$. Then $X\overset{d}{=}\text{Poi}(p\lambda)$ and $Y=N-X\overset{d}{=}\text{Poi}(q\lambda)$ and both X and Y are independent of one another.*

Proof. By assumption,

$$\mathbb{P}[N=n]=e^{-\lambda}\frac{\lambda^n}{n!}\text{ and}$$

$$\mathbb{P}[X=k|N=n]=\binom{n}{k}p^k(1-p)^{n-k}\text{ for }0\le k\le n.$$

Therefore the joint distribution is determined by

$$\mathbb{P}[X=k,N=n]=\binom{n}{k}p^k(1-p)^{n-k}e^{-\lambda}\frac{\lambda^n}{n!}$$
$$=\frac{(p\lambda)^k}{k!}\frac{(q\lambda)^{n-k}}{(n-k)!}e^{-\lambda}\text{ for }0\le k\le n<\infty.$$

Summing this equation on $n\in\mathbb{N}_0$ shows

$$\mathbb{P}[X=k]=\sum_{n=0}^{\infty}\mathbb{P}[X=k,N=n]=\sum_{n=k}^{\infty}\frac{(p\lambda)^k}{k!}\frac{(q\lambda)^{n-k}}{(n-k)!}e^{-\lambda}$$
$$=\frac{(p\lambda)^k}{k!}e^{-\lambda}\sum_{n=k}^{\infty}\frac{(q\lambda)^{n-k}}{(n-k)!}\quad(\text{let }\ell=n-k)$$
$$=\frac{(p\lambda)^k}{k!}e^{-\lambda}\sum_{\ell=0}^{\infty}\frac{(q\lambda)^{\ell}}{(\ell)!}$$
$$=\frac{(p\lambda)^k}{k!}e^{-\lambda}e^{q\lambda}=\frac{(p\lambda)^k}{k!}e^{-\lambda}e^{(1-p)\lambda}=e^{-p\lambda}\frac{(p\lambda)^k}{k!}$$

from which it follows that $X\overset{d}{=}\text{Poi}(p\lambda)$.

We then see that $(Y=N-X)$

$$\mathbb{P}[X=k,Y=\ell]=\mathbb{P}[X=k,N-X=\ell]$$
$$=\mathbb{P}[X=k,N=k+\ell]$$
$$=\frac{(p\lambda)^k}{k!}\frac{(q\lambda)^{k+\ell-k}}{(k+\ell-k)!}e^{-\lambda}=\frac{(p\lambda)^k}{k!}\frac{(q\lambda)^{\ell}}{\ell!}e^{-\lambda}$$
$$=\frac{(p\lambda)^k}{k!}e^{-p\lambda}\frac{(q\lambda)^{\ell}}{\ell!}e^{-q\lambda}$$

which shows X is independent of Y and $Y\overset{d}{=}\text{Poi}(q\lambda)$.

Note that

$$\mathbb{P}[X=k]=\sum_{\ell=0}^{\infty}\mathbb{P}[X=k,Y=\ell]=\frac{(p\lambda)^k}{k!}e^{-p\lambda}\sum_{\ell=0}^{\infty}\frac{(q\lambda)^{\ell}}{\ell!}e^{-q\lambda}$$
$$=\frac{(p\lambda)^k}{k!}e^{-p\lambda}e^{q\lambda}e^{-q\lambda}=\frac{(p\lambda)^k}{k!}e^{-p\lambda}.$$

■

3.2 Conditional Expectations

Definition 3.6. *We let $L^1(\mathbb{P})$ denote those random variables, $X:\Omega\to\mathbb{R}$, such that $\mathbb{E}|X|<\infty$. We say such a random variable is **integrable.***

Again suppose that T is a finite or at most countable set and let $Y:\Omega\to T$. For $X\in L^1(\mathbb{P})$ and $y\in T$, let

$$\mathbb{E}[X|Y=y]=\begin{cases}\frac{\mathbb{E}[X:Y=y]}{\mathbb{P}[Y=y]}=\frac{\mathbb{E}[X\cdot 1_{Y=y}]}{\mathbb{P}[Y=y]} & \text{if }\mathbb{P}[Y=y]\neq 0\\ \mathbb{E}X & \text{if }\mathbb{P}[Y=y]=0\end{cases}.$$

When $\mathbb{P}[Y=y]=0$, $\mathbb{E}[X|Y=y]$ is rather arbitrarily defined above. As there is no chance that the event $\{Y=y\}$ occurs the value we choose for $\mathbb{E}[X|Y=y]$ in this case is actually irrelevant. If $S:=X(\Omega)$ is a finite or countable set, then we may take

$$\mathbb{E}[X|Y=y]=\sum_{x\in X}f(x)\,p_{X|Y}(x|y).$$

Definition 3.7 (Conditional Expectation). *The **conditional expectation** ($\mathbb{E}[X|Y]$) **of** X given Y is the `random variable`,*

$$\mathbb{E}[X|Y]:=g(Y)\ \textit{ where }g(y):=\mathbb{E}[X|Y=y]\ \textit{ for }y\in T.$$

The next theorem summarizes the main properties of conditional expectations.

Theorem 3.8 (Basic properties). *Let X, $X_1,X_2\in L^1(\mathbb{P})$, $Y:\Omega\to T$, and $a\in\mathbb{R}$ be given. Then:*

1. *$\mathbb{E}[k|Y]=k$ when k is a constant.*
2. ***Linearity:***
$$\mathbb{E}(X_1+aX_2|Y)=\mathbb{E}(X_1|Y)+a\mathbb{E}(X_2|Y).$$
3. ***Pullout Property:*** *for all bounded functions g,*
$$\mathbb{E}(g(Y)X|Y)=g(Y)\mathbb{E}(X|Y)$$
4. ***Tower Property/ Law of Total Probability / Law of the Forgetful Statistician (LFS):***
$$\mathbb{E}(\mathbb{E}(X|Y))=\mathbb{E}X.$$
5. ***Independence property:*** *If X and Y are independent then*
$$\mathbb{E}(X|Y)=\mathbb{E}X\quad a.s.$$
6. ***Best RMS[1] Approximation:*** *if $\mathbb{E}|X|^2<\infty$ and $\bar{X}:=\mathbb{E}[X|Y]=g(Y)$ and $h:T\to\mathbb{R}$ is such that $\mathbb{E}|h(Y)|^2<\infty$, then*
$$\mathbb{E}(X-h(Y))^2\ge\mathbb{E}\left(X-\bar{X}\right)^2=\mathbb{E}(X-g(Y))^2.\tag{3.2}$$
This shows that $\bar{X}=g(Y)$ is the best approximation to X among all functions of the form $h(Y)$ with $\mathbb{E}|h(Y)|^2<\infty$, see Figure 3.1.

[1] RMS stands for root-mean-square.

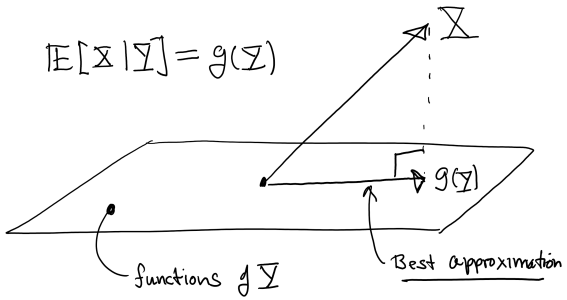


Fig. 3.1. Geometric description of $\mathbb{E}[X|Y]$.

Proof. We will take each item in turn. We will use over and over again that

$$\Omega = \sum_{y \in T} \{Y = y\} = \cup_{y \in T} \{Y = y\}.$$

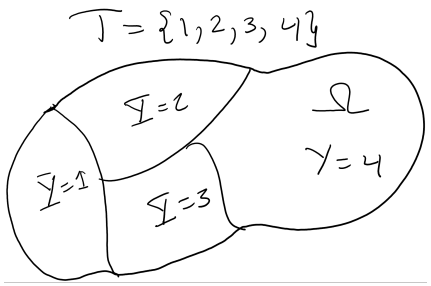


Fig. 3.2. Showing how Y partitions Ω.

$$\mathbb{E}[X|Y] = \mathbb{E}[X|Y = j] \text{ on the set } \{Y = j\}.$$

1. & 2. On the event, $\{Y = y\}$ we have

$$\mathbb{E}(k|Y) = \mathbb{E}(k|Y = y) = k$$

$$\mathbb{E}(X_1 + aX_2|Y) = \mathbb{E}(X_1 + aX_2|Y = y) = \mathbb{E}(X_1|Y = y) + a\mathbb{E}(X_2|Y = y).$$

As the events $\{Y = y\}$ for $y \in T$ partitions Ω, these identities suffice to prove the first two items.

3. Similarly, on the event, $\{Y = y\}$

$$\begin{aligned}\mathbb{E}(g(Y)X|Y) &= \mathbb{E}(g(Y)X|Y = y) = \mathbb{E}(g(y)X|Y = y) \\ &= g(y)\mathbb{E}(X|Y = y) = g(Y)\mathbb{E}(X|Y).\end{aligned}$$

4. The proof of the **LFS** is contained in the following simple computation;

$$\begin{aligned}\mathbb{E}(\mathbb{E}(X|Y)) &= \sum_{y \in T} \mathbb{E}(\mathbb{E}(X|Y) : Y = y) = \sum_{y \in T} \mathbb{E}(\mathbb{E}(X|Y = y) : Y = y) \\ &= \sum_{y \in T} \mathbb{E}(\mathbb{E}(X|Y = y) \cdot 1_{Y=y}) \\ &= \sum_{y \in T} \mathbb{E}(X|Y = y) \cdot \mathbb{P}(Y = y) = \sum_{y \in T} \mathbb{E}(X : Y = y) \\ &= \mathbb{E}[X].\end{aligned}$$

5. If X and Y are independent and $y \in T$, then

$$\mathbb{E}(X|Y) = \frac{\mathbb{E}[X\mathbf{1}_{Y=y}]}{\mathbb{P}[Y = y]} = \frac{\mathbb{E}[X] \cdot \mathbb{E}[\mathbf{1}_{Y=y}]}{\mathbb{P}[Y = y]} = \mathbb{E}X.$$

6. Taking expectations of the identity;

$$\begin{aligned}(X - h(Y))^2 &= (X - g(Y) + g(Y) - h(Y))^2 \\ &= (X - g(Y))^2 + (g(Y) - h(Y))^2 + 2(X - g(Y))(g(Y) - h(Y)),\end{aligned}$$

shows,

$$\mathbb{E}(X - h(Y))^2 = \mathbb{E}\left(X - \bar{X}\right)^2 + \mathbb{E}\left(\bar{X} - h(Y)\right)^2 + 2\mathbb{E}[(X - g(Y))(g(Y) - h(Y))]. \tag{3.3}$$

However, by the pull-out property,

$$\mathbb{E}[(X - g(Y))|Y] = \mathbb{E}[X|Y] - g(Y)\mathbb{E}[1|Y] = 0 \text{ and so}$$
$$\mathbb{E}[(X - g(Y))(g(Y) - h(Y))|Y] = (g(Y) - h(Y))\mathbb{E}[(X - g(Y))|Y] = 0$$

and so by the tower property,

$$\begin{aligned}\mathbb{E}[(X - g(Y))(g(Y) - h(Y))] &= \mathbb{E}(\mathbb{E}[(X - g(Y))(g(Y) - h(Y))|Y]) \\ &= \mathbb{E}(0) = 0.\end{aligned} \tag{3.4}$$

So from Eqs. (3.3) and (3.4);

$$\mathbb{E}(X - h(Y))^2 = \mathbb{E}\left(X - \bar{X}\right)^2 + \mathbb{E}\left(\bar{X} - h(Y)\right)^2 \geq \mathbb{E}\left(X - \bar{X}\right)^2$$

and Equation (3.2) is proved.

■

Example 3.9 (Example 3.2 Cont.). Recall that in Example 3.2 that let $Y \in [6] = \{1, 2, 3, 4, 5, 6\}$ be the result of a fair die toss and then let X be the number of heads resulting from tossing a fair coin Y-times. Since

$$\mathbb{E}[X|Y=y]=\frac{1}{2}y \implies \mathbb{E}[X|Y]=\frac{1}{2}Y$$

it follows by the LFS that

$$\mathbb{E}[X]=\mathbb{E}(\mathbb{E}[X|Y])=\frac{1}{2}\mathbb{E}Y=\frac{1}{2}\cdot\frac{1}{6}(1+\cdots+6)=\frac{21}{12}=\frac{7}{4}.$$

This is in agreement with what we already found in Example 3.2.

Exercise 3.1 (See Durrett, #8, p. 213). Suppose that X and Y are two integrable random variables such that

$$\mathbb{E}[X|Y]=18-\frac{3}{5}Y \text{ and } \mathbb{E}[Y|X]=10-\frac{1}{3}X.$$

Find $\mathbb{E}X$ and $\mathbb{E}Y$.

3.3 Random Length Random Sums

See P.K. Section 2.3. Let $\{\xi_i\}_{i=1}^{\infty}$ be i.i.d random variables and for $n\in\mathbb{N}_0$ let

$$X_n=\xi_1+\cdots+\xi_n:=\begin{cases}\sum_{i=1}^n\xi_i & \text{if } n\geq 1\\ 0 & \text{if } n=0\end{cases} \text{ for } n\in\mathbb{N}_0\,.$$

If $N:\Omega\to\mathbb{N}_0$ is a discrete random variable we also let

$$X=X_N=\xi_1+\cdots+\xi_N:=\begin{cases}\sum_{i=1}^N\xi_i & \text{if } N\geq 1\\ 0 & \text{if } N=0\end{cases}$$

which is now a random sum of random length, N. We further let

$$\begin{aligned}\mu&=\mathbb{E}\xi_i, \quad \sigma^2=\operatorname{Var}(\xi_i),\\ \nu&=\mathbb{E}N=\nu, \quad \tau^2=\operatorname{Var}(N)\end{aligned}$$

which we assume to be finite.

Proposition 3.10 (Random Length Random Sums). *If N is independent of $\{\xi_i\}_{i=1}^{\infty}$, then*

$$\mathbb{E}X=\mu\nu \text{ and } \operatorname{Var}(X)=\nu\sigma^2+\mu^2\tau^2.$$

Proof. For a general function, $f(x)$, with $x\in\mathbb{R}$ we have

$$\begin{aligned}\mathbb{E}f(X)&=\sum_{n=0}^{\infty}\mathbb{E}[f(X):N=n]=\sum_{n=0}^{\infty}\mathbb{E}[f(X_n):N=n]\\ &=\sum_{n=0}^{\infty}\mathbb{E}[f(X_n)]\cdot\mathbb{P}[N=n]\,.\end{aligned}$$

Since $\mathbb{E}X_n=n\cdot\mu$, taking $f(x)=x$ we find,

$$\mathbb{E}X=\sum_{n=0}^{\infty}\mathbb{E}X_n\cdot\mathbb{P}[N=n]=\sum_{n=0}^{\infty}\mu\cdot n\cdot\mathbb{P}[N=n]=\mu\mathbb{E}N=\mu\nu.$$

Similarly taking $f(x)=x^2$ while using,

$$\mathbb{E}X_n^2=\operatorname{Var}(X_n)+(\mathbb{E}X_n)^2=n\cdot\sigma^2+n^2\mu^2,$$

it follows that

$$\begin{aligned}\mathbb{E}X^2&=\sum_{n=0}^{\infty}\mathbb{E}\left[X_n^2\right]\cdot\mathbb{P}[N=n]\\ &=\sum_{n=1}^{\infty}\left(n\cdot\sigma^2+n^2\mu^2\right)\cdot\mathbb{P}[N=n]\\ &=\sigma^2\mathbb{E}N+\mu^2\mathbb{E}N^2=\sigma^2\nu+\mu^2\left(\tau^2+\nu^2\right)\end{aligned}$$

and so

$$\operatorname{Var}(X)=\mathbb{E}X^2-(\mathbb{E}X)^2=\mathbb{E}X^2-(\mu\nu)^2=\nu\sigma^2+\mu^2\tau^2.$$

■

Example 3.11 (Examples 3.2 and 3.9 revisited). Recall that in Example 3.2 that let $N:=Y\in[6]=\{1,2,3,4,5,6\}$ be the result of a fair die toss and then let X be the number of heads resulting from tossing a fair coin Y-times. Thus if we let $\{\xi_j\}_{j=1}^{\infty}$ be i.i.d. with $\mathbb{P}[\xi_j=1]=\frac{1}{2}=\mathbb{P}[\xi_j=0]$, then we may represent $X=\xi_1+\cdots+\xi_N$. In this case

$$\begin{aligned}\mu&=\mathbb{E}\xi_i=\frac{1}{2}\\ \sigma^2&=\operatorname{Var}(\xi_i)=\frac{1}{2}\cdot\frac{1}{2}=\frac{1}{4},\\ \nu&=\mathbb{E}N=\frac{1}{6}(1+2+\cdots+6)=\frac{21}{6}=\frac{7}{2},\\ \tau^2&=\operatorname{Var}(N)=\frac{91}{6}-\left(\frac{21}{6}\right)^2=\frac{35}{12}\end{aligned}$$

where we used[2] $\mathbb{E}N^2=\frac{1}{6}\sum_{j=1}^{6}j^2=\frac{91}{6}$. Thus we find,

$$\mathbb{E}X=\mu\nu=\frac{7}{4} \text{ and}$$

$$\operatorname{Var}(X)=\nu\sigma^2+\mu^2\tau^2=\frac{7}{2}\cdot\frac{1}{4}+\left(\frac{1}{2}\right)^2\frac{35}{12}=\frac{77}{48}.$$

[2] See see the end of Appendix B for general formula of this type.

3.4 Wald's Equation and Gambler's Ruin

See the Wikipedia page;

$$https://en.wikipedia.org/wiki/Wald\%27s\_equation$$

for a more general version of the following theorem.

Theorem 3.12 (Wald's Equation). *Lets $\{\xi_j\}_{j=1}^{\infty}$ be a sequence of random variables and $N \in \mathbb{N}_0$ be a random time and suppose that;*

1. *$\mathbb{E}N < \infty$ and $\mathbb{E}|\xi_j| < \infty$ for each j.*
2. *$\mu = \mathbb{E}\xi_j$ and $\tilde{\mu} := \mathbb{E}|\xi_j|$ are independent of $j \in \mathbb{N}$.*
3. *$1_{j\le N}$ and ξ_j are independent for each $j \in \mathbb{N}$.*

If, as above, $X_N := \sum_{j=1}^{\infty} 1_{j\le N}\cdot\xi_j$, then

$$\mathbb{E}X_N = \mu\cdot\mathbb{E}N = \mathbb{E}\xi_1\cdot\mathbb{E}N. \tag{3.5}$$

***Moreover**; if $\xi_j \ge 0$ for all j, the Eq. (3.5) holds even when $\mathbb{E}N = \infty$. [In this latter case one should interpret $\mu\cdot\mathbb{E}N := 0$ even when $\mu = \mathbb{E}\xi_1 = 0$ or $\mathbb{E}N = \infty$.]*

Proof. Here is the basic calculation;

$$\begin{aligned}\mathbb{E}[X_N] &= \mathbb{E}\left[\sum_{j=1}^{\infty} 1_{j\le N}\cdot\xi_j\right] = \sum_{j=1}^{\infty}\mathbb{E}[1_{j\le N}\cdot\xi_j]\\ &= \sum_{j=1}^{\infty}\mathbb{E}[\xi_j]\,\mathbb{E}[1_{j\le N}] = \sum_{j=1}^{\infty}\mu\cdot\mathbb{E}[1_{j\le N}]\\ &= \mu\cdot\mathbb{E}\left[\sum_{j=1}^{\infty}1_{j\le N}\right] = \mu\cdot\mathbb{E}N.\end{aligned}$$

The only question is whether all of the interchanges of infinite sums and expectations are justified. By general measure theory summarized in Appendix A, the above calculation with ξ_j replaced by $|\xi_j|$ is always valid and this implies

$$\mathbb{E}\left[\sum_{j=1}^{\infty}1_{j\le N}\cdot|\xi_j|\right] = \tilde{\mu}\cdot\mathbb{E}N < \infty.$$

It then again follows by the general theory of the expectations that this is what is needed in order to justify the previous calculation. ∎

Corollary 3.13. *Suppose that $\{\xi_j\}_{j=1}^{\infty} \subset L^1(\mathbb{P})$ are i.i.d. random variables, $\mu = \mathbb{E}\xi_i$, and $N \in \mathbb{N}_0$ is a random "**stopping**" time, i.e.*

$$\{j \le N\} \text{ depends only on } \{\xi_1,\dots,\xi_{j-1}\} \text{ for each } j \in \mathbb{N}. \tag{3.6}$$

If we further assume that $\mathbb{E}N < \infty$ or that $\xi_j \ge 0$ for all j, then

$$\mathbb{E}X_N = \mu\cdot\mathbb{E}N.$$

Proof. The stopping time assumption in Eq. (3.6) means (more precisely) that

$$1_{j\le N} = \text{a function of } (\xi_1,\dots,\xi_{j-1}).$$

Since $\{\xi_j\}_{j=1}^{\infty}$ are independent, it follows that ξ_j is independent of $1_{j\le N}$ and hence the corollary follows from Wald's identity, Theorem 3.12. ∎

Example 3.14. Let $\{\xi_j\}_{j=1}^{\infty}$ be $\{0,1\}$-valued random variables and

$$N := \min\{j : \xi_1+\cdots+\xi_j = 10\}.$$

[Flip a coin sides labeled by 0 and 1. Then N is the first time you have flipped 10 ones.] Since

$$\{j \le N\} = \#\{k \le j-1 : \xi_k = 1\} < 10$$

we see that N is a stopping time.

Further assume $0 < p \le 1$ and $\{\xi_j\}_{j=1}^{\infty}$ are i.i.d. with $\xi_j \overset{d}{=} Bern(p)$ - random variables, i.e. $\mathbb{P}(\xi_j = 1) = p$ and $\mathbb{P}(\xi_j = 0) = q = 1-p$. Then

$$10 = \mathbb{E}\left[\sum_{j=1}^{N}\xi_j\right] = \mathbb{E}\xi_1\cdot\mathbb{E}N = p\cdot\mathbb{E}N \implies \mathbb{E}N = \frac{10}{p}. \tag{3.7}$$

[Note that $\mathbb{P}[N=\infty] =$"q^{∞}"$= 0$, so that $\sum_{j=1}^{N}\xi_j = 10$ a.s.]

Remark 3.15. The Random time, N, in Example 3.14 may be written as $N = W_{10} + 10$ where W_{10} is the negative binomial distribution as in Definition 2.7 with $r = 10$. It then follows by Corollary 2.10, that

$$\mathbb{E}N = \mathbb{E}W_{10} + 10 = 10\cdot\frac{1-p}{p} + 10 = \frac{10}{p}$$

in agreement with Eq. (3.7).

Example 3.16 (Gambler's ruin). Let $\{\xi_j\}_{j=1}^{\infty}$ be i.i.d. such that $\mathbb{P}(\xi_j = -1) = \mathbb{P}(\xi_j = 1) = 1/2$ and let

$$N := \min\{j : \xi_1+\cdots+\xi_j = 1\}.$$

[So N represents the first time a gambler is ahead by 1\$ in a betting game based on the flips of a fair cone.] Notice that

$$\{j \le N\} = \cap_{k<j} \{\xi_1 + \cdots + \xi_k \le 0\}$$

which shows N is a stopping time.

Claim: $\mathbb{E}N = \infty$.

Proof. If $\mathbb{E}N < \infty$, then $N < \infty$ a.s. and hence

$$\xi_1 + \cdots + \xi_N = 1 \text{ a.s..}$$

Taking expectations while using Wald's equation and $\mathbb{E}\xi_j = 0$ would lead to the following contradiction,

$$1 = \mathbb{E}\left[\sum_{j=1}^{N} \xi_j\right] = \mathbb{E}\xi_1 \cdot \mathbb{E}N = 0 \cdot \mathbb{E}N = 0.$$

Hence it must be that

$$\mathbb{E}N = \mathbb{E}\left[\text{first time that a gambler is ahead by } 1\right] = \infty.$$

3.5 A Review of Correlation and Independence

Notation 3.17 (Means, Variances, etc.) *Given square integrable random variables X and Y, let*

1. a) $\mu_X := \mathbb{E}X$ be the **mean** of X.
   b) $\operatorname{Var}(X) := \mathbb{E}\left[(X-\mu_X)^2\right] = \mathbb{E}X^2 - \mu_X^2$ be the **variance** of X.
   c) $\sigma_X = \sigma(X) := \sqrt{\operatorname{Var}(X)}$ be the **standard deviation** of X.
   d) $\operatorname{Cov}(X,Y) := \mathbb{E}\left[(X-\mu_X)(Y-\mu_Y)\right] = \mathbb{E}[XY] - \mu_X\mu_Y$ be the **covariance** of X and Y.
   e) The **correlation** of X and Y is defined to be

$$\operatorname{Corr}(X,Y) := \frac{\operatorname{Cov}(X,Y)}{\sqrt{\operatorname{Var}(X)\cdot\operatorname{Var}(Y)}} = \frac{\operatorname{Cov}(X,Y)}{\sigma_X \cdot \sigma_Y} \in [-1,1].$$

   f) The random variables X and Y are uncorrelated if $\operatorname{Cov}(X,Y) = 0$ or equivalently if $\operatorname{Corr}(X,Y) = 0$. We further say that a collection $\{X_k\}_{k=1}^n$ are **uncorrelated** if $\operatorname{Cov}(X_k, X_l) = 0$ for all $k \ne l$.

The following Lemma lists the basic properties of variances and covariances that I hope you already basically know.

Lemma 3.18. *The following properties hold.*

1. $\operatorname{Var}(X) = \operatorname{Cov}(X,X)$.
2. $\operatorname{Cov}(X,Y) = \operatorname{Cov}(Y,Y)$.
3. $\operatorname{Cov}(X,Y) = 0$ *if either X or Y is constant.*
4. $\operatorname{Cov}(X,Y)$ *is bilinear in X and Y, i.e.*

$$\operatorname{Cov}(X_1 + \lambda X_2, Y) = \operatorname{Cov}(X_1,Y) + \lambda \operatorname{Cov}(X_1,Y) \tag{3.8}$$

   where $\lambda \in \mathbb{R}$ and $\{X_1, X_2, Y\}$ are square integrable random variables.
5. *For any constant $\lambda \in \mathbb{R}$,*

$$\operatorname{Var}(X+\lambda) = \operatorname{Var}(X) \text{ and } \operatorname{Var}(\lambda X) = \lambda^2 \operatorname{Var}(X).$$

6. *If $\{X_j\}_{j=1}^n$ are uncorrelated $L^2(P)$ – random variables, then*

$$\operatorname{Var}(X_1 + \cdots + X_n) = \sum_{j=1}^{n} \operatorname{Var}(X_j). \tag{3.9}$$

Proof. I will leave most of these results to the reader and only verify Eqs. (3.8) and (3.9). For Eq. (3.8) we have using the linearity of the expectation that,

$$\begin{aligned}
\operatorname{Cov}(X_1 + \lambda X_2, Y) &= \mathbb{E}\left[(X_1 + \lambda X_2)\cdot Y\right] - \mathbb{E}\left[(X_1 + \lambda X_2)\right]\cdot \mathbb{E}Y \\
&= \mathbb{E}[X_1 Y] + \lambda \mathbb{E}[X_2 Y] - [\mathbb{E}X_1 + \lambda \mathbb{E}X_2]\cdot \mathbb{E}Y \\
&= (\mathbb{E}[X_1 Y] - \mathbb{E}X_1 \cdot \mathbb{E}Y) + \lambda(\mathbb{E}[X_2 Y] - \mathbb{E}X_2 \cdot \mathbb{E}Y) \\
&= \operatorname{Cov}(X_1,Y) + \lambda \operatorname{Cov}(X_1,Y).
\end{aligned}$$

To prove Eq. (3.9) let $Y := X_1 + \cdots + X_n$, then (by what we just proved and induction),

$$\begin{aligned}
\operatorname{Var}(Y) &= \operatorname{Cov}(Y,Y) = \operatorname{Cov}(X_1 + \cdots + X_n, Y) \\
&= \sum_{i=1}^{n} \operatorname{Cov}(X_i, Y)
\end{aligned}$$

and similarly,

$$\operatorname{Cov}(X_i, Y) = \operatorname{Cov}(X_i, X_1 + \cdots + X_n) = \sum_{j=1}^{n} \operatorname{Cov}(X_i, X_j).$$

Combining the last two equations shows, in general, that

$$\operatorname{Var}(X_1 + \cdots + X_n) = \sum_{i,j=1}^{n} \operatorname{Cov}(X_i, X_j).$$

If the $\{X_j\}_{j=1}^n$ are uncorrelated, then $\mathrm{Cov}(X_i, X_j) = 0$ unless $i = j$ and so the above sum reduces to

$$\mathrm{Var}(X_1 + \cdots + X_n) = \sum_{j=1}^n \mathrm{Cov}(X_j, X_j) = \sum_{j=1}^n \mathrm{Var}(X_j).$$

Bonus proof, let me also show that $\mathrm{Var}(X + \lambda) = \mathrm{Var}(X)$ when λ is a constant;

$$\begin{aligned}\mathrm{Var}(X + \lambda) &= \mathrm{Cov}(X + \lambda, X + \lambda) = \mathrm{Cov}(X + \lambda, X) + \mathrm{Cov}(X + \lambda, \lambda) \\ &= \mathrm{Cov}(X + \lambda, X) = \mathrm{Cov}(X, X) + \mathrm{Cov}(\lambda, X) \\ &= \mathrm{Cov}(X, X) = \mathrm{Var}(X),\end{aligned}$$

wherein we have used the bilinearity of $\mathrm{Cov}(\cdot, \cdot)$ and the property (you should verify) that $\mathrm{Cov}(Y, \lambda) = 0$ whenever λ is a constant. ∎

Theorem 3.19 (Independence = Uncorrelated on Steroids). *If $X, Y : \Omega \to \mathbb{R}$ are random variables then X and Y are independent iff $f(X)$ and $g(Y)$ are uncorrelated (i.e. $\mathrm{Cov}(f(X), g(Y)) = 0$) for all functions $f, g : \mathbb{R} \to \mathbb{R}$ such that $f(X)$ and $g(Y)$ are square integrable.*

***Note well** that X and Y being uncorrelated is a much weaker condition than X and Y being independent as it only requires that $\mathrm{Cov}(X, Y) = 0$, i.e. we only perform the independence test with $f(x) = g(x) = x$!*

Proof. This is just a simple rewriting of item 3. of Theorem 2.4. ∎

Corollary 3.20. *If $\{X_j\}_{j=1}^n$ are independent random functions and f_j are real valued functions on the range of X_j such that $\mathbb{E}|f_j(X_j)|^2 < \infty$ for each j, then*

$$\mathrm{Var}(f_1(X_1) + \cdots + f_n(X_n)) = \sum_{j=1}^n \mathrm{Var}(f_j(X_j)).$$

Proof. Combine item 6. of Lemma 3.18 with Theorem 3.19. ∎

Part II

Appendix

A

Basics of Probabilities and Expectations

The goal of this appendix is to describe modern probability with "sufficient" precision to allow us to do the required computations for this course. We will thus be neglecting some technical details involving measures and σ – algebras. The knowledgeable reader should be able to fill in the missing hypothesis while the less knowledgeable readers should not be too harmed by the omissions to follow.

1. $(\Omega, \mathbb{P})$ will denote a **probability space** and S will denote a set which is called **state space**. Informally put, Ω is a set (often the sample space) and $\mathbb{P}$ is a function on all[1] subsets of Ω (subsets of Ω are called **events)** with the following properties;
   a) $\mathbb{P}(A) \in [0,1]$ for all $A \subset \Omega$,
   b) $\mathbb{P}(\Omega) = 1$ and $\mathbb{P}(\emptyset) = 0$.
   c) $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$ is $A \cap B = \emptyset$. More generally, if $A_n \subset \Omega$ for all n with $A_n \cap A_m = \emptyset$ for $m \neq n$ we have

$$\mathbb{P}\left(\cup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(A_n).$$

2. A **random variable,** Z, is a function from Ω to $\mathbb{R}$ or perhaps some other range space. For example if $A \subset \Omega$ is an event then the **indicator function of** A,

$$1_A(\omega) := \begin{cases} 1 \text{ if } \omega \in A \\ 0 \text{ if } \omega \notin A, \end{cases}$$

   is a random variable.

[1] This is often a lie! Nevertheless, for our purposes it will be reasonably safe to ignore this lie.

3. Note that every real value random variable, Z, may be approximated by the discrete random variables

$$Z_\varepsilon := \sum_{n \in \mathbb{Z}} n\varepsilon \cdot 1_{\{n\varepsilon \le Z < (n+1)\varepsilon\}} \text{ for all } \varepsilon > 0. \tag{A.1}$$

   As we usually do in probability, $\{n\varepsilon \le Z < (n+1)\varepsilon\}$, stands for the event more precisely written as;

$$\{\omega \in \Omega : n\varepsilon \le Z(\omega) < (n+1)\varepsilon\}.$$

4. $\mathbb{E}Z$ will denote the **expectation** of a random variable, $Z : \Omega \to \mathbb{R}$ which is defined as follows. If Z only takes on a finite number of real values $\{z_1, \dots, z_m\}$ we define

$$\mathbb{E}Z = \sum_{i=1}^{m} z_i \mathbb{P}(Z = z_i).$$

   For general $Z \ge 0$ we set $\mathbb{E}Z = \lim_{n\to\infty} \mathbb{E}Z_n$ where $\{Z_n\}_{n=1}^{\infty}$ is any sequence of discrete random variables such that $0 \le Z_n \uparrow Z$ as $n \uparrow \infty$. Finally if Z is real valued with $\mathbb{E}|Z| < \infty$ (in which case we say Z is **integrable)** we set $\mathbb{E}Z = \mathbb{E}\mathbb{Z}_+ - \mathbb{E}Z_-$ where $Z_\pm = \max(\pm Z, 0)$. With these definition one eventually shows via the dominated convergence theorem below; if $f : \mathbb{R} \to \mathbb{R}$ is a bounded continuous function, then

$$\mathbb{E}[f(Z)] = \lim_{\Delta \to 0} \sum_{n \in \mathbb{Z}} f(n\Delta)\, \mathbb{P}(n\Delta < Z \le (n+1)\Delta).$$

   We summarize this informally[2] by writing;

$$\mathbb{E}[f(Z)] = \text{“} \int_{\mathbb{R}} f(z)\, \mathbb{P}(z < Z \le z + dz) \text{.”}$$

[2] Think of $z = n\Delta$ and $dz = \Delta$.

5. The expectation has the following basic properties;
   a) **Expectations of indicator functions:** $\mathbb{E}1_A = \mathbb{P}(A)$ for all events $A \subset \Omega$.
   b) **Linearity**: if X and Y are integrable random variables and $c \in \mathbb{R}$, then
   $$\mathbb{E}[X + cY] = \mathbb{E}X + c\mathbb{E}Y.$$
   c) **Monotinicity:** if $X, Y : \Omega \to \mathbb{R}$ are integrable with $\mathbb{P}(X \le Y) = 1$, then $\mathbb{E}X \le \mathbb{E}Y$. In particular if $X = Y$ **almost surely** (a.s.) (i.e. $\mathbb{P}(X = Y) = 1$), then $\mathbb{E}X = \mathbb{E}Y$. [What happens on sets of probability 0 are typically irrelevant.]
   d) **Finite expectation $\implies$ finite random variable.** If $Z : \Omega \to [0, \infty]$ is a random variable such that $\mathbb{E}Z < \infty$ then $\mathbb{P}(Z = \infty) = 0$, i.e. $\mathbb{P}(Z < \infty) = 1$.
   e) **MCT**: the **monotone convergence theorem** holds; if $0 \le Z_n \uparrow Z$ then
   $$\uparrow \lim_{n\to\infty} \mathbb{E}[Z_n] = \mathbb{E}[Z] \text{ (with } \infty \text{ allowed as a possible value).}$$
   Example 1: If $\{A_n\}_{n=1}^\infty$ is a sequence of events such that $A_n \uparrow A$ (i.e. $A_n \subset A_{n+1}$ for all n and $A = \cup_{n=1}^\infty A_n$), then
   $$\mathbb{P}(A_n) = \mathbb{E}[1_{A_n}] \uparrow \mathbb{E}[1_A] = \mathbb{P}(A) \text{ as } n \to \infty$$
   Example 2: If $X_n : \Omega \to [0, \infty]$ for $n \in \mathbb{N}$ then
   $$\mathbb{E}\sum_{n=1}^\infty X_n = \mathbb{E}\lim_{N\to\infty}\sum_{n=1}^N X_n = \lim_{N\to\infty}\mathbb{E}\sum_{n=1}^N X_n = \lim_{N\to\infty}\sum_{n=1}^N \mathbb{E}X_n = \sum_{n=1}^\infty \mathbb{E}X_n.$$
   Example 3: Suppose S is a finite or countable set and $X : \Omega \to S$ is a random function. Then for any $f : S \to [0, \infty]$,
   $$\mathbb{E}[f(X)] = \sum_{s\in S} f(s)\,\mathbb{P}(X = s).$$
   Indeed, we have
   $$f(X) = \sum_{s\in S} f(s)\,1_{\{X=s\}}$$
   and so by Example 2. above,
   $$\begin{aligned}\mathbb{E}[f(X)] &= \sum_{s\in S}\mathbb{E}\left[f(s)\,1_{\{X=s\}}\right] \\ &= \sum_{s\in S} f(s)\,\mathbb{E}\left[1_{\{X=s\}}\right] = \sum_{s\in S} f(s)\,\mathbb{P}(X = s).\end{aligned}$$
   f) **DCT**: the **dominated convergence theorem** holds**,** if
   $$\mathbb{E}\left[\sup_n |Z_n|\right] < \infty \text{ and } \lim_{n\to\infty} Z_n = Z, \text{ then}$$
   $$\mathbb{E}\left[\lim_{n\to\infty} Z_n\right] = \mathbb{E}Z = \lim_{n\to\infty}\mathbb{E}Z_n.$$
   Example 1: If $\{A_n\}_{n=1}^\infty$ is a sequence of events such that $A_n \downarrow A$ (i.e. $A_n \supset A_{n+1}$ for all n and $A = \cap_{n=1}^\infty A_n$), then
   $$\mathbb{P}(A_n) = \mathbb{E}[1_{A_n}] \downarrow \mathbb{E}[1_A] = \mathbb{P}(A) \text{ as } n \to \infty.$$
   The dominating function is 1 here.
   Example 2: If $\{X_n\}_{n=1}^\infty$ is a sequence of real valued random variables such that
   $$\mathbb{E}\sum_{n=1}^\infty |X_n| = \sum_{n=1}^\infty \mathbb{E}|X_n| < \infty,$$
   then; 1) $Z := \sum_{n=1}^\infty |X_n| < \infty$ a.s. and hence $\sum_{n=1}^\infty X_n = \lim_{N\to\infty}\sum_{n=1}^N X_n$ exist a.s., 2) $\left|\sum_{n=1}^N X_n\right| \le Z$ and $\mathbb{E}Z < \infty$, and so 3) by DCT,
   $$\mathbb{E}\sum_{n=1}^\infty X_n = \mathbb{E}\lim_{N\to\infty}\sum_{n=1}^N X_n = \lim_{N\to\infty}\mathbb{E}\sum_{n=1}^N X_n = \lim_{N\to\infty}\sum_{n=1}^N \mathbb{E}X_n = \sum_{n=1}^\infty \mathbb{E}X_n.$$
   g) **Fatou's Lemma: Fatou's lemma** holds; if $0 \le Z_n \le \infty$, then
   $$\mathbb{E}\left[\liminf_{n\to\infty} Z_n\right] \le \liminf_{n\to\infty}\mathbb{E}[Z_n].$$
   This may be proved as an application of MCT.
6. **Discrete distributions.** If S is a discrete set, i.e. finite or countable and $X : \Omega \to S$ we let
   $$\rho_X(s) := \mathbb{P}(X = s).$$
   Notice that if $f : S \to \mathbb{R}$ is a function, then $f(X) = \sum_{s\in S} f(s)\,1_{\{X=s\}}$ and therefore,
   $$\mathbb{E}f(X) = \sum_{s\in S} f(s)\,\mathbb{E}1_{\{X=s\}} = \sum_{s\in S} f(s)\,\mathbb{P}(X = s) = \sum_{s\in S} f(s)\,\rho_X(s).$$
   More generally if $X_i : \Omega \to S_i$ for $1 \le i \le n$ we let
   $$\rho_{X_1,\dots,X_n}(\mathbf{s}) := \mathbb{P}(X_1 = s_1, \dots, X_n = s_n)$$
   for all $\mathbf{s} = (s_1, \dots, s_n) \in S_1 \times \cdots \times S_n$ and
   $$\mathbb{E}f(X_1, \dots, X_n) = \sum_{\mathbf{s}=(s_1,\dots,s_n)} f(\mathbf{s})\,\rho_{X_1,\dots,X_n}(\mathbf{s}).$$

7. **Continuous density functions.** If S is $\mathbb{R}$ or $\mathbb{R}^n$, we say $X : \Omega \to S$ is a "**continuous random variable,**" if there exists a **probability density function**, $\rho_X : S \to [0, \infty)$ such that for all bounded (or positive) functions, $f : S \to \mathbb{R}$, we have

$$\mathbb{E}[f(X)] = \int_S f(x)\, \rho_X(x)\, dx.$$

8. Given random variables X and Y we let;
   a) $\mu_X := \mathbb{E}X$ be the **mean** of X.
   b) $\mathrm{Var}(X) := \mathbb{E}\left[(X - \mu_X)^2\right] = \mathbb{E}X^2 - \mu_X^2$ be the **variance** of X.
   c) $\sigma_X = \sigma(X) := \sqrt{\mathrm{Var}(X)}$ be the **standard deviation** of X.
   d) $\mathrm{Cov}(X, Y) := \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] = \mathbb{E}[XY] - \mu_X \mu_Y$ be the **covariance** of X and Y.
   e) $Corr(X, Y) := \mathrm{Cov}(X, Y) / (\sigma_X \sigma_Y)$ be the **correlation** of X and Y.
9. **Tonelli's theorem;** if $f : \mathbb{R}^k \times \mathbb{R}^l \to \mathbb{R}_+$, then

$$\int_{\mathbb{R}^k} dx \int_{\mathbb{R}^l} dy f(x, y) = \int_{\mathbb{R}^l} dy \int_{\mathbb{R}^k} dx f(x, y) \text{ (with } \infty \text{ being allowed).}$$

10. **Fubini's theorem;** if $f : \mathbb{R}^k \times \mathbb{R}^l \to \mathbb{R}$ is a function such that

$$\int_{\mathbb{R}^k} dx \int_{\mathbb{R}^l} dy |f(x, y)| = \int_{\mathbb{R}^l} dy \int_{\mathbb{R}^k} dx |f(x, y)| < \infty,$$

then

$$\int_{\mathbb{R}^k} dx \int_{\mathbb{R}^l} dy f(x, y) = \int_{\mathbb{R}^l} dy \int_{\mathbb{R}^k} dx f(x, y).$$

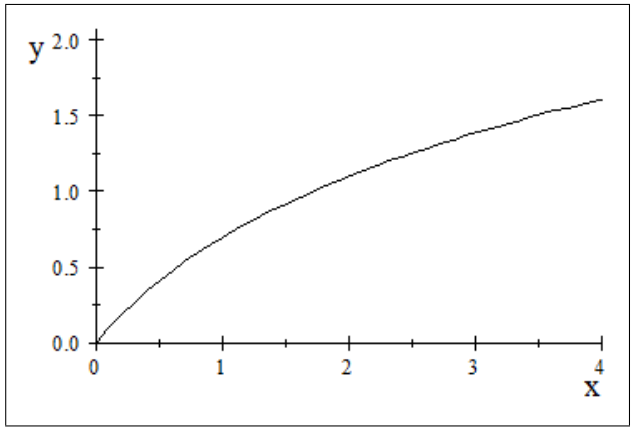
B

Analytic Facts

B.1 A Stirling's Formula Like Approximation

Theorem B.1. *Suppose that $f : (0,\infty) \to \mathbb{R}$ is an increasing concave down function (like $f(x) = \ln x$) and let $s_n := \sum_{k=1}^{n} f(k)$, then*

$$\begin{aligned} s_n - \frac{1}{2}(f(n) + f(1)) &\le \int_1^n f(x)\,dx \\ &\le s_n - \frac{1}{2}[f(n+1) + 2f(1)] + \frac{1}{2}f(2) \\ &\le s_n - \frac{1}{2}[f(n) + 2f(1)] + \frac{1}{2}f(2). \end{aligned}$$



Proof. On the interval, $[k-1,k]$, we have that $f(x)$ is larger than the straight line segment joining $(k-1, f(k-1))$ and $(k, f(k))$ and thus

$$\frac{1}{2}(f(k) + f(k-1)) \le \int_{k-1}^{k} f(x)\,dx.$$

Summing this equation on $k = 2,\dots,n$ shows,

$$\begin{aligned} s_n - \frac{1}{2}(f(n) + f(1)) &= \sum_{k=2}^{n} \frac{1}{2}(f(k) + f(k-1)) \\ &\le \sum_{k=2}^{n} \int_{k-1}^{k} f(x)\,dx = \int_1^n f(x)\,dx. \end{aligned}$$

For the upper bound on the integral we observe that $f(x) \le f(k) - f'(k)(x-k)$ for all x and therefore,

$$\int_{k-1}^{k} f(x)\,dx \le \int_{k-1}^{k} [f(k) - f'(k)(x-k)]\,dx = f(k) - \frac{1}{2}f'(k).$$

Summing this equation on $k = 2,\dots,n$ then implies,

$$\int_1^n f(x)\,dx \le \sum_{k=2}^{n} f(k) - \frac{1}{2}\sum_{k=2}^{n} f'(k).$$

Since $f''(x) \le 0$, $f'(x)$ is decreasing and therefore $f'(x) \le f'(k-1)$ for $x \in [k-1,k]$ and integrating this equation over $[k-1,k]$ gives

$$f(k) - f(k-1) \le f'(k-1).$$

Summing the result on $k = 3,\dots,n+1$ then shows,

$$f(n+1) - f(2) \le \sum_{k=2}^{n} f'(k)$$

and thus ti follows that

$$\begin{aligned}\int_1^n f(x)\,dx &\le \sum_{k=2}^n f(k) - \frac{1}{2}\left(f(n+1)-f(2)\right)\\ &= s_n - \frac{1}{2}\left[f(n+1)+2f(1)\right] + \frac{1}{2}f(2)\\ &\le s_n - \frac{1}{2}\left[f(n)+2f(1)\right] + \frac{1}{2}f(2)\end{aligned}$$

■

Example B.2 (Approximating n!). Let us take $f(n)=\ln n$ and recall that

$$\int_1^n \ln x dx = n\ln n - n + 1.$$

Thus we may conclude that

$$s_n - \frac{1}{2}\ln n \le n\ln n - n + 1 \le s_n - \frac{1}{2}\ln n + \frac{1}{2}\ln 2.$$

Thus it follows that

$$\left(n+\frac{1}{2}\right)\ln n - n + 1 - \ln\sqrt{2} \le s_n \le \left(n+\frac{1}{2}\right)\ln n - n + 1.$$

Exponentiating this identity then gives the following upper and lower bounds on $n!$;

$$\frac{e}{\sqrt{2}}\cdot e^{-n}n^{n+1/2} \le n! \le e\cdot e^{-n}n^{n+1/2}.$$

These bound compare well with Strirling's formula (Theorem B.5) which implies,

$$n!\sim\sqrt{2\pi}e^{-n}n^{n+1/2} \overset{\text{by definition}}{\iff} \lim_{n\to\infty}\frac{n!}{e^{-n}n^{n+1/2}} = \sqrt{2\pi}.$$

Observe that

$$\frac{e}{\sqrt{2}} \cong 1.922\,1 \le \sqrt{2\pi} \cong 2.506 \le e \cong 2.718\,3.$$

Definition B.3 (Gamma Function). *The **Gamma function**, $\Gamma:\mathbb{R}_+\to\mathbb{R}_+$ is defined by*

$$\Gamma(x) := \int_0^\infty u^{x-1}e^{-u}du \tag{B.1}$$

(The reader should check that $\Gamma(x)<\infty$ for all $x>0$.)

Here are some of the more basic properties of this function.

Example B.4 (Γ – function properties). Let Γ be the gamma function, then;

1. $\Gamma(1)=1$ as is easily verified.
2. $\Gamma(x+1)=x\Gamma(x)$ for all $x>0$ as follows by integration by parts;

$$\begin{aligned}\Gamma(x+1) &= \int_0^\infty e^{-u}\,u^{x+1}\,\frac{du}{u} = \int_0^\infty u^x\left(-\frac{d}{du}\,e^{-u}\right)du\\ &= x\int_0^\infty u^{x-1}\,e^{-u}\,du = x\,\Gamma(x).\end{aligned}$$

In particular, it follows from items 1. and 2. and induction that

$$\Gamma(n+1)=n! \text{ for all } n\in\mathbb{N}. \tag{B.2}$$

3. $\Gamma(1/2)=\sqrt{\pi}$. This last assertion is a bit trickier. One proof is to make use of the fact (proved below in Lemma D.1) that

$$\int_{-\infty}^\infty e^{-ar^2}dr = \sqrt{\frac{\pi}{a}} \text{ for all } a>0. \tag{B.3}$$

Taking $a=1$ and making the change of variables, $u=r^2$ below implies,

$$\sqrt{\pi} = \int_{-\infty}^\infty e^{-r^2}dr = 2\int_0^\infty u^{-1/2}e^{-u}du = \Gamma(1/2).$$

$$\begin{aligned}\Gamma(1/2) &= 2\int_0^\infty e^{-r^2}dr = \int_{-\infty}^\infty e^{-r^2}dr\\ &= I_1(1) = \sqrt{\pi}.\end{aligned}$$

4. A simple induction argument using items 2. and 3. now shows that

$$\Gamma\left(n+\frac{1}{2}\right) = \frac{(2n-1)!!}{2^n}\sqrt{\pi}$$

where $(-1)!!:=1$ and $(2n-1)!!=(2n-1)(2n-3)\dots3\cdot1$ for $n\in\mathbb{N}$.

Theorem B.5 (Stirling's formula). *The Gamma function (see Definition B.3), satisfies Stirling's formula,*

$$\lim_{x\to\infty}\frac{\Gamma(x+1)}{\sqrt{2\pi}e^{-x}x^{x+1/2}} = 1. \tag{B.4}$$

In particular, if $n\in\mathbb{N}$, we have

$$n! = \Gamma(n+1) \sim \sqrt{2\pi}e^{-n}n^{n+1/2}$$

where we write $a_n\sim b_n$ to mean, $\lim_{n\to\infty}\frac{a_n}{b_n}=1$.

B.2 Formula for integer valued unifrom distributions

By various means (and easily proved by induction) one finds the formulas;

$$\sum_{j=1}^{n} j = \frac{1}{2} n\,(n+1)$$
$$\sum_{j=1}^{n} j^2 = \frac{1}{6} n\,(2n+1)\,(n+1)\,.$$

Suppose now that $Y \in [n] = \{1, 2, \ldots, n\}$ is chosen uniformly at random, then

$$\mathbb{E}Y = \frac{1}{n} \cdot \sum_{j=1}^{n} j = \frac{n+1}{2},$$
$$\mathbb{E}Y^2 = \frac{1}{n} \cdot \sum_{j=1}^{n} j^2 = \frac{1}{6}\,(2n+1)\,(n+1)\,,$$

and hence

$$\begin{aligned}
\mathrm{Var}\,(Y) &= \frac{1}{6}\,(2n+1)\,(n+1) - \left(\frac{n+1}{2}\right)^2 \\
&= (n+1)\left[\frac{1}{6}\,(2n+1) - \frac{n+1}{4}\right] \\
&= \frac{1}{12}\,(n-1)\,(n+1) = \frac{1}{12}\left(n^2-1\right).
\end{aligned}$$

So for example if $n = 6$ (for dice) then (as we have used in the text),

$$\mathbb{E}Y = \frac{7}{2} \text{ and } \mathrm{Var}\,(Y) = \frac{1}{12}\left(6^2-1\right) = \frac{35}{12}.$$

C

Independence

Definition C.1. *We say that an event, A, is **independent** of an event, B, iff*[1]

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

*We further say a collection of events $\{A_j\}_{j\in J}$ are **independent** iff*

$$\mathbb{P}(\cap_{j\in J_0} A_j) = \prod_{j\in J_0} \mathbb{P}(A_j)$$

for any finite subset, J_0, of J.

Lemma C.2. *If $\{A_j\}_{j\in J}$ is an independent collection of events then so is $\{A_j, A_j^c\}_{j\in J}$.*

Proof. First consider the case of two independent events, A and B. By assumption, $\mathbb{P}(A\cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$. Since A is the disjoint union of $A\cap B$ and $A\cap B^c$, the additivity of $\mathbb{P}$ implies,

$$\mathbb{P}(A) = \mathbb{P}(A\cap B) + \mathbb{P}(A\cap B^c) = \mathbb{P}(A)\,\mathbb{P}(B) + \mathbb{P}(A\cap B^c).$$

Solving this identity for $\mathbb{P}(A\cap B^c)$ gives,

$$\mathbb{P}(A\cap B^c) = \mathbb{P}(A)\,[1-\mathbb{P}(B)] = \mathbb{P}(A)\,\mathbb{P}(B^c).$$

Thus if $\{A, B\}$ are independent then so is $\{A, B^c\}$. Similarly we may show $\{A^c, B\}$ are independent and then that $\{A^c, B^c\}$ are independent. That is $\mathbb{P}\left(A^\varepsilon \cap B^\delta\right) = \mathbb{P}(A^\varepsilon)\,\mathbb{P}\left(B^\delta\right)$ where ε, δ is either "nothing" or "c."

[1] Shortly we will consider conditional probabilities, $\mathbb{P}(\cdot|B)$. With this notation, A is independent of B iff $\mathbb{P}(A|B) = \mathbb{P}(A)$, i.e. given the information gained by B occurring does not affect the likelihood that A occurred.

The general case now easily follows similarly. Indeed, if $\{A_1,\dots,A_n\} \subset \{A_j\}_{j\in J}$ we must show that

$$\mathbb{P}(A_1^{\varepsilon_1}\cap\cdots\cap A_n^{\varepsilon_n}) = \mathbb{P}(A_1^{\varepsilon_1})\dots\mathbb{P}(A_n^{\varepsilon_n})$$

where $\varepsilon_j = c$ or $\varepsilon_j =$ " ". But this follows from above. For example, $\{A_1\cap\cdots\cap A_{n-1}, A_n\}$ are independent implies that $\{A_1\cap\cdots\cap A_{n-1}, A_n^c\}$ are independent and hence

$$\begin{aligned}\mathbb{P}(A_1\cap\cdots\cap A_{n-1}\cap A_n^c) &= \mathbb{P}(A_1\cap\cdots\cap A_{n-1})\,\mathbb{P}(A_n^c)\\ &= \mathbb{P}(A_1)\dots\mathbb{P}(A_{n-1})\,\mathbb{P}(A_n^c).\end{aligned}$$

Thus we have shown it is permissible to add A_j^c to the list for any $j\in J$. ∎

Lemma C.3. *If $\{A_n\}_{n=1}^\infty$ is a sequence of independent events, then*

$$\mathbb{P}(\cap_{n=1}^\infty A_n) = \prod_{n=1}^\infty \mathbb{P}(A_n) := \lim_{N\to\infty}\prod_{n=1}^N \mathbb{P}(A_n).$$

Proof. Since $\cap_{n=1}^N A_n \downarrow \cap_{n=1}^\infty A_n$, it follows that

$$\mathbb{P}(\cap_{n=1}^\infty A_n) = \lim_{N\to\infty}\mathbb{P}\left(\cap_{n=1}^N A_n\right) = \lim_{N\to\infty}\prod_{n=1}^N \mathbb{P}(A_n),$$

where we have used the independence assumption for the last equality.

The convergence assertion used above follows from DCT Indeed, $1_{\cap_{n=1}^N A_n}\downarrow 1_{\cap_{n=1}^\infty A_n}$ and all functions are dominated by 1 and therefore,

$$\mathbb{P}(\cap_{n=1}^\infty A_n) = \mathbb{E}\left[1_{\cap_{n=1}^\infty A_n}\right] = \lim_{N\to\infty}\mathbb{E}\left[1_{\cap_{n=1}^N A_n}\right] = \lim_{N\to\infty}\mathbb{P}\left(\cap_{n=1}^N A_n\right).$$

∎

C.1 Borel Cantelli Lemmas

Definition C.4 (A_n i.o.). *Suppose that $\{A_n\}_{n=1}^{\infty}$ is a sequence of events. Let*

$$\{A_n \ i.o.\} := \left\{\sum_{n=1}^{\infty} 1_{A_n} = \infty\right\}$$

denote the event where infinitely many of the events, A_n, occur. The abbreviation, "i.o." stands for ***infinitely often****.*

For example if X_n is H or T depending on whether a heads or tails is flipped at the n^{th} step, then $\{X_n = H \text{ i.o.}\}$ is the event where an infinite number of heads was flipped.

Lemma C.5 (The First Borell – Cantelli Lemma). *If $\{A_n\}$ is a sequence of events such that $\sum_{n=0}^{\infty} \mathbb{P}(A_n) < \infty$, then*

$$\mathbb{P}(\{A_n \ i.o.\}) = 0.$$

Proof. Since

$$\infty > \sum_{n=0}^{\infty} \mathbb{P}(A_n) = \sum_{n=0}^{\infty} \mathbb{E}1_{A_n} = \mathbb{E}\left[\sum_{n=0}^{\infty} 1_{A_n}\right]$$

it follows that $\sum_{n=0}^{\infty} 1_{A_n} < \infty$ almost surely (a.s.), i.e. with probability 1 only finitely many of the $\{A_n\}$ can occur. ∎

Under the additional assumption of independence we have the following strong converse of the first Borel-Cantelli Lemma.

Lemma C.6 (Second Borel-Cantelli Lemma). *If $\{A_n\}_{n=1}^{\infty}$ are independent events, then*

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty \quad \Longrightarrow \quad \mathbb{P}(\{A_n \ i.o.\}) = 1. \tag{C.1}$$

Proof. We are going to show $\mathbb{P}(\{A_n \text{ i.o.}\}^c) = 0$. Since,

$$\{A_n \text{ i.o.}\}^c = \left\{\sum_{n=1}^{\infty} 1_{A_n} = \infty\right\}^c = \left\{\sum_{n=1}^{\infty} 1_{A_n} < \infty\right\},$$

we see that $\omega \in \{A_n \text{ i.o.}\}^c$ iff there exists $n \in \mathbb{N}$ such that $\omega \notin A_m$ for all $m \geq n$. Thus we have shown, if $\omega \in \{A_n \text{ i.o.}\}^c$ then $\omega \in B_n := \cap_{m \geq n} A_m^c$ for some n and therefore,

$$\{A_n \text{ i.o.}\}^c = \cup_{n=1}^{\infty} B_n.$$

As $B_n \uparrow \{A_n \text{ i.o.}\}^c$ we have

$$\mathbb{P}(\{A_n \text{ i.o.}\}^c) = \lim_{n \to \infty} \mathbb{P}(B_n).$$

But making use of the independence (see Lemmas C.2 and C.3) and the estimate, $1 - x \leq e^{-x}$, see Figure C.1 below, we find

$$\mathbb{P}(B_n) = \mathbb{P}(\cap_{m \geq n} A_m^c) = \prod_{m \geq n} \mathbb{P}(A_m^c) = \prod_{m \geq n} [1 - \mathbb{P}(A_m)]$$

$$\leq \prod_{m \geq n} e^{-\mathbb{P}(A_m)} = \exp\left(-\sum_{m \geq n} \mathbb{P}(A_m)\right) = e^{-\infty} = 0.$$

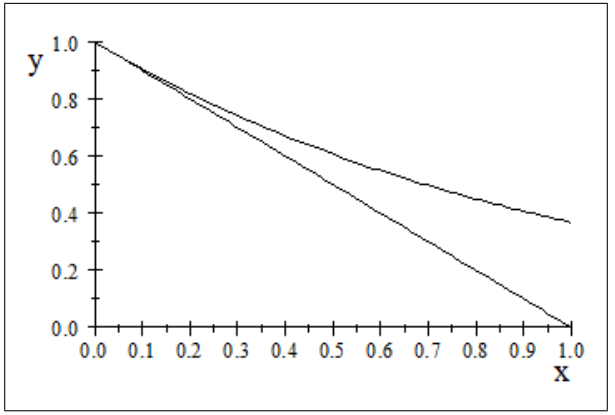


Fig. C.1. Comparing e^{-x} and $1 - x$.

∎

Combining the two Borel Cantelli Lemmas gives the following Zero-One Law.

Corollary C.7 (Borel's Zero-One law). *If $\{A_n\}_{n=1}^{\infty}$ are independent events, then*

$$\mathbb{P}(A_n \ i.o.) = \begin{cases} 0 \text{ if } \sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty \\ 1 \text{ if } \sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty \end{cases}.$$

Example C.8. If $\{X_n\}_{n=1}^{\infty}$ denotes the outcomes of the toss of a coin such that $\mathbb{P}(X_n = H) = p > 0$, then $\mathbb{P}(X_n = H \text{ i.o.}) = 1$.

Example C.9. If a monkey types on a keyboard with each stroke being independent and identically distributed with each key being hit with positive probability. Then eventually the monkey will type the text of the bible if she lives long enough. Indeed, let S be the set of possible key strokes and let $(s_1, \ldots, s_N)$ be the strokes necessary to type the bible. Further let $\{X_n\}_{n=1}^{\infty}$ be the strokes that the monkey types at time n. Then group the monkey's strokes as $Y_k := \left(X_{kN+1}, \ldots, X_{(k+1)N}\right)$. We then have

$$\mathbb{P}\left(Y_k = (s_1, \ldots, s_N)\right) = \prod_{j=1}^{N} \mathbb{P}\left(X_j = s_j\right) =: p > 0.$$

Therefore,

$$\sum_{k=1}^{\infty} \mathbb{P}\left(Y_k = (s_1, \ldots, s_N)\right) = \infty$$

and so by the second Borel-Cantelli lemma,

$$\mathbb{P}\left(\{Y_k = (s_1, \ldots, s_N)\} \text{ i.o. } k\right) = 1.$$

D

Multivariate Gaussians

The following basic Gaussian integration formula is needed in order to properly normalize Gaussian densities, see Definition D.2 below.

Lemma D.1. *Let $a > 0$ and for $d \in \mathbb{N}$ let,*

$$I_d(a) := \int_{\mathbb{R}^d} e^{-a|x|^2} dm(x).$$

Then $I_d(a) = (\pi/a)^{d/2}$.

Proof. By Tonelli's theorem and induction,

$$\begin{aligned} I_d(a) &= \int_{\mathbb{R}^{d-1}\times\mathbb{R}} e^{-a|y|^2} e^{-at^2} m_{d-1}(dy)\, dt \\ &= I_{d-1}(a) I_1(a) = I_1^d(a). \end{aligned} \tag{D.1}$$

So it suffices to compute:

$$I_2(a) = \int_{\mathbb{R}^2} e^{-a|x|^2} dm(x) = \int_{\mathbb{R}^2\setminus\{0\}} e^{-a(x_1^2+x_2^2)} dx_1 dx_2.$$

Using polar coordinates, we find,

$$\begin{aligned} I_2(a) &= \int_0^\infty dr\ r \int_0^{2\pi} d\theta\ e^{-ar^2} = 2\pi \int_0^\infty r e^{-ar^2} dr \\ &= 2\pi \lim_{M\to\infty} \int_0^M r e^{-ar^2} dr = 2\pi \lim_{M\to\infty} \frac{e^{-ar^2}}{-2a} \Big|_0^M = \frac{2\pi}{2a} = \pi/a. \end{aligned}$$

This shows that $I_2(a) = \pi/a$ and the result now follows from Eq. (D.1). ∎

D.1 Review of Gaussian Random Variables

Definition D.2 (Normal / Gaussian Random Variable). *A random variable, Y, **is normal with mean μ standard deviation σ^2** iff*

$$\mathbb{P}(Y \in (y, y+dy]) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y-\mu)^2} dy. \tag{D.2}$$

*We will abbreviate this by writing $Y \overset{d}{=} N(\mu, \sigma^2)$. When $\mu = 0$ and $\sigma^2 = 1$ we say Y is a **standard normal** random variable. We will often denote standard normal random variables by Z.*

Observe that Eq. (D.2) is equivalent to writing

$$\mathbb{E}[f(Y)] = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\mathbb{R}} f(y)\, e^{-\frac{1}{2\sigma^2}(y-\mu)^2} dy$$

for all bounded functions, $f : \mathbb{R} \to \mathbb{R}$. Also observe that $Y \overset{d}{=} N(\mu, \sigma^2)$ is equivalent to $Y \overset{d}{=} \sigma Z + \mu$. Indeed, by making the change of variable, $y = \sigma x + \mu$, we find

$$\begin{aligned} \mathbb{E}[f(\sigma Z + \mu)] &= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} f(\sigma x + \mu)\, e^{-\frac{1}{2}x^2} dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} f(y)\, e^{-\frac{1}{2\sigma^2}(y-\mu)^2} \frac{dy}{\sigma} = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\mathbb{R}} f(y)\, e^{-\frac{1}{2\sigma^2}(y-\mu)^2} dy. \end{aligned}$$

Lastly the constant, $\left(2\pi\sigma^2\right)^{-1/2}$ is chosen so that

$$\frac{1}{\sqrt{2\pi\sigma^2}} \int_{\mathbb{R}} e^{-\frac{1}{2\sigma^2}(y-\mu)^2} dy = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-\frac{1}{2}y^2} dy = 1.$$

Lemma D.3 (Integration by parts). *If $X \overset{d}{=} N\left(0,\sigma^2\right)$ for some $\sigma^2 \geq 0$ and $f:\mathbb{R}\to\mathbb{R}$ is a C^1 – function such that $Xf(X)$, $f'(X)$ and $f(X)$ are all integrable random variables and*[1] *$\lim_{z\to\pm\infty}\left[f(x)e^{-\frac{1}{2\sigma^2}x^2}\right]=0$, then*

$$\mathbb{E}\left[Xf(X)\right]=\sigma^2\mathbb{E}\left[f'(X)\right]=\mathbb{E}\left[X^2\right]\cdot\mathbb{E}\left[f'(X)\right]. \tag{D.3}$$

Proof. If $\sigma=0$ then $X=0$ a.s. and both sides of Eq. (D.3) are zero. So we now suppose that $\sigma>0$ and set $C:=1/\sqrt{2\pi\sigma^2}$. The result is a simple matter of using integration by parts;

$$\begin{aligned}\mathbb{E}\left[f'(X)\right]&=C\int_{\mathbb{R}}f'(x)\,e^{-\frac{1}{2\sigma^2}x^2}dx=C\lim_{M\to\infty}\int_{-M}^{M}f'(x)\,e^{-\frac{1}{2\sigma^2}x^2}dx\\&=C\lim_{M\to\infty}\left[f(x)\,e^{-\frac{1}{2\sigma^2}x^2}|_{-M}^{M}-\int_{-M}^{M}f(x)\,\frac{d}{dx}e^{-\frac{1}{2\sigma^2}x^2}dx\right]\\&=C\lim_{M\to\infty}\int_{-M}^{M}f(x)\,\frac{x}{\sigma^2}e^{-\frac{1}{2}x^2}dx=\frac{1}{\sigma^2}\mathbb{E}\left[Xf(X)\right].\end{aligned}$$

■

Example D.4. Suppose that $X\overset{d}{=}N(0,1)$ and define $\alpha_k:=\mathbb{E}\left[X^{2k}\right]$ for all $k\in\mathbb{N}_0$. By Lemma D.3,

$$\alpha_{k+1}=\mathbb{E}\left[X^{2k+1}\cdot X\right]=(2k+1)\,\alpha_k \text{ with } \alpha_0=1.$$

Hence it follows that

$$\alpha_1=\alpha_0=1,\ \alpha_2=3\alpha_1=3,\ \alpha_3=5\cdot 3$$

and by a simple induction argument,

$$\mathbb{E}X^{2k}=\alpha_k=(2k-1)!!,$$

where $(-1)!!:=0$.

Actually we can use the Γ – function to say more. Namely for any $\beta>-1$,

$$\mathbb{E}\left|X\right|^{\beta}=\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}}|x|^{\beta}\,e^{-\frac{1}{2}x^2}dx=\sqrt{\frac{2}{\pi}}\int_0^{\infty}x^{\beta}e^{-\frac{1}{2}x^2}dx.$$

Now make the change of variables, $y=x^2/2$ (i.e. $x=\sqrt{2y}$ and $dx=\frac{1}{\sqrt{2}}y^{-1/2}dy$) to learn,

$$\begin{aligned}\mathbb{E}\left|X\right|^{\beta}&=\frac{1}{\sqrt{\pi}}\int_0^{\infty}(2y)^{\beta/2}\,e^{-y}y^{-1/2}dy\\&=\frac{1}{\sqrt{\pi}}2^{\beta/2}\int_0^{\infty}y^{(\beta+1)/2}e^{-y}y^{-1}dy=\frac{1}{\sqrt{\pi}}2^{\beta/2}\Gamma\left(\frac{\beta+1}{2}\right).\end{aligned}$$

[1] This last hypothesis is actually unnecessary!

Exercise D.1. Let $q(x)$ be a polynomial[2] in x, $Z\overset{d}{=}N(0,1)$, and

$$u(t,x):=\mathbb{E}\left[q\left(x+\sqrt{t}Z\right)\right] \tag{D.4}$$

$$=\int_{\mathbb{R}}\frac{1}{\sqrt{2\pi t}}e^{-\frac{1}{2t}(y-x)^2}q(y)\,dy \tag{D.5}$$

Show u satisfies the heat equation,

$$\frac{\partial}{\partial t}u(t,x)=\frac{1}{2}\frac{\partial^2}{\partial x^2}u(t,x)\ \text{ for all } t>0 \text{ and } x\in\mathbb{R},$$
$$\text{with } u(0,x)=q(x).$$

Hints: Make use of Lemma D.3 along with the fact (which is easily proved here) that

$$\frac{\partial}{\partial t}u(t,x)=\mathbb{E}\left[\frac{\partial}{\partial t}q\left(x+\sqrt{t}Z\right)\right].$$

You will also have to use the corresponding fact for the x derivatives as well.

Exercise D.2. Let $q(x)$ be a polynomial in x, $Z\overset{d}{=}N(0,1)$, and $\Delta=\frac{d^2}{dx^2}$. Show

$$\mathbb{E}\left[q(Z)\right]=\left(e^{\Delta/2}q\right)(0):=\sum_{n=0}^{\infty}\frac{1}{n!}\left(\left(\frac{\Delta}{2}\right)^n q\right)(0)=\sum_{n=0}^{\infty}\frac{1}{n!}\frac{1}{2^n}\left(\Delta^n q\right)(0)$$

where the above sum is actually a finite sum since $\Delta^n q\equiv 0$ if $2n>\deg q$.
Hint: let $u(t):=\mathbb{E}\left[q\left(\sqrt{t}Z\right)\right]$. From your proof of Exercise D.1 you should be able to see that $\dot{u}(t)=\frac{1}{2}\mathbb{E}\left[(\Delta q)\left(\sqrt{t}Z\right)\right]$. This latter equation may be iterated in order to find $u^{(n)}(t)$ for all $n\geq 0$. With this information in hand you should be able to finish the proof with the aid of Taylor's theorem.

Example D.5. Suppose that $k\in\mathbb{N}$, then

$$\begin{aligned}\mathbb{E}\left[Z^{2k}\right]&=\left(e^{\frac{\Delta}{2}}x^{2k}\right)|_{x=0}=\sum_{n=0}^{\infty}\frac{1}{n!}\frac{1}{2^n}\left(\Delta^n x^{2k}\right)|_{x=0}\\&=\frac{1}{k!}\frac{1}{2^k}\Delta^k x^{2k}=\frac{(2k)!}{k!2^k}\\&=\frac{2k\cdot(2k-1)\cdot 2\,(k-1)\cdot(2k-3)\cdot\cdots\cdot(2\cdot 2)\cdot 3\cdot 2\cdot 1}{2^k k!}\\&=(2k-1)!!\end{aligned}$$

in agreement with Example D.4.

[2] Actually, $q(x)$ can be any twice continuously differentiable function which along with its derivatives grow slower than $e^{\varepsilon x^2}$ for any $\varepsilon>0$.

Example D.6. Let Z be a standard normal random variable and set $f(\lambda) := \mathbb{E}\left[e^{\lambda Z^2}\right]$ for $\lambda < 1/2$. Then $f(0) = 1$ and

$$\begin{aligned} f'(\lambda) &= \mathbb{E}\left[Z^2 e^{\lambda Z^2}\right] = \mathbb{E}\left[\frac{\partial}{\partial Z}\left(Z e^{\lambda Z^2}\right)\right] \\ &= \mathbb{E}\left[e^{\lambda Z^2} + 2\lambda Z^2 e^{\lambda Z^2}\right] \\ &= f(\lambda) + 2\lambda f'(\lambda). \end{aligned}$$

Solving for λ we find,

$$f'(\lambda) = \frac{1}{1-2\lambda} f(\lambda) \text{ with } f(0) = 1.$$

The solution to this equation is found in the usual way as,

$$\ln f(\lambda) = \int \frac{f'(\lambda)}{f(\lambda)} d\lambda = \int \frac{1}{1-2\lambda} d\lambda = -\frac{1}{2}\ln(1-2\lambda) + C.$$

By taking $\lambda = 0$ using $f(0) = 1$ we find that $C = 0$ and therefore,

$$\mathbb{E}\left[e^{\lambda Z^2}\right] = f(\lambda) = \frac{1}{\sqrt{1-2\lambda}} \text{ for } \lambda < \frac{1}{2}.$$

This can also be shown by directly evaluating the integral,

$$\mathbb{E}\left[e^{\lambda Z^2}\right] = \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(1-2\lambda)z^2} dz.$$

Exercise D.3. Suppose that $Z \overset{d}{=} N(0,1)$ and $\lambda \in \mathbb{R}$. Show

$$f(\lambda) := \mathbb{E}\left[e^{i\lambda Z}\right] = \exp\left(-\lambda^2/2\right). \tag{D.6}$$

Hint: You may use without proof that $f'(\lambda) = i\mathbb{E}\left[Z e^{i\lambda Z}\right]$ (i.e. it is permissible to differentiate past the expectation.) Assuming this use Lemma D.3 to see that $f'(\lambda)$ satisfies a simple ordinary differential equation.

Lemma D.7 (Gaussian tail estimates). *Suppose that X is a standard normal random variable, i.e.*

$$P(X \in A) = \frac{1}{\sqrt{2\pi}} \int_A e^{-x^2/2} dx \text{ for all } A \in \mathcal{B}_{\mathbb{R}},$$

then for all $x \geq 0$,

$$P(X \geq x) \leq \min\left(\frac{1}{2} - \frac{x}{\sqrt{2\pi}} e^{-x^2/2}, \frac{1}{\sqrt{2\pi}x} e^{-x^2/2}\right) \leq \frac{1}{2} e^{-x^2/2}. \tag{D.7}$$

Moreover (see [8, Lemma 2.5]),

$$P(X \geq x) \geq \max\left(1 - \frac{x}{\sqrt{2\pi}}, \frac{x}{x^2+1}\frac{1}{\sqrt{2\pi}} e^{-x^2/2}\right) \tag{D.8}$$

which combined with Eq. (D.7) proves Mill's ratio (see [3]);

$$\lim_{x\to\infty} \frac{P(X \geq x)}{\frac{1}{\sqrt{2\pi}x} e^{-x^2/2}} = 1. \tag{D.9}$$

Proof. See Figure D.1 where; the green curve is the plot of $P(X \geq x)$, the black is the plot of

$$\min\left(\frac{1}{2} - \frac{1}{\sqrt{2\pi}x} e^{-x^2/2}, \frac{1}{\sqrt{2\pi}x} e^{-x^2/2}\right),$$

the red is the plot of $\frac{1}{2} e^{-x^2/2}$, and the blue is the plot of

$$\max\left(\frac{1}{2} - \frac{x}{\sqrt{2\pi}}, \frac{x}{x^2+1}\frac{1}{\sqrt{2\pi}} e^{-x^2/2}\right).$$

The formal proof of these estimates for the reader who is not convinced by Figure D.1 is given below.

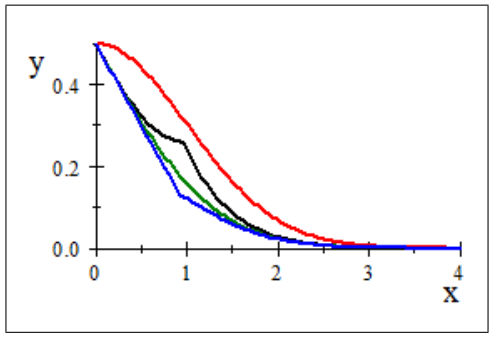


Fig. D.1. Plots of $P(X \geq x)$ and its estimates.

We begin by observing that

$$\begin{aligned} P(X \geq x) &= \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-y^2/2} dy \leq \frac{1}{\sqrt{2\pi}} \int_x^\infty \frac{y}{x} e^{-y^2/2} dy \\ &\leq -\frac{1}{\sqrt{2\pi}} \frac{1}{x} e^{-y^2/2}\Big|_x^\infty = \frac{1}{\sqrt{2\pi}} \frac{1}{x} e^{-x^2/2}. \end{aligned} \tag{D.10}$$

If we only want to prove Mill's ratio (D.9), we could proceed as follows. Let $\alpha > 1$, then for $x > 0$,

$$\begin{aligned} P(X \geq x) &= \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-y^2/2} dy \\ &\geq \frac{1}{\sqrt{2\pi}} \int_x^{\alpha x} \frac{y}{\alpha x} e^{-y^2/2} dy = -\frac{1}{\sqrt{2\pi}} \frac{1}{\alpha x} e^{-y^2/2} \Big|_{y=x}^{y=\alpha x} \\ &= \frac{1}{\sqrt{2\pi}} \frac{1}{\alpha x} e^{-x^2/2} \left[1 - e^{-\alpha^2 x^2/2}\right] \end{aligned}$$

from which it follows,

$$\liminf_{x\to\infty} \left[\sqrt{2\pi} x e^{x^2/2} \cdot P(X \geq x)\right] \geq 1/\alpha \uparrow 1 \text{ as } \alpha \downarrow 1.$$

The estimate in Eq. (D.10) shows $\limsup_{x\to\infty} \left[\sqrt{2\pi} x e^{x^2/2} \cdot P(X \geq x)\right] \leq 1$.

To get more precise estimates, we begin by observing,

$$\begin{aligned} P(X \geq x) &= \frac{1}{2} - \frac{1}{\sqrt{2\pi}} \int_0^x e^{-y^2/2} dy \\ &\leq \frac{1}{2} - \frac{1}{\sqrt{2\pi}} \int_0^x e^{-x^2/2} dy \leq \frac{1}{2} - \frac{1}{\sqrt{2\pi}} e^{-x^2/2} x. \end{aligned} \tag{D.11}$$

This equation along with Eq. (D.10) gives the first equality in Eq. (D.7). To prove the second equality observe that $\sqrt{2\pi} > 2$, so

$$\frac{1}{\sqrt{2\pi}} \frac{1}{x} e^{-x^2/2} \leq \frac{1}{2} e^{-x^2/2} \text{ if } x \geq 1.$$

For $x \leq 1$ we must show,

$$\frac{1}{2} - \frac{x}{\sqrt{2\pi}} e^{-x^2/2} \leq \frac{1}{2} e^{-x^2/2}$$

or equivalently that $f(x) := e^{x^2/2} - \sqrt{\frac{2}{\pi}} x \leq 1$ for $0 \leq x \leq 1$. Since f is convex $\left(f''(x) = \left(x^2+1\right) e^{x^2/2} > 0\right)$, $f(0) = 1$ and $f(1) \cong 0.85 < 1$, it follows that $f \leq 1$ on $[0,1]$. This proves the second inequality in Eq. (D.7).

It follows from Eq. (D.11) that

$$\begin{aligned} P(X \geq x) &= \frac{1}{2} - \frac{1}{\sqrt{2\pi}} \int_0^x e^{-y^2/2} dy \\ &\geq \frac{1}{2} - \frac{1}{\sqrt{2\pi}} \int_0^x 1 dy = \frac{1}{2} - \frac{1}{\sqrt{2\pi}} x \text{ for all } x \geq 0. \end{aligned}$$

So to finish the proof of Eq. (D.8) we must show,

$$\begin{aligned} f(x) &:= \frac{1}{\sqrt{2\pi}} x e^{-x^2/2} - \left(1+x^2\right) P(X \geq x) \\ &= \frac{1}{\sqrt{2\pi}} \left[x e^{-x^2/2} - \left(1+x^2\right) \int_x^\infty e^{-y^2/2} dy\right] \leq 0 \text{ for all } 0 \leq x < \infty. \end{aligned}$$

This follows by observing that $f(0) = -1/2 < 0$, $\lim_{x\uparrow\infty} f(x) = 0$ and

$$\begin{aligned} f'(x) &= \frac{1}{\sqrt{2\pi}} \left[e^{-x^2/2}\left(1-x^2\right) - 2xP(X \geq x) + \left(1+x^2\right) e^{-x^2/2}\right] \\ &= 2\left(\frac{1}{\sqrt{2\pi}} e^{-x^2/2} - xP(X \geq y)\right) \geq 0, \end{aligned}$$

where the last inequality is a consequence Eq. (D.7). ∎

D.2 Gaussian Random Vectors

Definition D.8 (Gaussian Random Vectors). *A random vector, $X \in \mathbb{R}^d$, is Gaussian iff*

$$\mathbb{E}\left[e^{i\lambda\cdot X}\right] = \exp\left(-\frac{1}{2}\operatorname{Var}(\lambda \cdot X) + i\mathbb{E}(\lambda \cdot X)\right) \text{ for all } \lambda \in \mathbb{R}^d. \tag{D.12}$$

In short, X is a Gaussian random vector iff $\lambda \cdot X$ is a Gaussian random variable for all $\lambda \in \mathbb{R}^d$. (Implicitly in this definition we are assuming that $\mathbb{E}\left|X_j^2\right| < \infty$ for $1 \leq j \leq d$.)

Notation D.9 *Let X be a random vector in $\mathbb{R}^d$ with second moments, i.e. $\mathbb{E}\left[X_k^2\right] < \infty$ for $1 \leq k \leq d$. The mean X is the vector $\mu = (\mu_1, \ldots, \mu_d)^{\mathrm{tr}} \in \mathbb{R}^d$ with $\mu_k := \mathbb{E}X_k$ for $1 \leq k \leq d$ and the covariance matrix $C = C(X)$ is the $d \times d$ matrix with entries,*

$$C_{kl} := \operatorname{Cov}(X_k, X_l) \text{ for } 1 \leq k, l \leq d. \tag{D.13}$$

Exercise D.4. Suppose that X is a random vector in $\mathbb{R}^d$ with second moments. Show for all $\lambda = (\lambda_1, \ldots, \lambda_d)^{\mathrm{tr}} \in \mathbb{R}^d$ that

$$\mathbb{E}[\lambda \cdot X] = \lambda \cdot \mu \text{ and } \operatorname{Var}(\lambda \cdot X) = \lambda \cdot C\lambda. \tag{D.14}$$

Corollary D.10. *If $Y \overset{d}{=} N\left(\mu, \sigma^2\right)$, then*

$$\mathbb{E}\left[e^{i\lambda Y}\right] = \exp\left(-\frac{1}{2}\lambda^2\sigma^2 + i\mu\lambda\right) \text{ for all } \lambda \in \mathbb{R}. \tag{D.15}$$

Conversely if Y is a random variable such that Eq. (D.15) holds, then $Y \overset{d}{=} N\left(\mu, \sigma^2\right)$.

Proof. ($\Longrightarrow$) From the remarks after Lemma D.2, we know that $Y \overset{d}{=} \sigma Z + \mu$ where $Z \overset{d}{=} N(0,1)$. Therefore,

$$\mathbb{E}\left[e^{i\lambda Y}\right] = \mathbb{E}\left[e^{i\lambda(\sigma Z+\mu)}\right] = e^{i\lambda\mu}\mathbb{E}\left[e^{i\lambda\sigma Z}\right] = e^{i\lambda\mu}e^{-\frac{1}{2}(\lambda\sigma)^2} = \exp\left(-\frac{1}{2}\lambda^2\sigma^2 + i\mu\lambda\right).$$

($\Longleftarrow$) This follows from the basic fact that the characteristic function or Fourier transform of a distribution uniquely determines the distribution. ∎

Remark D.11 (Alternate characterization of being Gaussian). Given Corollary D.10, we have Y is a Gaussian random variable iff $\mathbb{E}Y^2 < \infty$ and

$$\begin{aligned}\mathbb{E}\left[e^{i\lambda Y}\right] &= \exp\left(-\frac{1}{2}\operatorname{Var}(\lambda Y) + i\lambda\mathbb{E}Y\right)\\ &= \exp\left(-\frac{\lambda^2}{2}\operatorname{Var}(Y) + i\lambda\mathbb{E}Y\right) \text{ for all } \lambda\in\mathbb{R}.\end{aligned}$$

Exercise D.5. Suppose X_1 and X_2 are two independent Gaussian random variables with $X_i \overset{d}{=} N\left(0,\sigma_i^2\right)$ for $i = 1,2$. Show $X_1 + X_2$ is Gaussian and $X_1 + X_2 \overset{d}{=} N\left(0,\sigma_1^2+\sigma_2^2\right)$. (**Hint:** use Remark D.11.)

Exercise D.6. Suppose that $Z \overset{d}{=} N(0,1)$ and $t \in \mathbb{R}$. Show $\mathbb{E}\left[e^{tZ}\right] = \exp\left(t^2/2\right)$. (You could follow the hint in Exercise D.3 or you could use a completion of the squares argument along with the translation invariance of Lebesgue measure.)

Exercise D.7. Use Exercise D.6 to give another proof that $\mathbb{E}Z^{2k} = (2k-1)!!$ when $Z \overset{d}{=} N(0,1)$.

Exercise D.8. Let $Z \overset{d}{=} N(0,1)$ and $\alpha \in \mathbb{R}$, find $\rho : \mathbb{R}_+ \to \mathbb{R}_+ := (0,\infty)$ such that

$$\mathbb{E}\left[f\left(|Z|^\alpha\right)\right] = \int_{\mathbb{R}_+} f(x)\,\rho(x)\,dx$$

for all continuous functions, $f : \mathbb{R}_+ \to \mathbb{R}$ with compact support in $\mathbb{R}_+$.

In particular a random vector (X) in $\mathbb{R}^d$ with second moments a Gaussian random vector iff

$$\mathbb{E}\left[e^{i\lambda\cdot X}\right] = \exp\left(-\frac{1}{2}C\lambda\cdot\lambda + i\mu\cdot\lambda\right) \text{ for all } \lambda\in\mathbb{R}^d. \tag{D.16}$$

We abbreviate Eq. (D.16) by writing $X \overset{d}{=} N(\mu, C)$. Notice that it follows from Eq. (D.13) that $C^{\mathrm{tr}} = C$ and from Eq. (D.14) that $C \geq 0$, i.e. $\lambda\cdot C\lambda \geq 0$ for all $\lambda\in\mathbb{R}^d$.

Definition D.12. *Given a Gaussian random vector, X, we call the pair, (C,μ) appearing in Eq. (D.16) the* ***characteristics*** *of X.*

Lemma D.13. *Suppose that $X = \sum_{l=1}^k Z_l v_l + \mu$ where $\{Z_l\}_{l=1}^k$ are i.i.d. standard normal random variables, $\mu\in\mathbb{R}^d$ and $v_l\in\mathbb{R}^d$ for $1\leq l\leq k$. Then $X \overset{d}{=} N(\mu,C)$ where $C = \sum_{l=1}^k v_l v_l^{\mathrm{tr}}$.*

Proof. Using the basic properties of independence and normal random variables we find

$$\begin{aligned}\mathbb{E}\left[e^{i\lambda\cdot X}\right] &= \mathbb{E}\left[e^{i\sum_{l=1}^k Z_l\lambda\cdot v_l + i\lambda\cdot\mu}\right] = e^{i\lambda\cdot\mu}\prod_{l=1}^k\mathbb{E}\left[e^{iZ_l\lambda\cdot v_l}\right] = e^{i\lambda\cdot\mu}\prod_{l=1}^k e^{-\frac{1}{2}(\lambda\cdot v_l)^2}\\ &= \exp\left(-\frac{1}{2}\sum_{l=1}^k(\lambda\cdot v_l)^2 + i\lambda\cdot\mu\right).\end{aligned}$$

Since

$$\sum_{l=1}^k(\lambda\cdot v_l)^2 = \sum_{l=1}^k\lambda\cdot v_l\left(v_l^{\mathrm{tr}}\lambda\right) = \lambda\cdot\left(\sum_{l=1}^k v_l v_l^{\mathrm{tr}}\right)\lambda$$

we may conclude,

$$\mathbb{E}\left[e^{i\lambda\cdot X}\right] = \exp\left(-\frac{1}{2}C\lambda\cdot\lambda + i\lambda\cdot\mu\right),$$

i.e. $X \overset{d}{=} N(\mu,C)$. ∎

Exercise D.9 (Existence of Gaussian random vectors for all $C\geq 0$ and $\mu\in\mathbb{R}^d$). Suppose that $\mu\in\mathbb{R}^d$ and C is a symmetric non-negative $d\times d$ matrix. By the spectral theorem we know there is an orthonormal basis $\{u_j\}_{j=1}^d$ for $\mathbb{R}^d$ such that $Cu_j = \sigma_j^2 u_j$ for some $\sigma_j^2\geq 0$. Let $\{Z_j\}_{j=1}^d$ be i.i.d. standard normal random variables, show $X := \sum_{j=1}^d Z_j\sigma_j u_j + \mu \overset{d}{=} N(\mu,C)$.

Theorem D.14 (Gaussian Densities). *Suppose that $X \overset{d}{=} N(\mu,C)$ is an $\mathbb{R}^d$ – valued Gaussian random vector with $C>0$ (for simplicity). Then*

$$\mathbb{E}\left[f(X)\right] = \frac{1}{\sqrt{\det(2\pi C)}}\int_{\mathbb{R}^d} f(x)\exp\left(-\frac{1}{2}C^{-1}(x-\mu)\cdot(x-\mu)\right)dx \tag{D.17}$$

for bounded or non-negative functions, $f : \mathbb{R}^d\to\mathbb{R}$.

Proof. Let us continue the notation in Exercise D.9 and further let

$$A := [\sigma_1 u_1|\ldots|\sigma_n u_n] = U\Sigma \tag{D.18}$$

where

$$U = [u_1|\dots|u_n] \text{ and } \Sigma = \operatorname{diag}(\sigma_1,\dots,\sigma_n) = [\sigma_1 e_1|\dots|\sigma_n e_n],$$

where $\{e_i\}_{i=1}^d$ is the standard orthonormal basis for $\mathbb{R}^d$. With this notation we know that $X \overset{d}{=} AZ+\mu$ where $Z = (Z_1,\dots,Z_d)^{\mathrm{tr}}$ is a standard normal Gaussian vector. Therefore,

$$\mathbb{E}[f(X)] = (2\pi)^{-d/2}\int_{\mathbb{R}^d} f(Az+\mu)\, e^{-\frac{1}{2}\|z\|^2} dz \tag{D.19}$$

wherein we have used

$$\prod_{j=1}^d \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z_i^2} = (2\pi)^{-d/2} e^{-\frac{1}{2}\|z\|^2} \text{ with } \|z\|^2 := \sum_{j=1}^d z_j^2.$$

Making the change of variables $x = Az+\mu$ in Eq. (D.19) (i.e. $z = A^{-1}(x-\mu)$ and $dz = dx/\det A$) implies

$$\mathbb{E}[f(X)] = \frac{1}{(2\pi)^{d/2}\det A}\int_{\mathbb{R}^d} f(x)\, e^{-\frac{1}{2}\|A^{-1}(x-\mu)\|^2} dx. \tag{D.20}$$

Recall from your linear algebra class (or just check) that $CU = U\Sigma^2$, i.e. $C = U\Sigma^2 U^{-1} = U\Sigma^2 U^{\mathrm{tr}}$. Therefore[3],

$$AA^{\mathrm{tr}} = U\Sigma\Sigma U^{\mathrm{tr}} = U\Sigma^2 U^{-1} = C \tag{D.21}$$

which then implies $\det A = \sqrt{\det C}$ and for all $y \in \mathbb{R}^d$

$$\left\|A^{-1}y\right\|^2 = \left(A^{-1}\right)^{\mathrm{tr}} A^{-1}y \cdot y = \left(AA^{\mathrm{tr}}\right)^{-1} y \cdot y = C^{-1}y \cdot y.$$

Equation (D.17) follows from theses observations, Eq. (D.20), and the identity;

$$(2\pi)^{d/2}\det A = (2\pi)^{d/2}\sqrt{\det C} = \sqrt{(2\pi)^d \det C} = \sqrt{\det(2\pi C)}.$$

■

[3] Alternatively,

$$\begin{aligned} C_{ik} &= \operatorname{Cov}(X_i, X_k) = \operatorname{Cov}(Y_i, Y_k) \\ &= \sum_{j,m} \operatorname{Cov}(A_{ij}Z_j, A_{km}Z_m) = \sum_{j,m} A_{ij}A_{km}\operatorname{Cov}(Z_j, Z_m) \\ &= \sum_{j,m} A_{ij}A_{km}\delta_{jm} = \sum_j A_{ij}A_{kj} = \left(AA^{\mathrm{tr}}\right)_{ik}. \end{aligned}$$

Theorem D.15 (Gaussian Integration by Parts). *Suppose that $X = (X_1,\dots,X_d)$ is a mean zero Gaussian random vector and $C_{ij} = \operatorname{Cov}(X_i, Y_j) = \mathbb{E}[X_i X_j]$. Then for any smooth function $f : \mathbb{R}^d \to \mathbb{R}$ such that f and all its derivatives grows slower that $\exp(|x|^\alpha)$ for some $\alpha < 2$, we have*

$$\begin{aligned} \mathbb{E}[X_i f(X_1,\dots,X_d)] &= \sum_{k=1}^d C_{ik}\mathbb{E}\left[\frac{\partial}{\partial X_k} f(X_1,\dots,X_d)\right] \\ &= \sum_{k=1}^d \mathbb{E}[X_i X_k]\cdot\mathbb{E}\left[\frac{\partial}{\partial X_k} f(X_1,\dots,X_d)\right]. \end{aligned}$$

Here we write $\frac{\partial}{\partial X_k} f(X_1,\dots,X_d)$ for $(\partial_k f)(X_1,\dots,X_d)$ where

$$(\partial_k f)(x_1,\dots,x_d) := \frac{\partial}{\partial x_k} f(x_1,\dots,x_d) = \frac{d}{dt}|_0 f(\mathbf{x}+t\mathbf{e}_k)$$

where $\mathbf{x} := (x_1,\dots,x_d)$ and $\mathbf{e}_k$ is the k^{th} – standard basis vector for $\mathbb{R}^d$.

Proof. From Exercise D.9 we know $X \overset{d}{=} Y$ where $Y = \sum_{j=1}^d \sigma_j Z_j u_j$ where $\{u_j\}_{j=1}^d$ is an orthonormal basis for $\mathbb{R}^d$ such that $Cu_j = \sigma_j^2 u_j$ and $\{Z_j\}_{j=1}^d$ are i.i.d. standard normal random variables. To simplify notation we define $A := [\sigma_1 u_1|\dots|\sigma_d u_d]$ as in Eq. (D.18) so that $Y = AZ$ where $Z = (Z_1,\dots,Z_d)^{\mathrm{tr}}$ as in the proof of Theorem D.14. From our previous observations and a simple generalization of Lemma D.3, it follows that

$$\begin{aligned} \mathbb{E}[X_i f(X_1,\dots,X_d)] &= \mathbb{E}[Y_i f(Y_1,\dots,Y_d)] \\ &= \sum_j A_{ij}\mathbb{E}[Z_j f((AZ)_1,\dots(AZ_d))] \\ &= \sum_j A_{ij}\mathbb{E}\left[\frac{\partial}{\partial Z_j} f((AZ)_1,\dots(AZ_d))\right] \\ &= \sum_j A_{ij}\mathbb{E}\left[\sum_k (\partial_k f)((AZ)_1,\dots(AZ_d))\cdot\frac{\partial}{\partial Z_j}(AZ)_k\right] \\ &= \sum_{j,k} A_{ij}A_{kj}\mathbb{E}[(\partial_k f)(X_1,\dots,X_d)]. \end{aligned}$$

This completes the proof since, $\sum_j A_{ij}A_{kj} = \left(AA^{\mathrm{tr}}\right)_{ik} = C_{ik}$ as we saw in Eq. (D.21). ■

Theorem D.16 (Wick's Theorem). *If $X = (X_1,\dots,X_{2n})$ is a mean zero Gaussian random vector, then*

$$\mathbb{E}[X_1 \dots X_{2n}] = \sum_{pairings} C_{i_1 j_1} \dots C_{i_n j_n}$$

where the sum is over all perfect pairings of $\{1, 2, \dots, 2n\}$ *and*

$$C_{ij} = \operatorname{Cov}(X_i, X_j) = \mathbb{E}[X_i X_j].$$

Proof. From Theorem D.15,

$$\mathbb{E}[X_1 \dots X_{2n}] = \sum_j C_{1j} \mathbb{E}\left[\frac{\partial}{\partial X_j} X_2 \dots X_{2n}\right] = \sum_{j>2} C_{1j} \mathbb{E}\left[X_2 \dots \hat{X}_j \dots X_{2n}\right]$$

where the hat indicates a term to be omitted. The result now basically follows by induction. For example,

$$\begin{aligned}
\mathbb{E}[X_1 X_2 X_3 X_4] =& C_{12} \mathbb{E}\left[\frac{\partial}{\partial X_2}(X_2 X_3 X_4)\right] \\
&+ C_{13} \mathbb{E}\left[\frac{\partial}{\partial X_3}(X_2 X_3 X_4)\right] + C_{14} \mathbb{E}\left[\frac{\partial}{\partial X_4}(X_2 X_3 X_4)\right] \\
=& C_{12} \mathbb{E}[X_3 X_4] + C_{13} \mathbb{E}[X_2 X_4] + C_{14} \mathbb{E}[X_2 X_3] \\
=& C_{12} C_{34} + C_{13} C_{24} + C_{14} C_{23}.
\end{aligned}$$

■

Recall that if X_i and Y_j are independent, then $\operatorname{Cov}(X_i, Y_j) = 0$, i.e. independence implies uncorrelated. On the other hand, typically uncorrelated random variables are not independent. However, if the random variables involved are jointly Gaussian, then independence and uncorrelated are actually the same thing!

Lemma D.17. *Suppose that* $Z = (X, Y)^{\mathrm{tr}}$ *is a Gaussian random vector with* $X \in \mathbb{R}^k$ *and* $Y \in \mathbb{R}^l$. *Then* X *is independent of* Y *iff* $\operatorname{Cov}(X_i, Y_j) = 0$ *for all* $1 \le i \le k$ *and* $1 \le j \le l$.

Remark D.18. Lemma D.17 also holds more generally. Namely if $\{X^l\}_{l=1}^n$ is a sequence of random vectors such that $(X^1, \dots, X^n)$ is a Gaussian random vector. Then $\{X^l\}_{l=1}^n$ are independent iff $\operatorname{Cov}\left(X_i^l, X_k^{l'}\right) = 0$ for all $l \neq l'$ and i and k.

Exercise D.10. Prove Lemma D.17. **Hint:** by basic facts about the Fourier transform, it suffices to prove

$$\mathbb{E}\left[e^{ix\cdot X} e^{iy\cdot Y}\right] = \mathbb{E}\left[e^{ix\cdot X}\right] \cdot \mathbb{E}\left[e^{iy\cdot Y}\right] \text{ for all } x \in \mathbb{R}^k \text{ and } y \in \mathbb{R}^l.$$

If you get stuck, take a look at the proof of Corollary D.19 below.

Corollary D.19. *Suppose that* $X \in \mathbb{R}^k$ *and* $Y \in \mathbb{R}^l$ *are two independent random Gaussian vectors, then* (X, Y) *is also a Gaussian random vector. This corollary generalizes to multiple independent random Gaussian vectors.*

Proof. Let $x \in \mathbb{R}^k$ and $y \in \mathbb{R}^l$, then

$$\begin{aligned}
\mathbb{E}\left[e^{i(x,y)\cdot(X,Y)}\right] =& \mathbb{E}\left[e^{i(x\cdot X + y\cdot Y)}\right] = \mathbb{E}\left[e^{ix\cdot X} e^{iy\cdot Y}\right] = \mathbb{E}\left[e^{ix\cdot X}\right] \cdot \mathbb{E}\left[e^{iy\cdot Y}\right] \\
=& \exp\left(-\frac{1}{2}\operatorname{Var}(x\cdot X) + i\mathbb{E}(x\cdot X)\right) \\
&\times \exp\left(-\frac{1}{2}\operatorname{Var}(y\cdot Y) + i\mathbb{E}(y\cdot Y)\right) \\
=& \exp\left(-\frac{1}{2}\operatorname{Var}(x\cdot X) + i\mathbb{E}(x\cdot X) - \frac{1}{2}\operatorname{Var}(y\cdot Y) + i\mathbb{E}(y\cdot Y)\right) \\
=& \exp\left(-\frac{1}{2}\operatorname{Var}(x\cdot X + y\cdot Y) + i\mathbb{E}(x\cdot X + y\cdot Y)\right)
\end{aligned}$$

which shows that (X, Y) is again Gaussian. ■

Remark D.20 (Be careful). If X_1 and X_2 are two standard normal random variables, it is **not** generally true that (X_1, X_2) is a Gaussian random vector. For example suppose $X_1 \overset{d}{=} N(0,1)$ is a standard normal random variable and ε is an independent Bernoulli random variable with $\mathbb{P}(\varepsilon = \pm 1) = \frac{1}{2}$. Then $X_2 := \varepsilon X_1 \overset{d}{=} N(0,1)$ but $X := (X_1, X_2)$ is **not** a Gaussian random vector as we now verify.

If $\lambda = (\lambda_1, \lambda_2) \in \mathbb{R}^2$, then

$$\begin{aligned}
\mathbb{E}\left[e^{i\lambda\cdot X}\right] &= \mathbb{E}\left[e^{i(\lambda_1 X_1 + \lambda_2 X_2)}\right] = \mathbb{E}\left[e^{i(\lambda_1 X_1 + \lambda_2 \varepsilon X_1)}\right] \\
&= \frac{1}{2}\sum_{\tau=\pm 1} \mathbb{E}\left[e^{i(\lambda_1 X_1 + \lambda_2 \tau X_1)}\right] = \frac{1}{2}\sum_{\tau=\pm 1} \mathbb{E}\left[e^{i(\lambda_1 + \lambda_2\tau) X_1}\right] \\
&= \frac{1}{2}\sum_{\tau=\pm 1} \exp\left(-\frac{1}{2}(\lambda_1 + \lambda_2\tau)^2\right) \\
&= \frac{1}{2}\sum_{\tau=\pm 1} \exp\left(-\frac{1}{2}\left(\lambda_1^2 + \lambda_2^2 + 2\tau\lambda_1\lambda_2\right)\right) \\
&= \frac{1}{2} e^{-\frac{1}{2}(\lambda_1^2 + \lambda_2^2)} \cdot \left[\exp(-\lambda_1\lambda_2) + \exp(\lambda_1\lambda_2)\right] \\
&= e^{-\frac{1}{2}(\lambda_1^2 + \lambda_2^2)} \cosh(\lambda_1\lambda_2).
\end{aligned}$$

On the other hand, $\mathbb{E}\left[X_1^2\right] = \mathbb{E}\left[X_2^2\right] = 1$ and

$$\mathbb{E}[X_1 X_2] = \mathbb{E}\varepsilon \cdot \mathbb{E}\left[X_1^2\right] = 0 \cdot 1 = 0,$$

from which it follows that X_1 and X_2 are uncorrelated and $C^X = I_{2\times 2}$. Thus if X were Gaussian we would have,

$$\mathbb{E}\left[e^{i\lambda\cdot X}\right] = \exp\left(-\frac{1}{2}C^X\lambda\cdot\lambda\right) = e^{-\frac{1}{2}(\lambda_1^2+\lambda_2^2)}$$

which is just not the case!

Incidentally, this example also shows that two uncorrelated random variables need **not** be independent. For if $\{X_1, X_2\}$ were independent, then again we would have

$$\begin{aligned}\mathbb{E}\left[e^{i\lambda\cdot X}\right] &= \mathbb{E}\left[e^{i(\lambda_1X_1+\lambda_2X_2)}\right] = \mathbb{E}\left[e^{i\lambda_1X_1}e^{i\lambda_2X_2}\right]\\ &= \mathbb{E}\left[e^{i\lambda_1X_1}\right]\cdot\mathbb{E}\left[e^{i\lambda_2X_2}\right] = e^{-\frac{1}{2}\lambda_1^2}e^{-\frac{1}{2}\lambda_2^2} = e^{-\frac{1}{2}\left(\lambda_1^2+\lambda_2^2\right)},\end{aligned}$$

which is not the case.

The following theorem gives another useful way of computing Gaussian integrals of polynomials and exponential functions.

Theorem D.21. *Suppose $X \overset{d}{=} N(0,C)$ where C is a $N\times N$ symmetric positive definite matrix. Let $L = L^C := \sum_{i,j=1}^d C_{ij}\partial_i\partial_j$ (sum on repeated indices) where $\partial_i := \partial/\partial x_i$. Then for any polynomial function, $q:\mathbb{R}^N\to\mathbb{R}$,*

$$\mathbb{E}[q(X)] = \left(e^{\frac{1}{2}L}q\right)(0) := \sum_{n=0}^{\infty}\frac{1}{n!}\left(\left(\frac{L}{2}\right)^n q\right)(0) \quad \textit{(a finite sum).} \tag{D.22}$$

Proof. This is a fairly straight forward extension of Exercise D.2 and so I will only provide a short outline to the proof. 1) Let $u(t) := \mathbb{E}\left[q\left(\sqrt{t}X\right)\right]$. 2) Using Theorem D.15 one shows that $\dot{u}(t) = \frac{1}{2}\mathbb{E}\left[(Lq)\left(\sqrt{t}X\right)\right]$. 3) Iterating this result and then using Taylor's theorem finishes the proof just like in Exercise D.2. ∎

Corollary D.22. *The function $u(t,x) := \mathbb{E}\left[q\left(x+\sqrt{t}X\right)\right]$ solves the heat equation,*

$$\partial_t u(t,x) = \frac{1}{2}L^C u(t,x) \text{ with } u(0,x) = q(x).$$

If $X \overset{d}{=} N(1,0)$ we have

$$\begin{aligned}u(t,x) &= \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} q\left(x+\sqrt{t}z\right)e^{-\frac{1}{2}z^2}dz\\ &= \int_{\mathbb{E}} p_t(x,y)\,q(y)\,dy\end{aligned}$$

where

$$p_t(x,y) := \frac{1}{\sqrt{2\pi t}}\exp\left(-\frac{1}{2t}(y-x)^2\right).$$

D.3 Gaussian Conditioning

Notation D.23 *Let (X,Y) be a mean zero Gaussian random vector taking values in $\mathbb{R}^k\times\mathbb{R}^l$,*

$$C_X := \mathbb{E}\left[XX^{\mathrm{tr}}\right],\ C_Y := \mathbb{E}\left[YY^{\mathrm{tr}}\right],\ C_{X,Y} = \mathbb{E}\left[XY^{\mathrm{tr}}\right], \text{ and } C_{Y,X} = \mathbb{E}\left[YX^{\mathrm{tr}}\right]$$

so that C_X, C_Y, $C_{X,Y}$, and $C_{Y,X}$ are $k\times k$, $l\times l$, $k\times l$, and $l\times k$ matrices respectively. [Note that $C_{X,Y}^{\mathrm{tr}} = C_{Y,X}$ while $C_X^{\mathrm{tr}} = C_X$ and $C_Y^{\mathrm{tr}} = C_Y$.]

Definition D.24 (Pseudo-Inverses). *If C is a symmetric $k\times k$ matrix on $\mathbb{R}^k$, let C^{-1} be the $k\times k$ matrix uniquely determined by; $C^{-1}v = 0$ if $v\in\mathrm{Nul}(C)$ while if $v\in\mathrm{Nul}(C)^\perp = \mathrm{Ran}(C)$ we let $C^{-1}v = w$ where w is the unique element of $\mathrm{Ran}(C)$ such that $Cw = v$. [If C is invertible, then the pseudo inverse is the same as the inverse matrix.]*

Lemma D.25. *If X is a mean zero $\mathbb{R}^k$ – valued Gaussian random vector, then $X\in\mathrm{Ran}(C_X)$ a.s.*

Proof. If $\lambda\in\mathrm{Ran}(C_X)^\perp = \mathrm{Nul}(C_X^*) = \mathrm{Nul}(C_X)$, then

$$\mathbb{E}[\lambda\cdot X]^2 = C_X\lambda\cdot\lambda = 0 \implies \lambda\cdot X = 0 \text{ a.s.}.$$

Letting λ run through a basis for $\mathrm{Ran}(C_X)^\perp$, it then follows that $X\in\mathrm{Ran}(C_X)$ a.s. ∎

Lemma D.26. *If (X,Y) is a mean zero Gaussian random vector taking values in $\mathbb{R}^k\times\mathbb{R}^l$, then $Z := Y - C_{Y,X}C_X^{-1}X$ is independent of X, i.e.*

$$Y = C_{Y,X}C_X^{-1}X + Z \tag{D.23}$$

where Z is a mean zero Gaussian random vector independent of X such that

$$C_Z = C_Y - C_{Y,X}C_X^{-1}C_{X,Y}. \tag{D.24}$$

Proof. We look for a $l\times k$ matrix, A, so that $Z := Y - AX$ is independent of X. Since (X,Y) is Gaussian it suffices to find A so that

$$0 = \mathbb{E}\left[ZX^{\mathrm{tr}}\right] = \mathbb{E}\left[(Y-AX)X^{\mathrm{tr}}\right] = C_{Y,X} - AC_X,$$

i.e. we require $AC_X = C_{Y,X}$. This suggests that we let $A = C_{Y,X}C_X^{-1}$ but we must check this works even when C_X is not invertible. In this case

$$AC_X = C_{Y,X}C_X^{-1}C_X = C_{Y,X}P_X$$

where P_X is orthogonal projection onto $\mathrm{Ran}(C_X)$. The claim is that $C_{Y,X}P_X = C_{Y,X}$. The point is that if $\lambda\in\mathrm{Nul}(C_X)$, then

$$C_{Y,X}\lambda = \mathbb{E}\left[YX^{\mathrm{tr}}\lambda\right] = \mathbb{E}\left[(\lambda\cdot X)\,Y\right] = \mathbb{E}\left[0\right] = 0,$$

wherein we have used Lemma D.25 to conclude $\lambda \cdot X = 0$ a.s. Consequently, $C_{Y,X}$ vanishes on $\mathrm{Ran}\,(C_X)^{\perp}$ and hence $C_{Y,X}P_X = C_{Y,X}$.

So we have shown $Z := Y - C_{Y,X}C_X^{-1}X$ is independent of X. We now compute the covariance (C_Z) of Z;

$$\begin{aligned} C_Z &= \mathbb{E}\left[ZZ^{\mathrm{tr}}\right] = \mathbb{E}\left[Z\left[Y - C_{Y,X}C_X^{-1}X\right]^{\mathrm{tr}}\right] \\ &= \mathbb{E}\left[ZY^{\mathrm{tr}}\right] = \mathbb{E}\left[\left[Y - C_{Y,X}C_X^{-1}X\right]Y^{\mathrm{tr}}\right] \\ &= C_Y - C_{Y,X}C_X^{-1}C_{X,Y}, \end{aligned}$$

wherein the third equality we made use of the fact that Z and X were independent of (X,Z) is still jointly Gaussian. ∎

Theorem D.27 (Gaussian Conditioning). *Suppose that (X,Y) is a Gaussian vector taking values in $\mathbb{R}^k \times \mathbb{R}^l$. Then*

$$\mathbb{E}\left[f\left(Y\right)|X\right] = G\left(X\right) \tag{D.25}$$

where

$$G\left(x\right) := \mathbb{E}\left[f\left(C_{Y,X}C_X^{-1}x + Z\right)\right] \tag{D.26}$$

and Z is a $\mathbb{R}^l$ – mean zero Gaussian random vector with covariance matrix (C_Z) as in Eq. (D.24). As a special case, taking $f(y) = y$ above shows

$$\mathbb{E}\left[Y|X\right] = C_{Y,X}C_X^{-1}X.$$

Proof. This follows directly from Proposition **??** and the decomposition in Lemma D.26. ∎

Corollary D.28. *If (X,Y) is a Gaussian random vector taking values in $\mathbb{R}^k \times \mathbb{R}^l$ with $\mu_X = \mathbb{E}X$ and $\mu_Y = \mathbb{E}Y$, then*

$$Y = \mu_Y + C_{Y,X}C_X^{-1}\left(X - \mu_X\right) + Z$$

where now,

$$C_{Y,X} = \mathbb{E}\left[\left(Y - \mu_Y\right)\left(X - \mu_X\right)^{\mathrm{tr}}\right] = \mathbb{E}YX^{\mathrm{tr}} - \left(\mathbb{E}Y\right)\left(\mathbb{E}X\right)^{\mathrm{tr}}$$

with similar definitions for $C_{X,Y}$, C_X, and C_Y and Z is a mean zero Gaussian $\mathbb{R}^l$ – random vector independent of X with covariance matrix,

$$C_Z = C_Y - C_{Y,X}C_X^{-1}C_{X,Y}. \tag{D.27}$$

Moreover

$$\mathbb{E}\left[f\left(Y\right)|X\right] = G\left(X\right) \text{ where } G\left(x\right) := \mathbb{E}\left[f\left(C_{Y,X}C_X^{-1}x + \tilde{Z}\right)\right], \tag{D.28}$$

where $\tilde{Z}$ is a $\mathbb{R}^l$ - valued Gaussian random vector independent such that

$$\tilde{Z} = Z + \mu_Y - C_{Y,X}C_X^{-1}\mu_X \overset{d}{=} N\left(\mu_Y - C_{Y,X}C_X^{-1}\mu_X, C_Z\right). \tag{D.29}$$

Proof. Applying Lemma D.26 to $(X - \mu_X, Y - \mu_Y)$ shows,

$$Y - \mu_Y = C_{Y,X}C_X^{-1}\left(X - \mu_X\right) + Z \tag{D.30}$$

where the matrices $C_{Y,X}$ and C_X are now the covariances as described in the statement. The $\mathbb{R}^l$ – valued random vector, Z, is still a mean zero Gaussian vector with covariance given by Eq. (D.27) as described. We may rewrite Eq. (D.30) in the form $Y = C_{Y,X}C_X^{-1}X + \tilde{Z}$ where $\tilde{Z}$ is given as in Eq. (D.29). The formula for $\mathbb{E}\left[f\left(Y\right)|X\right]$ given in Eq. (D.28) follows from this decomposition along with Proposition **??**. ∎

D.4 Independent Random Variables

Definition D.29. *We say a collection of discrete random variables, $\{X_j\}_{j\in J}$, are **independent** if*

$$\mathbb{P}\left(X_{j_1} = x_1, \ldots, X_{j_n} = x_n\right) = \mathbb{P}\left(X_{j_1} = x_1\right)\cdots\mathbb{P}\left(X_{j_n} = x_n\right) \tag{D.31}$$

for all possible choices of $\{j_1, \ldots, j_n\} \subset J$ and all possible values x_k of X_{j_k}.

Proposition D.30. *A sequence of discrete random variables, $\{X_j\}_{j\in J}$, is independent iff*

$$\mathbb{E}\left[f_1\left(X_{j_1}\right)\ldots f_n\left(X_{j_n}\right)\right] = \mathbb{E}\left[f_1\left(X_{j_1}\right)\right]\ldots\mathbb{E}\left[f_n\left(X_{j_n}\right)\right] \tag{D.32}$$

for all choices of $\{j_1, \ldots, j_n\} \subset J$ and all choice of bounded (or non-negative) functions, $f_1, \ldots, f_n$. Here n is arbitrary.

Proof. ($\Longrightarrow$) If $\{X_j\}_{j\in J}$, are independent then

$$\begin{aligned} \mathbb{E}\left[f\left(X_{j_1}, \ldots, X_{j_n}\right)\right] &= \sum_{x_1,\ldots,x_n} f\left(x_1, \ldots, x_n\right)\mathbb{P}\left(X_{j_1} = x_1, \ldots, X_{j_n} = x_n\right) \\ &= \sum_{x_1,\ldots,x_n} f\left(x_1, \ldots, x_n\right)\mathbb{P}\left(X_{j_1} = x_1\right)\cdots\mathbb{P}\left(X_{j_n} = x_n\right). \end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{E}\left[f_1\left(X_{j_1}\right)\dots f_n\left(X_{j_n}\right)\right] &= \sum_{x_1,\dots,x_n} f_1\left(x_1\right)\dots f_n\left(x_n\right)\mathbb{P}\left(X_{j_1}=x_1\right)\cdots\mathbb{P}\left(X_{j_n}=x_n\right)\\ &= \left(\sum_{x_1} f_1\left(x_1\right)\mathbb{P}\left(X_{j_1}=x_1\right)\right)\cdots\left(\sum_{x_n} f\left(x_n\right)\mathbb{P}\left(X_{j_n}=x_n\right)\right)\\ &= \mathbb{E}\left[f_1\left(X_{j_1}\right)\right]\dots\mathbb{E}\left[f_n\left(X_{j_n}\right)\right].\end{aligned}$$

($\Longleftarrow$) Now suppose that Eq. (D.32) holds. If $f_j := \delta_{x_j}$ for all j, then

$$\mathbb{E}\left[f_1\left(X_{j_1}\right)\dots f_n\left(X_{j_n}\right)\right] = \mathbb{E}\left[\delta_{x_1}\left(X_{j_1}\right)\dots\delta_{x_n}\left(X_{j_n}\right)\right] = \mathbb{P}\left(X_{j_1}=x_1,\dots,X_{j_n}=x_n\right)$$

while

$$\mathbb{E}\left[f_k\left(X_{j_k}\right)\right] = \mathbb{E}\left[\delta_{x_k}\left(X_{j_k}\right)\right] = \mathbb{P}\left(X_{j_k}=x_k\right).$$

Therefore it follows from Eq. (D.32) that Eq. (D.31) holds, i.e. $\{X_j\}_{j\in J}$ is an independent collection of random variables. ∎

Using this as motivation we make the following definition.

Definition D.31. *A collection of arbitrary random variables, $\{X_j\}_{j\in J}$, are **independent** iff*

$$\mathbb{E}\left[f_1\left(X_{j_1}\right)\dots f_n\left(X_{j_n}\right)\right] = \mathbb{E}\left[f_1\left(X_{j_1}\right)\right]\dots\mathbb{E}\left[f_n\left(X_{j_n}\right)\right]$$

for all choices of $\{j_1,\dots,j_n\}\subset J$ and all choice of bounded (or non-negative) functions, $f_1,\dots,f_n$.

Fact D.32 *To check independence of a collection of real valued random variables, $\{X_j\}_{j\in J}$, it suffices to show*

$$\mathbb{P}\left(X_{j_1}\le t_1,\dots,X_{j_n}\le t_n\right) = \mathbb{P}\left(X_{j_1}\le t_1\right)\dots\mathbb{P}\left(X_{j_n}\le t_n\right)$$

for all possible choices of $\{j_1,\dots,j_n\}\subset J$ and all possible $t_k\in\mathbb{R}$. Moreover, one can replace $\le$ by $<$ or reverse these inequalities in the the above expression.

Theorem D.33 (Groupings of independent RVs). *If $\{X_j\}_{j\in J}$, are **independent** random variables and J_0, J_1 are finite disjoint subsets in J, then*

$$\mathbb{E}\left[f_0\left(\{X_j\}_{j\in J_0}\right)\cdot f_1\left(\{X_j\}_{j\in J_1}\right)\right] = \mathbb{E}\left[f_0\left(\{X_j\}_{j\in J_0}\right)\right]\cdot\mathbb{E}\left[f_1\left(\{X_j\}_{j\in J_1}\right)\right]. \tag{D.33}$$

This holds more generally for any $\{J_k\}_{k=0}^n\subset J$ with $J_k\cap J_l=\emptyset$ and $\#\left(J_k\right)<\infty$.

***In words;** disjoint groupings of independent random variables are still independent random vectors.*

Proof. Discrete case example. Suppose $\{X_1,\dots,X_5\}$ are independent discrete random variables. Then

$$\begin{aligned}&\mathbb{P}\left(X_1=s_1,X_2=s_2,X_3=s_3,X_4=s_4,X_5=s_5\right)\\ &\quad=\mathbb{P}\left(X_1=s_1\right)\mathbb{P}\left(X_2=s_2\right)\mathbb{P}\left(X_3=s_3\right)\mathbb{P}\left(X_4=s_4\right)\mathbb{P}\left(X_5=s_5\right)\\ &\quad=\mathbb{P}\left(X_1=s_1,X_2=s_2\right)\mathbb{P}\left(X_3=s_3,X_4=s_4,X_5=s_5\right)\\ &\quad=:\rho_{1,2}\left(s_1,s_2\right)\cdot\rho_{3,4,5}\left(s_3,s_4,s_5\right)\end{aligned}$$

and therefore,

$$\begin{aligned}&\mathbb{E}\left[f\left(X_1,X_2\right)g\left(X_3,X_4,X_5\right)\right]\\ &=\sum_{\mathbf{s}=(s_1,\dots,s_5)} f\left(s_1,s_2\right)g\left(s_3,s_4,s_5\right)\mathbb{P}\left(X_1=s_1,\dots,X_5=s_5\right)\\ &=\sum_{\mathbf{s}=(s_1,\dots,s_5)} f\left(s_1,s_2\right)g\left(s_3,s_4,s_5\right)\cdot\rho_{1,2}\left(s_1,s_2\right)\cdot\rho_{3,4,5}\left(s_3,s_4,s_5\right)\\ &=\sum_{\mathbf{s}=(s_1,s_2)} f\left(s_1,s_2\right)\rho_{1,2}\left(s_1,s_2\right)\cdot\sum_{\mathbf{s}=(s_3,s_4,s_5)} g\left(s_3,s_4,s_5\right)\rho_{3,4,5}\left(s_3,s_4,s_5\right)\\ &=\mathbb{E}\left[f\left(X_1,X_2\right)\right]\cdot\mathbb{E}\left[g\left(X_3,X_4,X_5\right)\right].\end{aligned}$$

General Case. Equation (D.33) is easy to verify when f_0 and f_1 are themselves product functions. The general result is then deduced from this observation along with measure theoretic arguments which go under the name of Dynkin's multiplicative systems theorem. ∎

Proposition D.34 (Disintegration I.). *Suppose that X is an $\mathbb{R}^k$ – valued random variable, Y is an $\mathbb{R}^l$ – valued random variable independent of X, and $f:\mathbb{R}^k\times\mathbb{R}^l\to\mathbb{R}_+$ then (assuming X and Y have continuous distributions $\rho_X(x)$ and $\rho_Y(y)$ respectively),*

$$\mathbb{E}\left[f\left(X,Y\right)\right] = \int_{\mathbb{R}^k}\mathbb{E}\left[f\left(x,Y\right)\right]\rho_X\left(x\right)dx \text{ and}$$
$$\mathbb{E}\left[f\left(X,Y\right)\right] = \int_{\mathbb{R}^l}\mathbb{E}\left[f\left(X,y\right)\right]\rho_Y\left(y\right)dy.$$

Proof. It is a fact that independence implies that the joint probability distribution, $\rho_{(X,Y)}(x,y)$, for (X,Y) must be given by

$$\rho_{(X,Y)}\left(x,y\right) = \rho_X\left(x\right)\rho_Y\left(y\right).$$

Therefore,

$$\begin{aligned}\mathbb{E}\left[f\left(X,Y\right)\right] &= \int_{\mathbb{R}^k\times\mathbb{R}^l} f\left(x,y\right)\rho_X\left(x\right)\rho_Y\left(y\right)dxdy\\ &= \int_{\mathbb{R}^k}\left[\int_{\mathbb{R}^l} dy f\left(x,y\right)\rho_Y\left(y\right)\right]\rho_X\left(x\right)dx\\ &= \int_{\mathbb{R}^k}\mathbb{E}\left[f\left(x,Y\right)\right]\rho_X\left(x\right)dx.\end{aligned}$$

■

One of the key theorems involving independent random variables is the strong law of large numbers. The other is the central limit theorem.

Theorem D.35 (Kolmogorov's Strong Law of Large Numbers). *Suppose that $\{X_n\}_{n=1}^\infty$ are i.i.d. random variables and let $S_n := X_1 + \cdots + X_n$. Then there exists $\mu \in \mathbb{R}$ such that $\frac{1}{n}S_n \to \mu$ a.s. iff X_n is integrable and in which case $\mathbb{E}X_n = \mu$.*

Remark D.36. If $\mathbb{E}|X_1| = \infty$ but $\mathbb{E}X_1^- < \infty$, then $\frac{1}{n}S_n \to \infty$ a.s. To prove this, for $M > 0$ let

$$X_n^M := \min(X_n, M) = \begin{cases} X_n \text{ if } X_n \le M \\ M \text{ if } X_n \ge M \end{cases}$$

and $S_n^M := \sum_{i=1}^n X_i^M$. It follows from Theorem D.35 that $\frac{1}{n}S_n^M \to \mu^M := \mathbb{E}X_1^M$ a.s.. Since $S_n \ge S_n^M$, we may conclude that

$$\liminf_{n\to\infty} \frac{S_n}{n} \ge \liminf_{n\to\infty} \frac{1}{n}S_n^M = \mu^M \text{ a.s.}$$

Since $\mu^M \to \infty$ as $M \to \infty$, it follows that $\liminf_{n\to\infty} \frac{S_n}{n} = \infty$ a.s. and hence that $\lim_{n\to\infty} \frac{S_n}{n} = \infty$ a.s.

Here is a crude special case of Theorem D.35 which however does come with a rate estimate. We will do considerably better later in Corollary **??**.

Proposition D.37. *Let $k \in \mathbb{N}$ with $k \ge 2$ and $\{X_n\}_{n=1}^\infty$ be i.i.d. random variables with $\mathbb{E}X_n = 0$ and $\mathbb{E}X_n^{2k} < \infty$. Then for every $p > \frac{1}{2} + \frac{1}{2k}$,*

$$\lim_{n\to\infty} \frac{S_n}{n^p} = 0 \text{ a.s.}$$

In other words for any $\varepsilon > 0$ small we have

$$\frac{S_n}{n} = \frac{S_n}{n^p}\frac{1}{n^{1-p}} = O\left(\frac{1}{n^{1-p}}\right) = O\left(\frac{1}{n^{\frac{1}{2}\left(1-\frac{1}{k}-\varepsilon\right)}}\right).$$

Proof. We start with the identity,

$$\mathbb{E}\left[\frac{1}{n}S_n\right]^{2k} = \frac{1}{n^{2k}} \sum_{j_1,\dots j_{2k}=1}^n \mathbb{E}\left[X_{j_1}\dots X_{j_{2k}}\right].$$

Using $\mathbb{E}\left[X_{j_1}\dots X_{j_{2k}}\right] = 0$ if there is any one index, j_l, distinct from the others, we conclude that the above sum can contain at most $C_k n^k$ non-zero terms for some $C_k < \infty$ and all of these terms are bounded by a constant C depending on $\mathbb{E}X_n^{2k}$. For example if $k = 2$ we have $\mathbb{E}\left[X_{j_1}X_{j_2}X_{j_3}X_{j_4}\right] = 0$ unless $j_1 = j_2 = j_3 = j_4$ (of which there are n such terms) or $j_1 = j_2$ and $j_3 = j_4$ (or similar with permuted indices) of which there are $3n^2$ – terms.

From the previous observations it follows that

$$\mathbb{E}\left[\frac{1}{n}S_n\right]^{2k} \le \frac{Cn^k}{n^{2k}} = C\frac{1}{n^k}.$$

Therefore if $0 < \alpha < 1$, then

$$\begin{aligned}\mathbb{E}\left(\sum_{n=1}^\infty \left[n^\alpha \frac{1}{n}S_n\right]^{2k}\right) &= \sum_{n=1}^\infty \mathbb{E}\left[n^\alpha \frac{1}{n}S_n\right]^{2k} \\ &\le \sum_{n=1}^\infty C\frac{1}{n^k}n^{\alpha 2k} = \sum_{n=1}^\infty C\frac{1}{n^{k(1-2\alpha)}} < \infty\end{aligned}$$

provided $k(1-2\alpha) > 1$, i.e. $1 - 2\alpha > \frac{1}{k}$, i.e. $\alpha < \frac{1}{2}\left(1 - \frac{1}{k}\right)$. For such an α we have

$$\sum_{n=1}^\infty \left[n^\alpha \frac{1}{n}S_n\right]^{2k} < \infty \text{ a.s.} \implies \lim_{n\to\infty} \frac{1}{n^{1-\alpha}}S_n = 0 \text{ a.s.}.$$

Tracing through the inequalities shows $p := 1 - \alpha > 1 - \frac{1}{2}\left(1 - \frac{1}{k}\right) = \frac{1}{2} + \frac{1}{2k}$ is the required restriction on p. ■

Often times for practical importance, the following weak law of large numbers is in fact more useful. For the proof we will need the following simple but very useful inequality.

Lemma D.38 (Chebyshev's Inequality). *If X is a random variable, $\delta > 0$, and $p > 0$, then*

$$P(\{|X| \ge \delta\}) = \mathbb{E}\left[1_{|X|\ge\delta}\right] \le \mathbb{E}\left[\frac{|X|^p}{\delta^p}1_{|X|\ge\delta}\right] \le \delta^{-p}\mathbb{E}|X|^p. \tag{D.34}$$

Proof. Taking expectations of the following pointwise inequalities,

$$1_{|X|\ge\delta} \le \frac{|X|^p}{\delta^p}1_{|X|\ge\delta} \le \delta^{-p}|X|^p,$$

immediately gives Eq. (D.34). ■

Theorem D.39. *Let $\{X_n\}_{n=1}^\infty$ be uncorrelated random square integrable random variables, then*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{m=1}^n (X_m - \mathbb{E}X_m)\right| \ge \delta\right) \le \frac{1}{\delta^2 n^2}\sum_{m=1}^n \operatorname{Var}(X_m).$$

If we further assume that $\mathbb{E}X_m = \mu$ *and* $\operatorname{Var}(X_m) = \sigma^2$ *are independent of* m, *then*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{m=1}^{n} X_m - \mu\right| \geq \delta\right) \leq \frac{\sigma^2}{\delta^2}\frac{1}{n}.$$

Proof. By Chebyshev's inequality and the assumption that $\operatorname{Cov}(X_m, X_k) = \delta_{mk}\operatorname{Var}(X_m)$, we find

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{1}{n}\sum_{m=1}^{n}(X_m - \mathbb{E}X_m)\right| \geq \delta\right) \\
&\quad\leq \frac{1}{\delta^2}\mathbb{E}\left|\frac{1}{n}\sum_{m=1}^{n}(X_m - \mathbb{E}X_m)\right|^2 \\
&\quad= \frac{1}{\delta^2 n^2}\mathbb{E}\sum_{m.k=1}^{n}\left[(X_m - \mathbb{E}X_m)(X_k - \mathbb{E}X_k)\right] \\
&\quad= \frac{1}{\delta^2 n^2}\sum_{m.k=1}^{n}\operatorname{Cov}(X_m, X_k) = \frac{1}{\delta^2 n^2}\sum_{m=1}^{n}\operatorname{Var}(X_m).
\end{aligned}$$

■

References

1. Persi Diaconis and J. W. Neuberger, *Numerical results for the Metropolis algorithm*, Experiment. Math. **13** (2004), no. 2, 207–213. MR 2068894
2. Richard Durrett, *Probability: theory and examples*, second ed., Duxbury Press, Belmont, CA, 1996. MR MR1609153 (98m:60001)
3. Robert D. Gordon, *Values of Mills' ratio of area to bounding ordinate and of the normal probability integral for large values of the argument*, Ann. Math. Statistics **12** (1941), 364–366. MR MR0005558 (3,171e)
4. Olav Kallenberg, *Foundations of modern probability*, second ed., Probability and its Applications (New York), Springer-Verlag, New York, 2002. MR MR1876169 (2002m:60002)
5. K. L. Mengersen and R. L. Tweedie, *Rates of convergence of the Hastings and Metropolis algorithms*, Ann. Statist. **24** (1996), no. 1, 101–121. MR 1389882 (98c:60081)
6. Sean Meyn and Richard L. Tweedie, *Markov chains and stochastic stability*, second ed., Cambridge University Press, Cambridge, 2009, With a prologue by Peter W. Glynn. MR 2509253 (2010h:60206)
7. J. R. Norris, *Markov chains*, Cambridge Series in Statistical and Probabilistic Mathematics, vol. 2, Cambridge University Press, Cambridge, 1998, Reprint of 1997 original. MR MR1600720 (99c:60144)
8. Yuval Peres, *An invitation to sample paths of brownian motion*, stat-www.berkeley.edu/ peres/bmall.pdf (2001), 1–68.