Stat 204 Notes

Matin Ghavamizadeh

Fall 2019

1 Probability Spaces

Definition. A probability space is a triple $(\Omega, \mathcal{F}, \mathbb{P})$ where:

- Ω is a set known as the **sample space**.
- $\mathcal{F} \subseteq 2^\Omega$ is a σ-field with each $A \in \mathcal{F}$ known as an **event**.
- $\mathbb{P}$ is a measure over $\mathcal{F}$ so that $\mathbb{P}(\Omega) = 1$ (such a measure is called a **probability measure**).

This definition relies on several other definitions. Most prominently that of a σ-field.

Definition. A σ-field (or σ-algebra) $\mathcal{F}$ over the set Ω is a subset of 2^Ω so that:

1. $\varnothing \in \mathcal{F}$;
2. $A \in \mathcal{F} \implies A^c := \Omega \setminus A \in \mathcal{F}$;
3. If $\{A_i\}$ is a countable collection of sets in $\mathcal{F}$ then $\cup_i A_i \in \mathcal{F}$.

That is, $\mathcal{F}$ contains the empty set and is closed under complements and countable unions.

De Morgan's laws imply that any σ-field is closed under countable intersections, since

$$\bigcap_i A_i = \left(\bigcup_i A_i^c \right)^c.$$

σ-fields are the suitable domains over which we can define the notion of measure of sets. Since $\mathbb{P}$ is a measure, it is fitting to give a general definition.

Definition. A measure μ over the σ-field $\mathcal{F}$ is a map $\mu : \mathcal{F} \to \mathbb{R}$ so that

1. $\mu(A) \geq 0$ for all $A \in \mathcal{F}$;
2. If $\{A_i\}$ is a countable collection of pairwise disjoint sets in $\mathcal{F}$ then $\mu(\cup_i A_i) = \sum_i \mu(A_i)$.

That is, a measure is a non-negative, countably additive set function over a σ-field.

1.1 Basic Properties of Measures

The following basic properties of measures are relevant in our study of probability theory.

Measure of the Empty Set Since for any $A \in \mathcal{F}$ we have $A \cap \varnothing = \varnothing$ we can use countable additivity to conclude

$$\mu(A \cup \varnothing) = \mu(A) + \mu(\varnothing)$$

but $A \cup \varnothing = A$ so

$$\mu(A) = \mu(A) + \mu(\varnothing) \implies \mu(\varnothing) = 0.$$

Monotonicity If $A \subseteq B$ for $A, B \in \mathcal{F}$ we have

$$B = (B \setminus A) \cup A \qquad (B \setminus A) \cap A = \varnothing$$

therefore

$$\mu(B) = \mu(B \setminus A) + \mu(A)$$

since $\mu(B \setminus A) \geq 0$ we conclude $\mu(B) \geq \mu(A)$.

Subadditivity Suppose $\{A_i\}$ is a countable collection of sets in $\mathcal{F}$. Let

$$B_n = A_n \setminus \bigcup_{i=1}^{n-1} A_i.$$

Clearly $\{B_i\}$ are pairwise disjoint and $\cup_i A_i = \cup_i B_i$, hence

$$\mu(\cup_i A_i) = \mu(\cup_i B_i) = \sum_i \mu(B_i).$$

At the same time $B_i \subseteq A_i$ so by monotonicity $\mu(B_i) \leq \mu(A_i)$. We therefore conclude

$$\mu(\cup_i A_i) \leq \sum_i \mu(A_i).$$

Continuity from Below Suppose $A_n \uparrow A$. Let $B_1 = A_1$ and

$$B_n = A_n \setminus A_{n-1} \qquad n \geq 2.$$

Clearly B_n are disjoint and

$$\cup_{n=1}^{\infty} B_n = \cup_{n=1}^{\infty} A_n = A.$$

Therefore,

$$\mu(A) = \mu(\cup_{n=1}^{\infty} B_n) = \sum_{n=1}^{\infty} \mu(B_n) = \lim_{N \to \infty} \sum_{n=1}^{N} \mu(B_n) = \lim_{N \to \infty} \mu(A_N).$$

Hence, $\mu(A_n) \uparrow \mu(A)$.

Continuity from Above Suppose $A_n \downarrow A$. Then $A_1 \setminus A_n \uparrow A_1 \setminus A$. Hence

$$\mu(A_1 \backslash A_n) \uparrow \mu(A_1 \backslash A) \implies \mu(A_1) - \mu(A_n) \uparrow \mu(A_1) - \mu(A) \implies \mu(A_n) \downarrow \mu(A).$$

Remark. It is reasonable to question why $\mathcal{F}$ needs to be explicitly identified in the definition of a probability space. That is, why can't we always take $\mathcal{F} = 2^{\Omega}$? The reason behind this requirement is that in many cases it is impossible to define a well-behaving measure that is defined for every subset of Ω. For instance, any "reasonable" measure cannot be defined for every subset of the unit interval. That is, if μ is a measure that is invariant under translation and

$$\mu([a, b]) = b - a$$

then it cannot be defined on every set. To illustrate this point, we use a simple construction by Vitali. Define the equivalence relation $\sim$ on the unit interval so that

$$x \sim y \iff y - x \in \mathbb{Q}.$$

This equivalence relation will partition $[0, 1]$. Utilizing the axiom of choice we can construct a set V that contains a point from every equivalence class of $\sim$. Let $\{q_n\}$ be an enumeration of the rationals in the unit interval. It is easy to confirm that the translated sets $V + q_n$ are disjoint and that

$$[0, 1] \subseteq \bigcup_n (V + q_n) \subseteq [0, 2].$$

Using countable additivity, monotonicity, and translation invariance of μ we conclude that

$$\mu([0, 1]) \subseteq \mu\left(\bigcup_n (V + q_n)\right) \subseteq \mu([0, 2]) \implies 1 \leq \sum_n \mu(V) \leq 2.$$

It is obvious that no value of $\mu(V)$ can satisfy the above inequality, therefore μ cannot be defined on V. A more general and elegant result is given by Banach and Tarski which we omit.

Remark. In the case where Ω is countable it is possible to take $\mathcal{F} = 2^{\Omega}$ and specify $\mathbb{P}$ by assigning probabilities to each $\omega \in \Omega$ so that

$$\sum_{\omega \in \Omega} \mathbb{P}(\omega) = 1.$$

Proposition. *Given a collection of σ-fields $\{\mathcal{F}_\alpha\}$, their intersection $\cap \mathcal{F}_\alpha$ is a σ-field.*

The above proposition allows us to give the following definition.

Definition. The **Borel σ-field** of a topology (Ω, τ), denoted by $\mathscr{B}((\Omega, \tau))$, is the intersection of all σ-fields on Ω that contain τ. That is, the Borel σ-field on a topology is the smallest σ-field that contains all the open sets of the topology.

Remark. All "reasonable" sets one can think of in $\mathbb{R}^d$ are Borel sets. In fact, all the known constructions of non-Borel sets utilize the axiom of choice in one way or another. Because of this, the Borel σ-algebra on the d-dimensional Euclidean space with the usual topology (denoted by $\mathscr{B}(\mathbb{R}^d)$) will play a central role in our study of random variables.

2 Random Variables

Definition. Let $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$ be measurable spaces. A **measurable map** $f : \Omega_1 \to \Omega_2$ satisfies

$$\forall A \in \mathcal{F}_2; f^{-1}(A) \in \mathcal{F}_1.$$

Definition. Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a d-dimensional random vector X is a map $X : \Omega \to \mathbb{R}^d$ so that

$$\forall A \in \mathscr{B}(\mathbb{R}^d); X^{-1}(A) \in \mathcal{F}.$$

That is, X is a measurable map from $(\Omega, \mathcal{F})$ to $(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d))$.

Definition. The distribution function F_X of a random vector X is a map $F_X : \mathscr{B}(\mathbb{R}^d) \to [0, 1]$ so that

$$\forall A \in \mathscr{B}(\mathbb{R}^d); \mathbb{P}\left(X^{-1}(A)\right) = F_X(A).$$

Remark. Note that F_X is a probability measure on $(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d))$:

1. Clearly F_X is non-negative since $\mathbb{P}$ is non-negative.
2. Given disjoint sets $A_i \in \mathscr{B}(\mathbb{R}^d)$ we have

   $$\mathbb{P}\left(X^{-1}\left(\bigcup_i A_i\right)\right) = F_X\left(\bigcup_i A_i\right)$$

   by definition. Also,

   $$\begin{aligned}\mathbb{P}\left(X^{-1}\left(\bigcup_i A_i\right)\right) &= \mathbb{P}\left(\bigcup_i X^{-1}(A_i)\right) \\ &= \sum_i \mathbb{P}\left(X^{-1}(A_i)\right) = \sum_i F_X(A_i)\end{aligned}$$

   so F_X is countably additive.
3. Clearly

   $$F_X(\mathbb{R}^d) = \mathbb{P}\left(X^{-1}(\mathbb{R}^d)\right) = \mathbb{P}(\Omega) = 1.$$

Remark. We use the following conventional notation when denoting probabilities of events involving random variables:

$$\begin{aligned} \mathbb{P}\left(X^{-1}(A)\right) &:= \mathbb{P}\left(X \in A\right) && (A \in \mathcal{F}) \\ \mathbb{P}\left(X^{-1}(\omega)\right) &:= \mathbb{P}\left(X = \omega\right) && (\omega \in \Omega) \\ \mathbb{P}\left(X^{-1}((-\infty, x])\right) &:= \mathbb{P}\left(X \leq x\right) && (\mathrm{Im}(X) = \mathbb{R}, x \in \mathbb{R}) \end{aligned}$$

$\mathbb{P}(X < x)$, $\mathbb{P}(X \geq x)$, and $\mathbb{P}(X > x)$ are defined in the same manner as the last definition above.

Definition. The distribution of a random vector X is said to be **absolutely continuous** if a map $f_X : \mathbb{R}^d \to \mathbb{R}^{\geq 0}$ exists so that

$$\forall B \in \mathscr{B}(\mathbb{R}^d); \mathbb{P}\left(X \in B\right) = \int_B f_X d\mu$$

where μ is the Lebesgue measure on $\mathbb{R}^d$. Such f_X is called the **density function** of X,

Proposition. *Given F_X we can find f_X by*

$$f_X(x) = \lim_B \frac{F_X(B)}{\mu(B)}$$

where B ranges over neighborhoods of x with diameter $\downarrow 0$.

2.1 Expectation

Definition. If X is a d-dimensional random vector with density f_X and $\phi : \mathbb{R}^d \to \mathbb{R}$, then the **expectation** of the random variable $\phi(X)$ is defined by

$$\mathbb{E}\left[\phi(X)\right] = \int_{\mathbb{R}^d} \phi \cdot f_X d\mu$$

where μ is the Lebesgue measure. We require ϕ to be a measurable map for $\phi(X)$ to be a random variable. We also require that

$$\int_{\mathbb{R}^d} |\phi| \cdot f_X d\mu < \infty$$

for the expectation to be well-defined.

2.2 Marginal, Joint, and Conditional Densities

Suppose $X = (X_1, \ldots, X_d)$ is a d-dimensional random vector, and let $Y = (X_1, \ldots, X_k)$ and $Z = (X_{k+1}, \ldots, X_d)$; i.e. $X = (Y, Z)$. Let $f_X := f_{Y,Z}$ be the density of X. Note that for any $B \in \mathscr{B}(\mathbb{R}^k)$ we have

$$\mathbb{P}\left(Y \in B\right) = \mathbb{P}\left((Y, Z) \in B \times \mathbb{R}^{d-k}\right) = \int_{B \times \mathbb{R}^{d-k}} f_X d\mu$$

where μ is the Lebesgue measure on $\mathbb{R}^d$. Noting that any density function is non-negative by definition and that the Lebesgue measure is σ-finite, we can apply Fubini's theorem to the above integral to get

$$\therefore \mathbb{P}(Y \in B) = \int_B \int_{\mathbb{R}^{d-k}} f_{Y,Z}(y,z)\mu_1(dz)\mu_2(dy).$$

where μ_1 and μ_2 are Lebesgue measure on the appropriate spaces. This shows that Y must be absolutely continuous with density

$$f_Y(y) = \int_{\mathbb{R}^{d-k}} f_{Y,Z}(y,z)dz.$$

The density $f_X = f_{Y,Z}$ is known as the **joint density** of random vectors Y and Z. The density f_Y is known as the **marginal density** of Y. Now consider the points

$$y = (y_1, \ldots, y_k) \qquad z = (z_1, \ldots, z_{d-k})$$

and let Δz and Δy be small boxes including these points; that is

$$\Delta y = \Delta y_1 \times \Delta y_2 \times \ldots \times \Delta y_k \qquad \Delta z = \Delta z_1 \times \Delta z_2 \times \ldots \times \Delta z_{d-k}$$

where each Δy_i is a segment containing y_i, ditto Δz_i. Using the definition of conditional probability we have

$$\mathbb{P}(z \in \Delta z | y \in \Delta y) = \frac{\mathbb{P}(z \in \Delta z, y \in \Delta y)}{\mathbb{P}(y \in \Delta y)}.$$

When Δy and Δz are small, we can intuitively write the above equation as

$$\frac{\mathbb{P}(z \in \Delta z, y \in \Delta y)}{\mathbb{P}(y \in \Delta y)} \approx \frac{f_{Y,Z}(y,z) \cdot \mu(\Delta y \times \Delta z)}{f_Y(y) \cdot \mu(\Delta y)} = \frac{f_{Y,Z}(y,z)}{f_Y(y)} \cdot \mu(\Delta z).$$

In summary, we have

$$\mathbb{P}(z \in \Delta z | y \in \Delta y) \approx \frac{f_{Y,Z}(y,z)}{f_Y(y)} \cdot \mu(\Delta z).$$

This motivates the following definition:

Definition. Given two random vectors Y and Z with joint density $f_{Y,Z}$ if the marginal density of Y, f_Y, is positive we define **the conditional density of Z given Y** as

$$f_{Z|Y}(z|y) = \frac{f_{Z,Y}(z,y)}{f_Y(y)}.$$

Another way to motivate the above definition, is by looking at the expectation of $\phi(Y,Z)$:

$$\mathbb{E}[\phi(Y,Z)] = \int_{\mathbb{R}^d} \phi(y,z) f_{Y,Z}(y,z) d\mu$$

since ϕ has to be absolutely integrable we can again apply Fubini's theorem to see

$$\mathbb{E}\left[\phi(Y,Z)\right] = \int_{\mathbb{R}^k}\int_{\mathbb{R}^{d-k}} \phi(y,z)\frac{f_{Y,Z}(y,z)}{f_Y(y)}d\mu_1(z)f_Y(y)d\mu_2(y)$$

which again suggests we set

$$f_{Z|Y}(z|y) = \frac{f_{Z,Y}(z,y)}{f_Y(y)}.$$

2.3 Independence

Definition. Two events $A, B \in \mathcal{F}$ are said to be **independent** (denoted $A \perp\!\!\!\perp B$) iff

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

If the events have non-zero probabilities we have the following equivalent conditions that are more intuitive

$$A \perp\!\!\!\perp B \iff \mathbb{P}\left(A|B\right) = \mathbb{P}\left(A\right) \iff \mathbb{P}\left(B|A\right) = \mathbb{P}\left(B\right).$$

Similarly, we say two random variables X and Y are independent if for any two appropriate Borel sets A and B we have

$$\mathbb{P}\left(X \in A, Y \in B\right) = \mathbb{P}\left(X \in A\right)\mathbb{P}\left(Y \in B\right).$$

Note that in this case we will have

$$f_{X,Y}(x,y) = f_X(x)f_Y(y)$$

which implies that

$$f_{X|Y}(x|y) = f_X(x) \qquad f_{Y|X}(y|x) = f_Y(y).$$

2.4 Change of Variables

Consider the problem of finding the density of the random variable $Y = g(X)$ where X is a random variable with known density f_X and $g : \mathbb{R} \to \mathbb{R}$ is a strictly increasing differentiable function. If f_Y is the density of Y, we must have

$$\mathbb{P}\left(Y \leq y\right) = \int_{-\infty}^{y} f_Y(t)dt.$$

Therefore, by applying the fundamental theorem of calculus we get

$$f_Y(y) = \frac{d}{dy}\mathbb{P}\left(Y \leq y\right). \tag{1}$$

On the other hand

$$\mathbb{P}\left(Y \leq y\right) = \mathbb{P}\left(g(X) \leq y\right).$$

Since g is strictly increasing it must be injective with an increasing inverse; hence,

$$\mathbb{P}\left(g(X) \leq y\right) = \mathbb{P}\left(X \leq g^{-1}(y)\right) = \int_{-\infty}^{g^{-1}(y)} f_X(s)ds.$$

At this point we can use (1) in addition to the fundamental theorem of calculus and the chain rule to find f_Y. This solution does not generalize well to higher dimensions, so instead we perform the change of variables

$$r = g(s) \iff s = g^{-1}(r)$$

which gives

$$\mathbb{P}\left(Y \leq y\right) = \int_{-\infty}^{y} f_X(g^{-1}(r)){g^{-1}}'(r)dr$$

and ultimately using (1) and the fundamental theorem of calculus

$$f_Y(y) = f_X(g^{-1}(y)){g^{-1}}'(y).$$

We can carry out a similar process for higher dimensional random vectors. If X is a d-dimensional random vector and $g : \mathbb{R}^d \to \mathbb{R}^d$ is injective and continuously differentiable (hence measurable), we can find the density of $Y = g(X)$ by noticing that if extant, it must satisfy

$$\mathbb{P}\left(Y \in B\right) = \int_B f_Y(y)\mu(dy)$$

for any Borel set B. Also,

$$\mathbb{P}\left(Y \in B\right) = \mathbb{P}\left(g(X) \in B\right) = \mathbb{P}\left(X \in g^{-1}(B)\right) \tag{2}$$

the last equality holds because g is injective and measurable. Given the density of X we can write

$$\mathbb{P}\left(X \in g^{-1}(B)\right) = \int_{g^{-1}(B)} f_X(x)\mu(dx).$$

Since g is injective and continuously differentiable we can perform the change of variables

$$x = g^{-1}(y) \qquad \mu(dx) = \mu(dg^{-1}(y)) = |\det J_{g^{-1}}|\mu(dy)$$

where $J_{g^{-1}}$ is the Jacobian determinant of g^{-1} (or equivalently the inverse Jacobian determinant of g given by

$$J_{g^{-1}}(i,j) = \frac{\partial x_i}{\partial y_j}$$

where

$$(x_1, \ldots, x_d) \overset{g}{\mapsto} (y_1, \ldots, y_d).$$

Hence,

$$\mathbb{P}\left(X \in g^{-1}(B)\right) = \int_B f_X(g^{-1}(y))|\det J_{g^{-1}}|\mu(dy)$$

which combined with (2) gives us

$$\therefore \mathbb{P}\left(Y \in B\right) = \int_B f_X(g^{-1}(y))|\det J_{g^{-1}}|\mu(dy);$$

i.e.

$$f_Y(y) = f_X(g^{-1}(y))|\det J_{g^{-1}}|.$$

We can summarize as follows:

Proposition 1. *Let X have density f_X and assume there is an open set $S \subseteq \mathbb{R}^d$ such that $\mathbb{P}(X \in S) = 1$ and on S the mapping $g : \mathbb{R}^d \to \mathbb{R}^d$ is one to one and continuously differentiable with Jacobian determinant nonzero at each point of S. Then the random vector $Y = g(X)$ has an absolutely continuous distribution with density f_Y given by*

$$f_Y(y) = f_X(g^{-1}(y))|\det J_{g^{-1}}|\mathbf{1}_{g(S)}(y).$$

3 The Dirichlet Distribution

The following property of the gamma distribution is central to our analysis of the Dirichlet distribution.

Proposition (Additivity of Gamma Random Variables)**.** *Suppose for independent random variables X_1 and X_2 we have*

$$\begin{aligned} X_1 &\sim \text{gamma}(\alpha_1, \beta), \\ X_2 &\sim \text{gamma}(\alpha_2, \beta), \\ Y &:= X_1 + X_2 \end{aligned}$$

then,

$$Y \sim \text{gamma}(\alpha_1 + \alpha_2, \beta)\,.$$

Proof. Since X_1 and X_2 are independent, the density of Y is given by the convolution

$$f_Y(y) = \int_{-\infty}^{+\infty} f_{X_1}(t) f_{X_2}(y-t)dt.$$

Both X_1 and X_2 have positive support, so

$$\begin{aligned} f_Y(y) &= \int_0^y f_{X_1}(t) f_{X_2}(y-t) dt \\ &= \int_0^y \frac{\beta^{\alpha_1}}{\Gamma(\alpha_1)} t^{\alpha_1 - 1} e^{-\beta t} \frac{\beta^{\alpha_2}}{\Gamma(\alpha_2)} (y-t)^{\alpha_2 - 1} e^{-\beta(y-t)} dt \\ &= \frac{\beta^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} e^{-\beta y} \int_0^y t^{\alpha_1 - 1} (y-t)^{\alpha_2 - 1} dt. \end{aligned} \tag{3}$$

Letting $t = ys$ in the last integral we get

$$\int_0^y t^{\alpha_1 - 1} (y-t)^{\alpha_2 - 1} dt = y^{\alpha_1+\alpha_2-1} \int_0^1 s^{\alpha_1 - 1} (1-s)^{\alpha_2 - 1} ds = y^{\alpha_1+\alpha_2-1} \mathrm{B}(\alpha_1, \alpha_2) \tag{4}$$

where $\mathrm{B}(\cdot, \cdot)$ denotes the beta function. Applying the identity

$$\mathrm{B}(\alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1 + \alpha_2)}$$

and combining (3) and (4) we get

$$\therefore f_Y(y) = \frac{\beta^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1 + \alpha_2)} y^{\alpha_1+\alpha_2-1} e^{-\beta y}.$$

□

Proposition 2. *Let $X_1, \ldots, X_{n+1}$ be independent and assume*

$$X_i \sim \mathrm{gamma}(\alpha_\mathrm{i}, \beta)$$

Let

$$S_i := X_1 + \ldots + X_i \qquad 1 \leq i \leq n+1$$

and

$$V_i := \frac{X_i}{S_{n+1}}.$$

Then $(V_1, \ldots, V_{n+1})$ and S_{n+1} are independent and the random vector $\mathbf{V} = (V_1, \ldots, V_n)$ has an absolutely continuous distribution with density

$$f_{\mathbf{V}}(v_1, \ldots, v_n) = \frac{1}{D(\alpha_1, \ldots, \alpha_{n+1})} \left(\prod_{i=1}^n v_i^{\alpha_i - 1} \right) \left(1 - \sum_{i=1}^n v_i \right)^{\alpha_{n+1} - 1} \mathbf{1}_{A_n}(v_1, \ldots, v_n) \tag{5}$$

where D is defined as

$$D(\alpha_1, \ldots, \alpha_{n+1}) = \frac{\Gamma(\alpha_1) \ldots \Gamma(\alpha_{n+1})}{\Gamma(\alpha_1 + \ldots + \alpha_{n+1})} \tag{6}$$

and A_n is the standard n-dimensional simplex given by

$$A_n = \left\{ (v_1, \ldots, v_n) \middle| v_i \geq 0, \sum_{i=1}^{n} v_i \leq 1 \right\}$$

Proof. First, note that by additivity of gamma random variables S_{n+1} must have density

$$f_{S_{n+1}}(s) = \frac{\beta^{\alpha_1+\ldots+\alpha_{n+1}}}{\Gamma(\alpha_1 + \ldots + \alpha_{n+1})} s^{\alpha_1+\ldots+\alpha_{n+1}-1} e^{-\beta s} \tag{7}$$

Now consider the map $g : \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$ so that

$$(x_1, \ldots, x_{n+1}) \mapsto (v_1, \ldots, v_n, s)$$

where

$$s = x_1 + \ldots + x_{n+1} \qquad v_i = \frac{x_i}{s} \quad 1 \leq i \leq n.$$

Note that g is absolutely continuous and invertible on A_{n+1}, moreover

$$x_i = v_i s \quad 1 \leq i \leq n \qquad x_{n+1} = s(1 - v_1 - \ldots - v_n). \tag{8}$$

Note that

$$(V_1, \ldots, V_n, S_{n+1}) = g(X_1, \ldots, X_{n+1})$$

so we can use proposition 1 to conclude that $(V_1, \ldots, V_n, S_{n+1})$ must have density

$$f(v_1, \ldots, v_n, s) = f_{X_1,\ldots,X_{n+1}}(g^{-1}(v_1, \ldots, v_n, s)) |\det J_{g^{-1}}(v_1, \ldots, v_n, s)| \tag{9}$$

Using (8) we can find the Jacobian determinant

$$\det J_{g^{-1}}(v_1, \ldots, v_n, s_{n+1}) = \begin{vmatrix} s & 0 & \ldots & 0 & v_1 \\ 0 & s & \ldots & 0 & v_2 \\ \vdots & & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & s & v_n \\ -s & -s & \ldots & -s & 1 - \sum_{i=1}^{n} v_i \end{vmatrix}$$

since adding a multiple of a row to another doesn't change the determinant, we can add the first n rows to the last row to get

$$\det J_{g^{-1}}(v_1, \ldots, v_n, s_{n+1}) = \begin{vmatrix} s & 0 & \ldots & 0 & v_1 \\ 0 & s & \ldots & 0 & v_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & s & v_n \\ 0 & 0 & \ldots & 0 & 1 \end{vmatrix} = s^n.$$

So we can simplify (9) to

$$f(v_1, \ldots, v_n, s) = f_{X_1,\ldots,X_{n+1}}(sv_1, \ldots, sv_n, s(1 - v_1 - \ldots - v_n)))s^n.$$

since X_i are independent we can write the joint density as a product of marginals to get

$$f(v_1, \ldots, v_n, s) = \left(\prod_{i=1}^{n} \frac{\beta^{\alpha_i}}{\Gamma(\alpha_i)}(sv_i)^{\alpha_i - 1}e^{-\beta s v_i}\right) \times \frac{\beta^{\alpha_{n+1}}}{\Gamma(\alpha_{n+i})}(s(1 - \sum_{i}^{n} v_i))^{\alpha_{n+1}-1}e^{-\beta s(1-\sum_i^n v_i)}s^n.$$

Simplifying and letting $\alpha_0 = \alpha_1 + \ldots + \alpha_{n+1}$ we get

$$f(v_1, \ldots, v_n, s) = \frac{1}{\prod_{i=1}^{n+1}\Gamma(\alpha_i)} v_i^{\alpha_i - 1} \ldots v_n^{\alpha_n - 1}(1 - \sum_{i}^{n} v_i))^{\alpha_{n+1}-1}\beta^{\alpha_0}s^{\alpha_0 - 1}e^{-\beta s}$$

multiplying and dividing by $\Gamma(\alpha_0)$ we finally get

$$f(v_1, \ldots, v_n, s) = \frac{\Gamma(\alpha_0)}{\prod_{i=1}^{n+1}\Gamma(\alpha_i)} v_i^{\alpha_i - 1} \ldots v_n^{\alpha_n - 1}(1 - \sum_{i}^{n} v_i))^{\alpha_{n+1}-1}\frac{\beta^{\alpha_0}}{\Gamma(\alpha_0)}s^{\alpha_0 - 1}e^{-\beta s}. \tag{10}$$

Noting that

$$\int_0^\infty \beta^{\alpha_0}s^{\alpha_0 - 1}e^{-\beta s}ds = \Gamma(\alpha_0)$$

we can marginalize s out and conclude

$$f(v_1, \ldots, v_n) = \frac{1}{D(\alpha_1, \ldots, \alpha_{n+1})} v_1^{\alpha_1 - 1} \ldots v_n^{\alpha_n - 1}(1 - \sum_{i}^{n} v_i))^{\alpha_{n+1}-1}. \tag{11}$$

Putting together (7), (10), and (11) we can see that the joint density of $(V_1, \ldots, V_n)$ and S_{n+1} factors into the product of marginal densities of $(V_1, \ldots, V_n)$ and S_{n+1}. Hence, $(V_1, \ldots, V_n)$ and S_{n+1} are independent. Since V_{n+1} is determined by $V_1, \ldots, V_n$ we finally see that $(V_1, \ldots, V_{n+1})$ and S_{n+1} are independent. □

Remark. Note that marginalizing v_i out of (10) gives an alternate proof of the additivity of independent gamma random variables.

We are now ready to give the definiton of the Dirichlet distribution:

Definition 1. The **Dirichlet distribution** on $\mathbb{R}^n$ having parameter $\alpha = (\alpha_1, \ldots, \alpha_{n+1})$ where $\alpha_i > 0$ is the absolutely continuous distribution having the density

$$\frac{1}{D(\alpha_1, \ldots, \alpha_{n+1})} v_1^{\alpha_1 - 1} \ldots v_n^{\alpha_n - 1}(1 - \sum_{i=1}^{n} v_i))^{\alpha_{n+1}-1}\mathbf{1}_{A_n}(\mathbf{v}).$$

Proposition 3. *Let* $\mathbf{V}$ *have the Dirichlet distribution on* $\mathbb{R}^n$ *with parameter* $\alpha = (\alpha_1, \ldots, \alpha_{n+1})$. *Then* $(V_{k+1}, \ldots, V_n), k < n$ *has the Dirichlet distribution on* $\mathbb{R}^{n-k}$ *with parameter*

$$\alpha_{\mathbf{k}} = (\alpha_{k+1}, \ldots, \alpha_n, \alpha_1 + \ldots + \alpha_k + \alpha_{n+1}).$$

Proof. Let

$$U = X_1 + \ldots + X_k + X_{n+1}$$

because of additivity of the gamma distribution we will have

$$U \sim \text{gamma}(\alpha_1 + \ldots + \alpha_k + \alpha_{n+1}, \beta)$$

and it will be independent of $X_{k+1}, \ldots, X_n$. Note that

$$S_{n+1} = X_k + \ldots + X_n + U.$$

Applying proposition 2 to the independent gamma random variables $X_k, \ldots, X_{k+1}$ and U finishes the proof. □

Proposition 4. *Let* $\mathbf{V}$ *have the Dirichlet distribution on* $\mathbb{R}^n$ *with parameter* $\alpha = (\alpha_1, \ldots, \alpha_{n+1})$. *Then the conditional distribution of* $(V_1, \ldots, V_k)$, *given* $V_{k+1} = v_{k+1}, \ldots, V_n = v_n$, *with*

$$v_{k+1} > 0, \ldots, v_n > 0, \qquad \sum_{i=k+1}^{n} v_i < 1 \qquad k < n,$$

is the distribution of $\left(1 - \sum_{i=k+1}^{n} v_i\right) W$, *where* W *has the Dirichlet distribution on* $\mathbb{R}^k$ *with parameters* $(\alpha_1, \ldots, \alpha_k, \alpha_{n+1})$.

Proof. Applying the definition of conditional density we have

$$f_{V_1,\ldots,V_k|V_{k+1},\ldots,V_n}(v_1, \ldots, v_k | v_{k+1}, \ldots, v_n) = \frac{f_{V_1,\ldots,V_n}(v_1, \ldots, v_n)}{f_{V_{k+1},\ldots,V_n}(v_{k+1}, \ldots, v_n)}.$$

By proposition 2

$$f_{V_1,\ldots,V_n}(v_1, \ldots, v_n) = \frac{1}{D(\alpha_1, \ldots, \alpha_{n+1})} v_1^{\alpha_1 - 1} \ldots v_n^{\alpha_n - 1} (1 - \sum_{i=1}^{n} v_i)^{\alpha_{n+1} - 1},$$

and by proposition 3

$$f_{V_{k+1},\ldots,V_n}(v_{k+1}, \ldots, v_n) = \frac{1}{D(\alpha_{k+1}, \ldots, \alpha_n, \alpha_{n+1} + \sum_{i=1}^{k} \alpha_i)} \times v_{k+1}^{\alpha_{k+1} - 1} \ldots v_n^{\alpha_n - 1} (1 - \sum_{i=k+1}^{n} v_i)^{\alpha_{n+1} - 1 + \sum_{i=1}^{k} \alpha_i}.$$

Taking the ratio we will see that the conditional density is of the form

$$\frac{\Gamma(\alpha_{n+1}+\sum_{i=1}^{k}\alpha_i)}{\Gamma(\alpha_1)\dots\Gamma(\alpha_k)\Gamma(\alpha_{n+1})}v_1^{\alpha_1-1}\dots v_k^{\alpha_k-1}\frac{(1-\sum_{i=1}^{n}v_i)^{\alpha_{n+1}-1}}{(1-\sum_{i=k+1}^{n}v_i)^{\alpha_{n+1}-1+\sum_{i=1}^{k}\alpha_i}}.$$

Now let $c_0 = \left(1-\sum_{i=k+1}^{n}v_i\right)$ and note that

$$\begin{aligned}
&\mathbb{P}\left(c_0 W \in (v_1-\delta/2, v_1+\delta/2)\times\ldots\times(v_k-\delta/2, v_k+\delta/2)\right)\\
=&\mathbb{P}\left(W \in (v_1/c_0-\delta/2c_0, v_1/c_0+\delta/2c_0)\times\ldots\times(v_k/c_0-\delta/2c_0, v_k/c_0+\delta/2c_0)\right)\\
\approx&\frac{1}{D(\alpha_1,\dots,\alpha_k,\alpha_{n+1})}(v_1/c_0)^{\alpha_1-1}\dots(v_k/c_0)^{\alpha_n-1}(1-\sum_{i=1}^{n}v_i/c_0)^{\alpha_{n+1}-1}\times\frac{\delta^k}{c_0^k}\\
=&\frac{\Gamma(\alpha_{n+1}+\sum_{i=1}^{k}\alpha_i)}{\Gamma(\alpha_1)\dots\Gamma(\alpha_k)\Gamma(\alpha_{n+1})}v_1^{\alpha_1-1}\dots v_k^{\alpha_k-1}\frac{(1-\sum_{i=1}^{n}v_i)^{\alpha_{n+1}-1}}{(1-\sum_{i=k+1}^{n}v_i)^{\alpha_{n+1}-1+\sum_{i=1}^{k}\alpha_i}}\delta^k.
\end{aligned}$$

Dividing by the volume of $(v_1-\delta/2, v_1+\delta/2)\times\ldots\times(v_k-\delta/2, v_k+\delta/2)$, i.e. δ^k, we can see that the two densities are the same. □

Proposition 5. *Let* $\mathbf{V} = (V_1,\dots,V_n)$ *have the Dirichlet distribution on* $\mathbb{R}^n$ *with parameter* $\alpha = (\alpha_1,\dots,\alpha_{n+1})$. *Let* $V_i^* = V_1+\ldots+V_i, 1\le i\le n$, *and let*

$$\mathbf{V}^* = (V_1^*,\dots,V_n^*).$$

Then $\mathbf{V}^*$ *has an absolutely continuous distribution with density*

$$f_{\mathbf{V}^*}(v_1,\dots,v_n) = \frac{v_1^{\alpha_1-1}\prod_{i=2}^{n}(v_i-v_{i-1})^{\alpha_i-1}(1-v_n)^{\alpha_{n+1}-1}\mathbf{1}_{0<v_1<\ldots<v_n<1}}{D(\alpha_1,\dots,\alpha_{n+1})}.$$

Proof. Notice that $\mathbf{V}^* = A\mathbf{V}$ where

$$A = \begin{bmatrix} 1 & 0 & 0 & \dots & 0\\ 1 & 1 & 0 & \dots & 0\\ 1 & 1 & 1 & \dots & 0\\ 1 & 1 & 1 & \dots & 1 \end{bmatrix}$$

so the inverse Jacobian determinant of the function that maps $\mathbf{V}$ to $\mathbf{V}^*$ is given by

$$\det(A^{-1}) = \frac{1}{\det(A)} = \frac{1}{1} = 1.$$

Hence, applying the change of variable formula (proposition 1) we get

$$\begin{aligned}
f_{\mathbf{V}^*}(v_1,\dots,v_n) &= f_{\mathbf{V}}(A^{-1}(v_1,\dots,v_n)')\times 1\times\mathbf{1}_{0<v_1<\ldots<v_n<1}\\
&= f_{\mathbf{V}}(v_1, v_2-v_1,\dots,v_{n-1}-v_{n-2},v_n)\mathbf{1}_{0<v_1<\ldots<v_n<1}.
\end{aligned}$$

Expanding $f_{\mathbf{V}}$ as the density of the Dirichlet distribution on $\mathbb{R}^n$ for finishes the proof. □

Definition 2. The distribution on $\mathbb{R}^n$ having the density

$$f(v_1,\ldots,v_n) = \frac{v_1^{\alpha_1-1}\prod_{i=2}^n (v_i - v_{i-1})^{\alpha_i-1}(1-v_n)^{\alpha_{n+1}-1}\mathbf{1}_{0<v_1<\ldots<v_n<1}}{D(\alpha_1,\ldots,\alpha_{n+1})}.$$

is called the **ordered Dirichlet distribution** on $\mathbb{R}^n$ with parameter $(\alpha_1,\ldots,\alpha_{n+1})$.

Proposition 6. *Let* $\mathbf{V}$ *have the ordered Dirichlet distribution on* $\mathbb{R}^n$ *with parameter* $(\alpha_1,\ldots,\alpha_{n+1})$ *and let* $1 \le i_1 < i_2 < \ldots < i_k \le n$. *Set*

$$\begin{aligned}
\gamma_1 &= \alpha_1 + \ldots + \alpha_{i_1} \\
\gamma_j &= \alpha_{i_{j-1}+1} + \ldots + \alpha_{i_j} \qquad\qquad 2 \le j < k \\
\gamma_k &= \alpha_{i_k} + \ldots + \alpha_{n+1}.
\end{aligned}$$

Then $(V_{i_1},\ldots,V_{i_k})$ *has the ordered Dirichlet distribution on* $\mathbb{R}^n$ *with parameter* $(\gamma_1,\ldots,\gamma_k)$.

Proof. Let S_j, X_j be as in proposition 2, and consider the random variables

$$\begin{aligned}
T_1 := S_{i_1} \qquad &= X_1 + \ldots + X_{i_1} \\
T_j := S_{i_j} - S_{i_{j-1}} \qquad &= X_{i_{j-1}+1} + \ldots + X_{i_j} \qquad 2 \le j \le k \\
T_{k+1} := S_{n+1} - S_{i_k} \qquad &= X_{i_k+1} + \ldots + X_{n+1}.
\end{aligned}$$

By the additivity of the gamma distribution and independence of X_j, we conclude that

$$T_k \sim \text{gamma}(\gamma_k, \beta).$$

Noting that

$$\sum_{i=1}^{k+1} T_i = S_{n+1}$$

we can apply proposition 2 to see that $(T_1/S_{n+1},\ldots,T_k/S_{k+1})$ has the Dirichlet distribution on $\mathbb{R}^k$ with parameter $(\gamma_1,\ldots,\gamma_{k+1})$. Hence, by proposition 5 we conclude that

$$\left(\frac{T_1}{S_{n+1}}, \frac{T_1+T_2}{S_{n+1}}, \ldots, \frac{T_1+\ldots+T_k}{S_{n+1}}\right) \sim \text{OrderedDirichlet}(\gamma_1,\ldots,\gamma_k).$$

Noting that

$$\sum_{i=1}^{j} T_i = S_{i_j}$$

and that $(V_{i_1},\ldots,V_{i_k})$ is distributed like $(S_{i_1}/S_{n+1},\ldots,S_{i_k}/S_{n+1})$ we can see that $(V_{i_1},\ldots,V_{i_k})$ has the ordered Dirichlet distribution on $\mathbb{R}^k$ with parameter $(\gamma_1,\ldots,\gamma_{k+1})$. □

Proposition 7. *Let* $\mathbf{V}$ *have the ordered Dirichlet distribution on* $\mathbb{R}^n$ *with parameter* $(\alpha_1, \ldots, \alpha_{n+1})$*. Then the conditional distribution of*

$$(V_1, \ldots, V_{k-1}, V_{k+1}, \ldots, V_n)$$

given $V_k = v_k$ *is like the distribution of* $(v_k\mathbf{X}, v_k + (1-v_k)\mathbf{Y})$*, where* $\mathbf{X}$ *and* $\mathbf{Y}$ *are independent,* $\mathbf{X}$ *has the ordered Dirichlet distribution on* $\mathbb{R}^{k-1}$ *with parameter* $(\alpha_1, \ldots, \alpha_k)$ *and* $\mathbf{Y}$ *has the ordered Dirichlet distribution on* $\mathbb{R}^{n-k}$ *with parameter* $(\alpha_{k+1}, \ldots, \alpha_{n+1})$*. Note that we must have* $0 < v_k < 1$ *for the conditional density to be well defined.*

Proof. First note that by definition of conditional density

$$f_{V_1,\ldots,V_{k-1},V_{k+1},\ldots,V_n|V_k}(v_1, \ldots, v_{k-1}, v_{k+1}, \ldots, v_n|v_k) = \frac{f_{V_1,\ldots,V_n}(v_1, \ldots, v_n)}{f_{V_k}(v_k)}$$

proposition 5 gives us an expression for the numerator and proposition 6 allows us to find an expression for the denominator. Applying these propositions we see that the conditional density must be of the form

$$\frac{\frac{1}{D(\alpha_1,\ldots,\alpha_{n+1})} v_1^{\alpha_1-1} \prod_{i=2}^{n} (v_i - v_{i-1})^{\alpha_i-1} (1-v_n)^{\alpha_{n+1}-1}}{\frac{1}{D(\alpha_1+\ldots+\alpha_k, \alpha_{k+1}+\ldots+\alpha_{n+1})} v_k^{\alpha_1+\ldots+\alpha_k-1} (1-v_k)^{\alpha_{k+1}+\ldots+\alpha_{n+1}-1}},$$

that is,

$$\frac{\Gamma(\alpha_1 + \ldots + \alpha_k)\Gamma(\alpha_{k+1} + \ldots + \alpha_{n+1})}{\Gamma(\alpha_1) \ldots \Gamma(\alpha_{n+1})} \frac{v_1^{\alpha_1-1} \prod_{i=2}^{n} (v_i - v_{i-1})^{\alpha_i-1} (1-v_n)^{\alpha_{n+1}-1}}{v_k^{\alpha_1+\ldots+\alpha_k-1} (1-v_k)^{\alpha_{k+1}+\ldots+\alpha_{n+1}-1}}. \tag{12}$$

Now let $\mathbf{Z} = (v_k\mathbf{X}, v_k + (1-v_k)\mathbf{Y})$ where $\mathbf{X}$ and $\mathbf{Y}$ are independent, $\mathbf{X}$ has the ordered Dirichlet distribution on $\mathbb{R}^{k-1}$ with parameter $(\alpha_1, \ldots, \alpha_k)$ and $\mathbf{Y}$ has the ordered Dirichlet distribution on $\mathbb{R}^{n-k}$ with parameter $(\alpha_{k+1}, \ldots, \alpha_{n+1})$. And let g be an operator on $\mathbb{R}^{n-1}$ that maps $(\mathbf{X}, \mathbf{Y})$ to $\mathbf{Z}$; i.e.

$$(x_1, \ldots, x_{k-1}, y_1, \ldots, y_{n-k}) \overset{g}{\mapsto} (z_1, \ldots, z_{n-1}).$$

It is easy to see that

$$\begin{aligned} x_j &= \frac{1}{v_k} x_j && 1 \le j \le k-1 \\ y_j &= \frac{z_{j+k-1} - v_k}{1 - v_k} && 1 \le j \le n-k \end{aligned}$$

so the inverse Jacobian determinant of g is given by

$$\det J_{g^{-1}} = \begin{vmatrix} \frac{1}{v_k} & & & & & & \\ & \ddots & & & & & \\ & & \frac{1}{v_k} & & & & \\ & & & \frac{1}{1-v_k} & & & \\ & & & & \ddots & & \\ & & & & & \frac{1}{1-v_k} & \end{vmatrix} = \frac{1}{v_k^{k-1}(1-v_k)^{n-k}}.$$

Using change of variables (proposition 1) we get

$$\begin{aligned} f_{\mathbf{Z}}(z_1,\ldots,z_{n-1}) &= f_{\mathbf{X},\mathbf{Y}}\left(g^{-1}(z_1,\ldots,z_{k-1})\right)|\det J_{g^{-1}}| \\ &= f_{\mathbf{X},\mathbf{Y}}\left(\frac{z_1}{v_k},\ldots,\frac{z_{k-1}}{v_k},\frac{z_{k+1}}{v_k},\ldots,\frac{z_{n-1}}{v_k}\right)\frac{1}{v_k^{k-1}(1-v_k)^{n-k}} \\ &= f_{\mathbf{X}}\left(\frac{z_1}{v_k},\ldots,\frac{z_{k-1}}{v_k}\right)f_{\mathbf{Y}}\left(\frac{z_{k+1}}{v_k},\ldots,\frac{z_{n-1}}{v_k}\right)\frac{1}{v_k^{k-1}(1-v_k)^{n-k}} \end{aligned}$$

where the last equality results from the independence of $\mathbf{X}$ and $\mathbf{Y}$. Expanding $f_{\mathbf{X}}$ and $f_{\mathbf{Y}}$ and simplifying we see that $f_{\mathbf{Z}}$ has the same form as (12). □

Remark. Note that proposition 7 implies that given the value of V_k, the random vectors $(V_1,\ldots,V_{k-1})$ and $(V_{k+1},\ldots,V_n)$ are conditionally independent.

Definition. Let $X_1,\ldots,X_n$ denote i.i.d random variables having a common density function f. Such a collection is called a **random sample of size** n. For each $\omega\in\Omega$, arrange the sample values $X_1(\omega),\ldots,X_n(\omega)$ in non-decreasing order $X_{(1)}(\omega)\leq\ldots\leq X_(n)(\omega)$, where $(1),(2),\ldots,(n)$ is a random (i.e. depending on ω) permutation of $1,2,\ldots,n$. The new variables $X_{(1)},\ldots,X_(n)(\omega)$ are called the **order statistics** of the random sample. They are also denoted $X_1^*,\ldots,X_n^*$ and are referred to as the **order statistics of the random sample of size** n.

Proposition 8. *Write f for the common density of the X_i. Then $(X_1^*,\ldots,X_n^*)$ has density*

$$f_{\mathbf{X}^*} = n!\left[\prod_{i=1}^n f(x_i)\right]\mathbf{1}_{x_1<\ldots<x_n}$$

Proof. Consider the event

$$\mathbb{P}\left(\bigcap_{i=1}^n\{X_i^*\in[x_i,x_i+\Delta x_i)\}\right)=\mathbb{P}\left(\bigcup_\sigma\bigcap_{i=1}^n\{X_{\sigma_i}\in[x_i,x_i+\Delta x_i)\}\right)$$

where σ ranges over the permutations of $1,\ldots,n$. Note that the events associated with different permuations are disjoint. Hence,

$$\mathbb{P}\left(\bigcup_\sigma\bigcap_{i=1}^n\{X_{\sigma_i}\in[x_i,x_i+\Delta x_i)\}\right)=\sum_\sigma\mathbb{P}\left(\bigcap_{i=1}^n\{X_{\sigma_i}\in[x_i,x_i+\Delta x_i)\}\right)$$

since X_i are independent

$$\sum_\sigma\mathbb{P}\left(\bigcap_{i=1}^n\{X_{\sigma_i}\in[x_i,x_i+\Delta x_i)\}\right)=\sum_\sigma\prod_{i=1}^n\mathbb{P}\left(\{X_{\sigma_i}\in[x_i,x_i+\Delta x_i)\}\right).$$

Using the common density we get

$$\sum_{\sigma}\prod_{i=1}^{n}\mathbb{P}\left(\{X_{\sigma_i}\in[x_i,x_i+\Delta x_i)\}\right)\approx\sum_{\sigma}\prod_{i=1}^{n}f(x_i)\Delta x_i$$

since the summands are identical and σ takes over $n!$ values we finally get

$$\mathbb{P}\left(\bigcap_{i=1}^{n}\{X_i^*\in[x_i,x_i+\Delta x_i)\}\right)\approx n!\prod_{i=1}^{n}f(x_i)\Delta x_i.$$

□

Proposition 8 implies that if $U_1,\ldots,U_n$ are i.i.d. Uniform$(0,1)$. Then

$$f_{\mathbf{U}^*}(u_1,\ldots,u_n)=n!\mathbf{1}_{u_1<\ldots<u_n}$$

which is the ordered Dirichlet distribution with parameter $\alpha_1=\ldots=\alpha_{n+1}=1$. We can summarize as

Proposition 9. *Let $Y_1,\ldots,Y_{n+1}$ be independent and exponentially distributed with common parameter β and let $S_i=Y_1+\ldots+Y_i, 1\le i\le n+1$. Then $(U_1^*,\ldots,U_n^*)$ and $(S_1/S_{n+1},\ldots,S_n/S_{n+1})$ have the same distribution.*

Remark. Throughout our analysis of the Dirichlet distribution rarely have we used the rate parameter (denoted by β) of the involved gamma random variables. The reason is that a gamma(α,β) random variable X, is distributed the same as Y/β where Y is a gamma$(\alpha,1)$ random variable. To see why simply consider

$$f_X(x)=\frac{d}{dx}\mathbb{P}\left(X\le x\right)=\frac{d}{dx}\int_0^x\frac{\beta^\alpha}{\Gamma(\alpha)}t^{\alpha-1}e^{-\beta t}dt$$

performing the change of variable $u=\beta t$ we get

$$f_X(x)=\frac{d}{dx}\int_0^{\beta x}\frac{1}{\Gamma(\alpha)}u^{\alpha-1}e^{-u}du=\frac{d}{dx}\mathbb{P}\left(\frac{1}{\beta}Y\le x\right)=f_{Y/\beta}(x).$$

4 The Gaussian Distribution

The following fact is used in our analysis of Gaussian distributions:

Proposition (Cramer-Wold Device)**.** *If $\mathbf{X}=(X_1,\ldots,X_n)'$ is a d-dimensional random (column) vector, then the distribution of $\mathbf{X}$ is uniquely determined by the distribution of the random variables $\mathbf{t}'\mathbf{X}=\sum_{i=1}^d t_iX_i$ for all $\mathbf{t}\in\mathbb{R}^d$. That is if $\mathbf{X}=(X_1,\ldots,X_n)'$ and $\mathbf{Y}=(Y_1,\ldots,Y_n)'$ are d-dimensional random vectors and for any $\mathbf{t}\in\mathbb{R}^d$ variables $\mathbf{t}'\mathbf{X}$ and $\mathbf{t}'\mathbf{Y}$ are identically distributed, then $\mathbf{X}$ and $\mathbf{Y}$ are identically distributed.*

4.1 Mean Vector and Variance-Covariance Matrix

Suppose $\mathbf{X} = (X_1, \ldots, X_d)'$ is a random vector. Then, the mean vector and the variance-covariance matrix of $\mathbf{X}$ are given by

$$\mathbb{E}[\mathbf{X}] := \begin{bmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_d] \end{bmatrix} \qquad \mathrm{Cov}(\mathbf{X}, \mathbf{X})_{ij} := \mathrm{Cov}(X_i, X_j),$$

assuming the relevant expectations exist. Recall the definition of the covariance of two random variables X and Y

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

Note that if A is a $n \times d$ matrix and $\mathbf{Y} = A\mathbf{X}$, linearity of expectation implies that

$$\mathbb{E}[\mathbf{Y}]_i = \mathbb{E}[Y_i] = \mathbb{E}\left[\sum_{j=1}^{d} A_{ij} X_j\right] = \sum_{j=1}^{d} A_{ij} \mathbb{E}[X_j] = \mathbb{E}[A\mathbf{X}]_i.$$

Similarly, bilinearity of covariance implies that

$$\begin{aligned} \mathrm{Cov}(\mathbf{Y}, \mathbf{Y})_{ij} = \mathrm{Cov}(Y_i, Y_j) &= \mathrm{Cov}\left(\sum_{k=1}^{d} A_{ik} X_k, \sum_{l=1}^{d} A_{jl} X_l\right) \\ &= \sum_{k=1}^{d} \sum_{l=1}^{d} A_{ik} A_{jl} \mathrm{Cov}(X_k, X_l) \\ &= \sum_{l=1}^{d} \sum_{k=1}^{d} A_{ik} \mathrm{Cov}(X_k, X_l) A'_{lj} \\ &= (A \mathrm{Cov}(\mathbf{X}, \mathbf{X}) A')_{ij}. \end{aligned}$$

4.2 Preliminary Results from Linear Algebra

Let A be a real $d \times d$ matrix. Recall that A is said to be **symmetric** if $A_{ij} = A_{ji}$. If in addition to being symmetric, A has the property that for any non-zero vector $\mathbf{t} \in \mathbb{R}^d$ we have $\mathbf{t}' A \mathbf{t} > 0$, we say A is a **positive definite** matrix. Similarly, if $\mathbf{t}' A \mathbf{t} \geq 0$, we say A is **non-negative definite** (positive semi-definite). A is said to be **orthogonal** if $A^{-1} = A'$; that is, $AA' = A'A = I$. If for a given $\lambda \in \mathbb{C}$ the matrix $A - \lambda I$ is singular (i.e. $\det(A - \lambda I) = 0$), λ is said to be an **eigenvalue** of A. If a non-zero vector $\mathbf{x}$ satisfies $A\mathbf{x} = \lambda \mathbf{x}$ it is said to be an **eigenvector** of A associated with the eigenvalue λ.

Remark. We can see that the variance-covariance matrix of a given random vector is always non-negative definite. Symmetry is clear from definition:

$$\mathrm{Cov}(\mathbf{X}, \mathbf{X})_{ij} = \mathrm{Cov}(X_i, X_j) = \mathrm{Cov}(X_j, X_i) = \mathrm{Cov}(\mathbf{X}, \mathbf{X})_{ji}.$$

Also, note that for any $\mathbf{t} \in \mathbb{R}^d$ we have

$$\mathbf{t}'\mathrm{Cov}\left(\mathbf{X}, \mathbf{X}\right)\mathbf{t} = \mathrm{Cov}\left(\mathbf{t}'\mathbf{X}, \mathbf{t}'\mathbf{X}\right) = \mathrm{Var}\left(\mathbf{t}'\mathbf{X}\right) \geq 0$$

based on the bilinearity property of covariance.

Proposition 10 (The Spectral Theorem)**.** *Suppose A is a real, symmetric, and non-negative definite matrix.*

- *The eigenvalues of A are all real and non-negative, and the eigenvectors are real.*
- *If $\lambda_0 \geq 0$ is an eigenvalue of A, then the dimension of the subspace (eigenspace) $\{\mathbf{x} : A\mathbf{x} = \lambda_0\mathbf{x}\}$ is the multiplicity of λ_0 as a root of the characteristic polynomial $P(\lambda) = \det(A - \lambda I)$.*
- *Eigenvectors for different eigenvalues are orthogonal, and we can choose an orthonormal basis for each eigenspace.*
- *Let $\lambda_1 \geq \ldots \geq \lambda_d \geq 0$ be the eigenvalues of A repeated according to multiplicity with orthonormal eigenvectors $\mathbf{x}_1, \ldots, \mathbf{x}_d$. Set $B = (\mathbf{x}_1, \ldots, \mathbf{x}_d)$. Then B is orthogonal and*

$$B'AB = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_d \end{bmatrix} \iff A = B \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_d \end{bmatrix} B'$$

4.3 Normal Random Vectors

Recall that a random variable X is said to have a normal distribution if $X = \mu$ for some $\mu \in \mathbb{R}$ or it is absolutely continuous with density given by

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

For $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}^+$. In the latter case we have

$$\mathbb{E}\left[X\right] = \mu \qquad \mathrm{Var}\left(X\right) = \sigma^2.$$

Definition. A random vector $\mathbf{X}$ on $\mathbb{R}^d$ is said to have a **normal distribution on** $\mathbb{R}^d$ (or to have a **multivariate (d-variate) normal distribution**) iff $\mathbf{t}'\mathbf{X}$ has a normal distribution for each $\mathbf{t} \in \mathbb{R}^d$.

Proposition 11. *Let $\mathbf{X}$ have a normal distribution on $\mathbb{R}^d$. Then*

1. *The distribution of $\mathbf{X}$ is uniquely determined by its mean vector $\mathbb{E}\left[\mathbf{X}\right]$ and its covariance matrix $Cov\left(\mathbf{X}, \mathbf{X}\right)$.*
2. *Let A be any $m \times d$ matrix and $\gamma \in \mathbb{R}^m$. Then $\mathbf{Y} = A\mathbf{X} + \gamma$ has a normal distribution on $\mathbb{R}^m$ with mean vector $A\mu + \gamma$ and covariance matrix $ACov\left(\mathbf{X}, \mathbf{X}\right)A'$.*

Proof. 1. Take any $\mathbf{t} \in \mathbb{R}^d$. Since $\mathbf{X}$ is a normal random vector, we know that $\mathbf{t}'\mathbf{X}$ will have a normal distribution $\mathcal{N}(\mu^*, \sigma^{*2})$. By our discussion of the properties of the mean vector and covariance matrix we have

$$\mu^* = \mathbb{E}[\mathbf{t}'\mathbf{X}] = \mathbf{t}'\mathbb{E}[\mathbf{X}] \qquad \sigma^{*2} = \mathrm{Cov}(\mathbf{t}'\mathbf{X}, \mathbf{t}'\mathbf{X}) = \mathbf{t}'\mathrm{Cov}(\mathbf{X}, \mathbf{X})\,\mathbf{t}.$$

Therefore, the mean vector and the covariance matrix completely determine the distribution of $\mathbf{t}'\mathbf{X}$ for any $\mathbf{t} \in \mathbb{R}^d$. Employing the Cramer-Wold device we conclude that the mean vector and the covariance matrix completely determine the distribution of $\mathbf{X}$.

2. Note that

$$\mathbf{t}'\mathbf{Y} = \mathbf{t}'(A\mathbf{X} + \gamma) = \mathbf{t}'A\mathbf{X} + \mathbf{t}'\gamma = (A'\mathbf{t})'\mathbf{X} + \mathbf{t}'\gamma.$$

Since $\mathbf{X}$ is a normal random vector and $A'\mathbf{t} \in \mathbb{R}^d$ we conclude that $(A'\mathbf{t})'\mathbf{X}$ is distributed normally. Also, note that $\mathbf{t}'\gamma$ is a scalar so $(A'\mathbf{t})'\mathbf{X} + \mathbf{t}'\gamma$ has a normal distribution. Hence, $\mathbf{Y}$ is a normal random vector. The values of the parameters are results of the linearity properties of expectation and covariance.

□

Example 2. Let $Z_1, \ldots, Z_d$ be independent standard normal random variables (i.e. $\mathbb{E}[Z_i] = 0$ and $\mathrm{Var}(Z_i) = 1$). Put $\mathbf{Z} = (Z_1, \ldots, Z_d)'$. We have

$$f_{\mathbf{Z}}(z_1, \ldots, z_d) = \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi}} e^{z_i^2/2} = \frac{1}{(2\pi)^{d/2}} e^{(z_1^2 + \ldots + z_d^2)/2}$$

more compactly

$$f_{\mathbf{Z}}(\mathbf{z}) = \frac{1}{(2\pi)^{d/2}} e^{|\mathbf{z}|^2/2}$$

where $z \in \mathbb{R}^d$ and $|\cdot|$ denotes the Euclidean norm. Now, let U be any orthogonal $d \times d$ matrix and consider the random vector $\mathbf{W} = U\mathbf{Z}$. Applying the change of variables formula we have

$$f_{\mathbf{W}}(\mathbf{w}) = f_{\mathbf{Z}}(U^{-1}\mathbf{w}) |\det U^{-1}|.$$

Since U is orthogonal we have $|\det U^{-1}| = 1$. Using this fact and expanding the right-hand side above we get

$$f_{\mathbf{W}}(\mathbf{w}) = \frac{1}{(2\pi)^{d/2}} e^{|U^{-1}\mathbf{w}|^2/2},$$

but

$$|U^{-1}\mathbf{w}|^2 = \left(U^{-1}\mathbf{w}\right)' U^{-1}\mathbf{w} = (U'\mathbf{w})' U'\mathbf{w} = \mathbf{w}'UU'\mathbf{w} = \mathbf{w}'\mathbf{w} = |\mathbf{w}|^2.$$

Hence,

$$f_{\mathbf{W}}(\mathbf{w}) = \frac{1}{(2\pi)^{d/2}} e^{|\mathbf{w}|^2/2};$$

that is, $\mathbf{Z}$ and $\mathbf{W}$ are identically distributed. Note that no restrictions were placed on our choice of U, other than it being orthogonal. Hence, by picking any $\mathbf{u} \in \mathbb{R}^d$ with $|\mathbf{u}| = 1$ and expanding it to an orthonormal basis using the Gram-Schmidt procedure, we can acquire a new U that has $\mathbf{u}'$ as its first row. Hence, we can see that $\mathbf{u}'\mathbf{Z}$ for any $\mathbf{u} \in \mathbb{R}^d$ with $|\mathbf{u}| = 1$ is distributed according to $\mathcal{N}(0, 1)$. Noting this, we observe that for any non-zero $\mathbf{t} \in \mathbb{R}^d$ we can write

$$\mathbf{t}'\mathbf{Z} = |\mathbf{t}| \left(\frac{1}{|\mathbf{t}|}\mathbf{t} \right)' \mathbf{Z}$$

where

$$\left(\frac{1}{|\mathbf{t}|}\mathbf{t} \right)' \mathbf{Z} \sim \mathcal{N}(0, 1).$$

so

$$\mathbf{t}'\mathbf{Z} = |\mathbf{t}| \left(\frac{1}{|\mathbf{t}|}\mathbf{t} \right)' \mathbf{Z} \sim \mathcal{N}(0, |\mathbf{t}|^2).$$

Therefore, we conclude that $\mathbf{Z}$ has the normal distribution on $\mathbb{R}^d$ with parameters

$$\mathbb{E}[\mathbf{Z}] = 0 \qquad \mathrm{Cov}(\mathbf{X}, \mathbf{X}) = I.$$

Proposition 12. *Let $\mathbf{X}$ have a normal distribution on $\mathbb{R}^d$ with mean μ and covariance matrix Σ. Assume Σ has rank r. Then there is a diagonal matrix D whose first r diagonal entries are positive, and the rest 0, and an orthogonal matrix B such that $\Sigma = B'DB$ and $\mathbf{X}$ is distributed like $B'\sqrt{D}\mathbf{Z} + \mu$ where $\mathbf{Z}$ has the standard normal distribution on $\mathbb{R}^d$.*

Proof. Note that by proposition 11, $B\sqrt{D}\mathbf{Z} + \mu$ is normally distributed on $\mathbb{R}^d$. Also, note that

$$\begin{aligned} \mathrm{Cov}\left(B'\sqrt{D}\mathbf{Z}, B'\sqrt{D}\mathbf{Z} \right) &= B'\sqrt{D}\mathrm{Cov}(\mathbf{Z}, \mathbf{Z})(B'\sqrt{D})' = B'\sqrt{D}I\sqrt{D}'B \\ &= B'DB = \Sigma. \end{aligned}$$

and

$$\mathbb{E}\left[B'\sqrt{D}\mathbf{Z} + \mu \right] = B'\sqrt{D}\mathbb{E}[\mathbf{Z}] + \mu = \mu.$$

Since the mean vector and the covariance matrix completely determine the distribution of a normal random vector, we conclude that $\mathbf{X}$ and $B'\sqrt{D}\mathbf{Z} + \mu$ are identically distributed. □

Remark. Note that only the first r rows of the matrix D are non-zero, hence we can actually represent $\mathbf{X}$ using a standard normal variable on $\mathbb{R}^r$.

We have the following immediate corollary

Proposition 13. *Let $\mu \in \mathbb{R}^d$ and let Σ be a $d \times d$ non-negative definite matrix. Then there is a multivariate normal random vector on $\mathbb{R}^d$ having mean μ and covariance matrix Σ.*

Proof. Since Σ is a non-negative definite matrix, by the spectral theorem, we can find $d \times d$ matrices B and D where B is orthogonal and D is diagonal with non-negative entries where $\Sigma = BDB'$. Let $\mathbf{Z}$ be a standard normal random vector on $\mathbb{R}^d$ and consider $\mathbf{X} = B\sqrt{D}\mathbf{Z} + \mu$. By proposition 11, $\mathbf{X}$ is normally distributed with mean μ and covariance

$$B\sqrt{D}I(B\sqrt{D})' = B\sqrt{D}I\sqrt{D}'B' = B\sqrt{D}\sqrt{D}B' = BDB' = \Sigma.$$

□

Definition 3. A normal distribution on $\mathbb{R}^d$ is called **nonsingular** iff its covariance matrix is nonsingular. Otherwise it is called **singular**. The **rank** of a singular distribution is the rank of its covariance matrix.

Proposition 14. *Let* $\mathbf{X}$ *have a normal distribution on* $\mathbb{R}^d$ *with mean* μ *and covariance matrix* Σ*. Then* $\mathbf{X}$ *has a nonsingular distribution on* $\mathbb{R}^d$ *iff* $\mathbf{X}$ *has an absolutely continuous distribution on* $\mathbb{R}^d$*. In that case the density of* $\mathbf{X}$ *is*

$$(2\pi)^{-d/2} (\det \Sigma)^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu)'\Sigma^{-1}(\mathbf{x}-\mu)\right].$$

Proof. First, suppose $\mathbf{X}$ is nonsingular; i.e. Σ is nonsingular. By proposition 13 we know that $\mathbf{X} = B\sqrt{D}\mathbf{Z} + \mu$ where $\mathbf{Z}$ is a standard normal random vector in $\mathbb{R}^d$. We can apply the change of variable formula:

$$f_{\mathbf{X}}(\mathbf{x}) = f_{\mathbf{Z}}((B\sqrt{D})^{-1}(\mathbf{x}-\mu)) \cdot \left|\det\left((B\sqrt{D})^{-1}\right)\right|. \tag{13}$$

Because Σ is positive semi-definite and nonsingular, the diagonal entries of D are positive; in addition, B is orthogonal. Hence,

$$(B\sqrt{D})^{-1} = \sqrt{D}^{-1}B'$$

and

$$\left|\det\left((B\sqrt{D})^{-1}\right)\right| = \left|\det(B)\det\left(\sqrt{D})^{-1}\right)\right| = |\det(B)|\frac{1}{\sqrt{\det D}}.$$

The last equality holds since D is diagonal with positive entries. Also, noting that B is orthogonal (hence $|\det B| = 1$) and that

$$\det\Sigma = \det(BDB') = \det B \det D \det B' = \det D,$$

we conclude that

$$\left|\det\left((B\sqrt{D})^{-1}\right)\right| = (\det\Sigma)^{-1/2}. \tag{14}$$

Using the density of the standard multivariate normal distribution given in example 2 we can combine equations (13) and (14).

$$f_{\mathbf{X}}(\mathbf{x}) = (2\pi)^{-d/2}\exp\left[-\frac{1}{2}\left|(B\sqrt{D})^{-1}(\mathbf{x}-\mu)\right|^2\right] \cdot (\det\Sigma)^{-1/2}. \tag{15}$$

We can simplify the argument of the exponential:

$$\begin{aligned}\left|(B\sqrt{D})^{-1}(\mathbf{x}-\mu)\right|^2 &= \left((B\sqrt{D})^{-1}(\mathbf{x}-\mu)\right)'\left((B\sqrt{D})^{-1}(\mathbf{x}-\mu)\right)\\ &= \left(D^{-1/2}B'(\mathbf{x}-\mu)\right)'\left(D^{-1/2}B'(\mathbf{x}-\mu)\right)\\ &= (\mathbf{x}-\mu)'BD^{-1/2}D^{-1/2}B(\mathbf{x}-\mu)\\ &= (\mathbf{x}-\mu)'BD^{-1}B(\mathbf{x}-\mu)\\ &= (\mathbf{x}-\mu)'\Sigma^{-1}(\mathbf{x}-\mu).\end{aligned}$$

The last equality holds because

$$\Sigma = BDB' \implies \Sigma^{-1} = (B')^{-1}D^{-1}B^{-1} = (B')'D^{-1}B' = BD^{-1}B'.$$

Substituting back in (15) we get our ultimate density formula

$$f_{\mathbf{X}}(\mathbf{x}) = (2\pi)^{-d/2}(\det\Sigma)^{-1/2}.\exp\left[-\frac{1}{2}(\mathbf{x}-\mu)'\Sigma^{-1}(\mathbf{x}-\mu)\right].$$

Now suppose Σ is singular and assume -for contradiction- that it has an absolutely continuous distribution. Consider the affine transformation $T:\mathbb{R}^d\to\mathbb{R}^d$ with the rule $T(\mathbf{x}) = B'\mathbf{x} - B'\mu$ and recall that $\mathbf{X}$ can be written as $B\sqrt{D}\mathbf{Z}+\mu$ where $\mathbf{Z}$ is a standard normal random vector. Let $\mathbf{Y} := T(\mathbf{X})$. We have

$$\mathbf{Y} := T(\mathbf{X}) = T(B\sqrt{D}\mathbf{Z}+\mu) = B'B\sqrt{D}\mathbf{Z} + B'\mu - B'\mu.$$

Noting that B is orthogonal and hence $B'B = I$ we can conclude that

$$\therefore \mathbf{Y} = \sqrt{D}\mathbf{Z}.$$

Since Σ is singular it has zero as an eigenvalue and hence D must have a zero row which causes the corresponding element of $\mathbf{Y}$, say Y_j, to be a constant random variable taking on 0 with probability 1. On the other hand, since we have assumer $\mathbf{X}$ is absolutely continuous and any affine map is continuously differentiable, $\mathbf{Y}$ must have an absolutely continuous distribution; moreover, our discussion of marginal densities shows that every one of the random variables forming $\mathbf{Y}$ has an absolutely continuous distribution. Specifically, Y_j has to have an absolutely continuous distribution. But we know that $\mathbb{P}(Y_j \in \{0\}) = 1$ while the set $\{0\}$ has Lebesgue measure zero. Hence Y_j cannot have an absolutely continuous distribution. Contradiction. Therefore, $\mathbf{X}$ cannot have an absolutely continuous distribution. □

The following result is helpful in establishing an important property of independent normal random vectors

Proposition. *Suppose $\Gamma \in \mathbb{R}^{k\times k}$ and $\Xi \in \mathbb{R}^{(d-k)\times(d-k)}$ are positive semi-definite matrices so that*

$$\Sigma = \begin{bmatrix}\Gamma & 0\\ 0 & \Xi\end{bmatrix}$$

is positive semi-definite. Take any $\mu \in \mathbb{R}^d$ and let $\mathbf{X}$ be a d-dimensional normal random vector with mean vector μ and covariance matrix Σ. If

$$\begin{bmatrix} \tilde{\mathbf{X}} \\ \tilde{\mathbf{Y}} \end{bmatrix} := \mathbf{X} \qquad \tilde{\mathbf{X}} \in \mathbb{R}^k, \tilde{\mathbf{Y}} \in \mathbb{R}^{d-k},$$

and

$$\begin{bmatrix} \nu \\ \rho \end{bmatrix} := \mu \qquad \nu \in \mathbb{R}^k, \rho \in \mathbb{R}^{d-k},$$

then $\tilde{\mathbf{X}} \sim \mathcal{N}(\nu, \Gamma)$, $\tilde{\mathbf{Y}} \sim \mathcal{N}(\rho, \Xi)$, and $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$ are independent.

Proof. First take any $\mathbf{t} \in \mathbb{R}^d$ and let $\mathbf{t} = [\mathbf{t_1}' \ \mathbf{t_2}']'$ where $\mathbf{t_1} \in \mathbb{R}^k$ and $\mathbf{t_2} \in \mathbb{R}^{d-k}$. Then

$$\mathbf{t}'\Sigma\mathbf{t} = \begin{bmatrix} \mathbf{t_1}' & \mathbf{t_2}' \end{bmatrix} \begin{bmatrix} \Gamma & 0 \\ 0 & \Xi \end{bmatrix} \begin{bmatrix} \mathbf{t_1} \\ \mathbf{t_2} \end{bmatrix} = \mathbf{t_1}'\Gamma\mathbf{t_1} + \mathbf{t_2}'\Xi\mathbf{t_2}$$

since both Γ and Ξ are positive semi-definite, the right-hand side must be non-negative, which renders Σ positive semi-definite.

Since both Γ and Ξ are symmetric, we can find orthogonal matrices B and C and diagonal matrices D and E of appropriate size so that

$$\Gamma = BDB' \qquad \Xi = CEC'.$$

Note that

$$\begin{bmatrix} B & 0 \\ 0 & C \end{bmatrix} \begin{bmatrix} B' & 0 \\ 0 & C' \end{bmatrix} = \begin{bmatrix} BB' & 0 \\ 0 & CC' \end{bmatrix} = I,$$

and also

$$\begin{bmatrix} B & 0 \\ 0 & C \end{bmatrix} \begin{bmatrix} D & 0 \\ 0 & E \end{bmatrix} \begin{bmatrix} B' & 0 \\ 0 & C' \end{bmatrix} = \begin{bmatrix} BDB' & 0 \\ 0 & CEC' \end{bmatrix} = \begin{bmatrix} \Gamma & 0 \\ 0 & \Xi \end{bmatrix} = \Sigma.$$

So we have

$$\mathbf{X} = \begin{bmatrix} B & 0 \\ 0 & C \end{bmatrix} \begin{bmatrix} \sqrt{D} & 0 \\ 0 & \sqrt{E} \end{bmatrix} \mathbf{Z},$$

where $\mathbf{Z} = [Z_1 \ \dots \ Z_d]'$ is a standard normal random vector on $\mathbb{R}^d$. Namely,

$$\mathbf{X} = \begin{bmatrix} B\sqrt{D} & 0 \\ 0 & C\sqrt{E} \end{bmatrix} \mathbf{Z}$$

which translates to

$$\tilde{\mathbf{X}} = B\sqrt{D} \begin{bmatrix} Z_1 \\ \vdots \\ Z_k \end{bmatrix} \qquad \tilde{\mathbf{Y}} = C\sqrt{E} \begin{bmatrix} Z_{k+1} \\ \vdots \\ Z_d \end{bmatrix}.$$

Since Z_is are independent and $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$ are composed of linear combinations of disjoint sets of Z_i, we conclude that they are independent. □

We can now easily establish the following result:

Proposition 15. *Let* $\mathbf{Y}$ *have a normal distribution on* $\mathbb{R}^d$ *and let* A *be any* $m \times d$ *and* B *any* $n \times d$ *matrix. Then* $A\mathbf{Y}$ *and* $B\mathbf{Y}$ *are independent iff they are uncorrelated. This is the case iff* $A\mathrm{Cov}(\mathbf{Y}, \mathbf{Y})\, B' = 0$.

Proof. Let

$$\mathbf{X} = \begin{bmatrix} \mathbf{Y} \\ \mathbf{Y} \end{bmatrix}.$$

It is easy to see that $\mathbf{X}$ is normally distributed on $\mathbb{R}^{2d}$, because for any $\mathbf{t} \in \mathbb{R}^{2d}$ we can write find $\mathbf{t_1}, \mathbf{t_2} \in \mathbb{R}^d$ so that $\mathbf{t} = [\mathbf{t_1}'\ \mathbf{t_2}']'$. Then,

$$\mathbf{t}'\mathbf{X} = \begin{bmatrix} \mathbf{t_1}' & \mathbf{t_2}' \end{bmatrix} \begin{bmatrix} \mathbf{Y} \\ \mathbf{Y} \end{bmatrix} = \mathbf{t_1}'\mathbf{Y} + \mathbf{t_2}'\mathbf{Y},$$

and since both $\mathbf{t_1}'\mathbf{Y}$ and $\mathbf{t_2}'\mathbf{Y}$ are normal random variables, and the sum of two normal random variable is another normal random variable, we conclude that $\mathbf{X}$ has a normal distribution. It is easy to see, by applying the relevant definitions, that the mean vector and the covariance matrix of $\mathbf{X}$ are given by

$$\mu_{\mathbf{X}} = \begin{bmatrix} \mu \\ \mu \end{bmatrix} \qquad \Sigma_{\mathbf{X}} = \begin{bmatrix} \Sigma & \Sigma \\ \Sigma & \Sigma \end{bmatrix}.$$

Now, let

$$\mathbf{W} = \begin{bmatrix} A\mathbf{Y} \\ B\mathbf{Y} \end{bmatrix} = \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix} \mathbf{X},$$

and note that the covariance matrix of $\mathbf{W}$ is given by

$$\begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix} \begin{bmatrix} \Sigma & \Sigma \\ \Sigma & \Sigma \end{bmatrix} \begin{bmatrix} A' & 0 \\ 0 & B' \end{bmatrix} = \begin{bmatrix} A\Sigma A' & A\Sigma B' \\ B\Sigma A' & B\Sigma B' \end{bmatrix}.$$

It is easy to see that since Σ is positive semi-definite, $A\Sigma A'$ and $B\Sigma B'$ are positive semi-definite. Applying the previous preposition we can see that if $A\Sigma B' = 0$, then $A\mathbf{Y}$ and $B\mathbf{Y}$ are independent. Also, since independent random variables are uncorrelated we can see that

$$A\mathbf{Y} \perp\!\!\!\perp B\mathbf{Y} \implies \mathrm{Cov}\,(A\mathbf{Y}, B\mathbf{Y}) = 0 \implies A\mathrm{Cov}\,(\mathbf{Y}, \mathbf{Y})\, B' = 0. \implies A\Sigma B' = 0.$$

□

Example 3. Let $X_1, \ldots, X_n$ be independent and normally distributed, $\mathbb{E}[X_i] = \mu_i$ and $\mathrm{Var}(X_i) = \sigma^2$. Show $\overline{X}_n = (1/n)\sum X_i$ and $\mathbf{X} - \mathbf{1}\overline{X}_n$ are independent, where $\mathbf{1} = (1, \ldots, 1)'$ and $\mathbf{X} = (X_1, \ldots, X_n)'$ and find the distribution of $\mathbf{X} - \mathbf{1}\overline{X}_n$.

Solution. We have

$$\overline{X}_n = \frac{1}{n}\mathbf{1}'\mathbf{X}$$

and

$$\mathbf{X} - \mathbf{1}\overline{X}_n = (I - \frac{1}{n}\mathbf{1}\mathbf{1}')\mathbf{X}. \tag{16}$$

Noting that

$$\mathrm{Cov}\,(\mathbf{X}, \mathbf{X}) = \begin{bmatrix} \sigma^2 & & \\ & \ddots & \\ & & \sigma^2 \end{bmatrix} = \sigma^2 I.$$

we have

$$\frac{1}{n}\mathbf{1}'\sigma^2 I(I - \frac{1}{n}\mathbf{1}\mathbf{1}') = \frac{\sigma^2}{n}\mathbf{1}'(I - \frac{1}{n}\mathbf{1}\mathbf{1}') = \frac{\sigma^2}{n}(\mathbf{1}' - \frac{1}{n}\mathbf{1}'\mathbf{1}\mathbf{1}') = 0.$$

Utilizing proposition 15 we can see that $\overline{X}_n$ and $\mathbf{X} - \mathbf{1}\overline{X}_n$ are independent. Equation (16) also shows that $\mathbf{X} - \mathbf{1}\overline{X}_n$ is normally distributed since example 2 implies that $\mathbf{X}$ is normally distributed. Hence, noting that

$$\overline{\mu} := \mathbb{E}\left[\overline{X}_n\right] = \frac{1}{n}\sum_i \mu_i,$$

we can easily find the mean vector

$$\mathbb{E}\left[\mathbf{X} - \mathbf{1}\overline{X}_n\right] = \mathbb{E}\left[\mathbf{X}\right] - \mathbb{E}\left[\mathbf{1}\overline{X}_n\right] = \begin{bmatrix} \mu_1 - \overline{\mu} \\ \vdots \\ \mu_n - \overline{\mu} \end{bmatrix}.$$

Also, applying (16) and the bilinearity of covariance we can find the covariance by observing that

$$\mathrm{Cov}\left(\mathbf{X} - \mathbf{1}\overline{X}_n, \mathbf{X} - \mathbf{1}\overline{X}_n\right) = (I - \frac{1}{n}\mathbf{1}\mathbf{1}')\mathrm{Cov}\,(\mathbf{X}, \mathbf{X})\,(I - \frac{1}{n}\mathbf{1}\mathbf{1}')' = \sigma^2(I - \frac{1}{n}\mathbf{1}\mathbf{1}')^2.$$

The last equality results from $I - \frac{1}{n}\mathbf{1}\mathbf{1}'$ being a symmetric matrix and $\mathrm{Cov}\,(\mathbf{X}, \mathbf{X}) = \sigma^2 I$. Simplifying we can see

$$(I - \frac{1}{n}\mathbf{1}\mathbf{1}')^2 = I - \frac{1}{n}\mathbf{1}\mathbf{1}' - \frac{1}{n}\mathbf{1}\mathbf{1}' + \frac{1}{n^2}(\mathbf{1}\mathbf{1}')^2 = I - \frac{1}{n}\mathbf{1}\mathbf{1}' - \frac{1}{n}\mathbf{1}\mathbf{1}' + \frac{n}{n^2}\mathbf{1}\mathbf{1}' = I - \frac{1}{n}\mathbf{1}\mathbf{1}'.$$

Ultimately,

$$\mathrm{Cov}\left(\mathbf{X} - \mathbf{1}\overline{X}_n, \mathbf{X} - \mathbf{1}\overline{X}_n\right) = \sigma^2\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}'\right).$$

Proposition 16 (Conditioning)**.** *Let* $\mathbf{X}$ *be normally distributed on* $\mathbb{R}^d$*. Let* A *be any* $r \times d$ *matrix and* B *any* $m \times d$ *matrix, and let* $\mathbf{Y} = A\mathbf{X}$ *and* $\mathbf{W} = B\mathbf{X}$*. Assume that* $\mathbf{W}$ *has a non-singular distribution. Then there is a unique* $r \times m$ *matrix* C *such that*

$$Cov(\mathbf{Y} - C\mathbf{W}, \mathbf{W}) = 0,$$

namely,

$$C = Cov(\mathbf{Y}, \mathbf{W})\, Cov(\mathbf{W}, \mathbf{W})^{-1},$$

and the conditional distribution of $\mathbf{Y}$ *given* $\mathbf{W} = \mathbf{w}$ *is normal with mean*

$$\mathbb{E}[\mathbf{Y}] - C(\mathbb{E}[\mathbf{W}] - \mathbf{w})$$

and covariance

$$Cov(\mathbf{Y}, \mathbf{Y}) - CCov(\mathbf{W}, \mathbf{W}).$$

Proof. By bilinearity of covariance

$$\begin{aligned} \operatorname{Cov}(\mathbf{Y} - C\mathbf{W}, \mathbf{W}) = 0 &\iff \operatorname{Cov}(\mathbf{Y}, \mathbf{W}) - C\operatorname{Cov}(\mathbf{W}, \mathbf{W}) = 0 \\ &\iff C\operatorname{Cov}(\mathbf{W}, \mathbf{W}) = \operatorname{Cov}(\mathbf{Y}, \mathbf{W}) \end{aligned}$$

given that $\mathbf{W}$ has a non-singular distribution, the last equation is equivalent to

$$C = \operatorname{Cov}(\mathbf{Y}, \mathbf{W}) \operatorname{Cov}(\mathbf{W}, \mathbf{W})^{-1}.$$

Noting that

$$\mathbf{Y} = (\mathbf{Y} - C\mathbf{W}) + C\mathbf{W} \implies (\mathbf{Y} - C\mathbf{W}) = \mathbf{Y} - C\mathbf{W}$$

so given $\mathbf{W} = w$, $\mathbf{Y} - C\mathbf{W}$ is distributed like $\mathbf{Y} - C\mathbf{w}$ which is an affine transformation of $\mathbf{X}$. Hence $\mathbf{Y}$ given $\mathbf{W} = w$ is normally distributed. We can easily see that

$$\mathbb{E}[\mathbf{Y}|\mathbf{W} = \mathbf{w}] = \mathbb{E}[(\mathbf{Y} - C\mathbf{W}) + C\mathbf{w}] = \mathbb{E}[\mathbf{Y}] - C(\mathbb{E}[\mathbf{W}] - \mathbf{w}).$$

Also, note that

$$\operatorname{Cov}((\mathbf{Y} - C\mathbf{W}) + C\mathbf{w}, (\mathbf{Y} - C\mathbf{W}) + C\mathbf{w}) = \operatorname{Cov}(\mathbf{Y} - C\mathbf{W}, \mathbf{Y} - C\mathbf{W})$$

Since $C\mathbf{w}$ is constant almost everywhere. In addition,

$$\operatorname{Cov}(\mathbf{Y} - C\mathbf{W}, \mathbf{Y} - C\mathbf{W}) = \operatorname{Cov}(\mathbf{Y} - C\mathbf{W}, \mathbf{Y}) - \operatorname{Cov}(\mathbf{Y} - C\mathbf{W}, \mathbf{W}) C'$$

and since the second term on the right-hand side is zero we conclude that

$$\operatorname{Cov}(\mathbf{Y}, \mathbf{Y})_{|\mathbf{W}=\mathbf{w}} = \operatorname{Cov}(\mathbf{Y} - C\mathbf{W}, \mathbf{Y}) = \operatorname{Cov}(\mathbf{Y}, \mathbf{Y}) - C\operatorname{Cov}(\mathbf{W}, \mathbf{Y}).$$

□

5 Sampling From a Distribution

Recall that if X is a random variable, then the **distribution function** of X (also known as the **cumulative distribution function (CDF)** of X) is a map $F_X : \mathbb{R} \to [0, 1]$ given by $F_X(x) = \mathbb{P}(X \leq x)$. F_X has the following properties:

Monotonicity F_X is non-decreasing.

Right-continuity $\lim_{y \downarrow x} F_X(y) = F_X(x)$.

First-type Left-discontinuity $F_X(x-) := \lim_{y \uparrow x} F_X(y)$ exists.

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

Remark. Note the similarity between these properties and the properties of probability measures discussed in section 1.1. Proofs are direct applications of the corresponding properties of probability measures. Also note that non-negativity, monotonicity, and right-continuity of F_X guarantee the existence of the Stieltjes measure associated with F_X. The normalization property of F_X also establishes that this measure is a probability measure on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$. Lastly, note that the fact that discontinuities of F_X are of the first kind results from its monotonicity. Hence, we can conclude that monotonicity, right-continuity, and normalization completely identify cumulative distribution functions.

We also have the following:

Proposition. *If F is a distribution function, then for some $0 \leq p \leq 1$ we have $F = pF_c + (1-p)F_d$ where F_c is a continuous function and F_d only increases in jumps.*

Proposition 17 (Inverse Transform Technique)**.** *Let F be a distribution function and let U be uniformly distributed on the unit interval. Define*

$$F^{-1}(u) = \inf\{x \in \mathbb{R}; u \leq F(x)\}.$$

Then,

1. *If F is a continuous function, the random variable $X = F^{-1}(U)$ has distribution function F.*
2. *Let F be the distribution function of a random variable taking non-negative integer values. The random variable X given by*

$$X = k \iff F(k-1) < U \leq F(k)$$

has distribution function F.

Example 4 (Stochastic Ordering)**.** If X and Y are random variables such that

$$F_X(x) \leq F_Y(x) \iff \mathbb{P}(X \leq x) \leq \mathbb{P}(Y \leq x) \iff \mathbb{P}(X > x) > \mathbb{P}(Y > x),$$

then we say X **dominates** Y **stochastically** and we write $X \geq_{\text{st}} Y$.

Remark. Note that X and Y need not be defined on the same probability space since the above definition is only concerned with their distribution functions.

Proposition 18. *Suppose that $X \geq_{st} Y$. There exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and two random variables X' and Y' on this space such that:*

1. *X and X' have the same distribution.*
2. *Y and Y' have the same distribution.*

3. $\mathbb{P}(X' \geq Y') = 1$.

Proof. Take $(\omega, \mathcal{F}, \mathbb{P})$ to be $([0,1], \mathscr{B}([0,1], \mu)$ where μ is the Lebesgue measure on $\mathbb{R}$. For any distribution function F define the random variable Z_F as

$$Z_F(\omega) = \inf\{x | \omega \leq F(x)\}.$$

Note that

$$x \leq Z_F(\omega) \iff F(x) \leq \omega$$

therefore

$$\mathbb{P}(Z_F \leq z) = \mathbb{P}([0, F(z)]) = F(z)$$

so Z_F has distribution function F. Now let $X' = Z_{F_X}$ and $Y' = Z_{F_Y}$, and note that

$$X'(\omega) = \inf\{x | \omega \leq F_X(x)\}$$

and since $F_X(x) \leq F_Y(x)$ we have

$$\{x | \omega \leq F_X(x)\} \subseteq \{x | \omega \leq F_Y(x)\}$$

so

$$X'(\omega) = \inf\{x | \omega \leq F_X(x)\} \leq \inf\{x | \omega \leq F_Y(x)\} = Y'(\omega).$$

Hence the event $X' \leq Y'$ consists of the entire sample space and, therefore, has probability one. □

6 Poisson Processes

Recall the definition of the Poisson distribution:

Definition. Let X be a random variable whose support is the set of non-negative integers. X is said to have the Poisson distribution with parameter $\lambda > 0$ if

$$\mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}.$$

We also make use of the following important property of Poisson random variables

Proposition (Additivity of independent Poisson random variables)**.** *If* $X \sim \text{Poisson}(\lambda)$*,* $Y \sim \text{Poisson}(\mu)$ *and* $X \perp\!\!\!\perp Y$ *then* $X + Y \sim \text{Poisson}(\lambda + \mu)$*.*

Proof.

$$\begin{aligned}
\mathbb{P}(X+Y=k) &= \sum_{j=0}^{\infty} \mathbb{P}(X+Y=k|Y=j)\,\mathbb{P}(Y=j) \\
&= \sum_{j=0}^{k} \mathbb{P}(X=k-j)\,\mathbb{P}(Y=j) \\
&= \sum_{j=0}^{k} \frac{1}{(k-j)!j!}\mu^j\lambda^{k-j}e^{-\lambda-\mu} \\
&= \frac{e^{-\lambda-\mu}}{k!}\sum_{j=0}^{k}\frac{k!}{j!(k-j)!}\mu^j\lambda^{k-j} \\
&= \frac{(\lambda+\mu)^k}{k!}e^{-\lambda-\mu}.
\end{aligned}$$

□

Consider a large number of independent events each having a small probability. The number of such events which actually occur has a distribution which is close to a Poisson distribution. For instance, suppose there are n such events $X_1, \ldots X_n$, and each event has success probability λ/n -i.e. $X_i \sim$ Bernoulli(λ/n). The number of total events occurring is given by $S = X_1 + \ldots + X_n$ with distribution Binomial$(n, \lambda/n)$. As n grows, the distribution of S converges point-wise to a Poisson distribution with parameter λ. The reason is

$$\begin{aligned}
\mathbb{P}(S=k) &= \binom{n}{k}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k} \\
&= \frac{\lambda^k}{k!}\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-k}\frac{n(n-1)\ldots(n-k+1)}{n^k},
\end{aligned}$$

and we have

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{-k} = 1, \qquad \lim_{n\to\infty}\frac{n(n-1)\ldots(n-k+1)}{n^k} = 1,$$

and,

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n = \lim_{n\to\infty} e^{n\ln\left(1-\frac{\lambda}{n}\right)} = \lim_{n\to\infty} e^{n\left(-\frac{\lambda}{n}+o(1/n)\right)} = e^{-\lambda}$$

so

$$\mathbb{P}(S=k) \to \frac{\lambda^k}{k!}e^{-\lambda} \qquad n\to\infty.$$

It is natural to ask whether a similar result holds if X_i had different success probabilities. Say X_i are independent and $X_i \sim$ Bernoulli(p_i) where p_i are "small." Our goal is to find a Poisson random variable $P \sim$ Poisson(λ) that

suitably approximates S. It is reasonable to expect that P's mean matches that of S. This helps us determine the parameter λ:

$$\lambda = \mathbb{E}[P] = \mathbb{E}[S] = p_1 + \ldots + p_n.$$

But note that in this case

$$\mathrm{Var}(P) = \lambda$$

while

$$\mathrm{Var}(S) = \sum p_i(1-p_i) = \sum p_i - \sum p_i^2$$

so we expect the quantity $\sum p_i^2$ to be small if we are to have an accurate approximation.

In order to further analyse the notion of convergence among distributions we need to specify an appropriate topology on the set of distributions.

Definition 4. Let F and G be the distribution functions of discrete distributions which place masses f_n and g_n at the points x_n, for $n \geq 1$, and define

$$d_{TV}(F,G) = \sum_{k\geq 1} |f_k - g_k|$$

The quantity d_{TV} is called the **total variation distance** between F and G. For random variables X and Y, we define $d_{TV}(X,Y) = d_{TV}(F_X, F_Y)$.

Remark. It is easy to verify that d_{TV} is a metric on the space of distribution functions of integer-valued random variables.

We can now give a formal result regarding our motivating remarks above:

Proposition 19. *Let $\{X_r : 1 \leq r \leq n\}$ be independent Bernoulli radnom variables with respective parameters $\{p_r : 1 \leq r \leq n\}$, and let $S = \sum_{r=1}^n X_r$. Then*

$$d_{TV}(S,P) \leq 2\sum_{r=1}^n p_r^2$$

where P is a random variable having the Poisson distribution with parameter $\lambda = \sum_{r=1}^n p_r$.

Proof. Note that by the law of total probability

$$\begin{aligned}
|\mathbb{P}(S=k) - \mathbb{P}(P=k)| = \quad & |\mathbb{P}(S=k, S=P) + \mathbb{P}(S=k, S\neq P) \\
& - \mathbb{P}(P=k, S=P) - \mathbb{P}(P=k, S\neq P)| \\
= {} & |\mathbb{P}(S=k, S\neq P) - \mathbb{P}(P=k, S\neq P)| \\
\leq {} & \mathbb{P}(S=k, S\neq P) + \mathbb{P}(P=k, S\neq P)
\end{aligned}$$

so

$$\begin{aligned}
\sum_{k=0}^{\infty} |\mathbb{P}(S=k) - \mathbb{P}(P=k)| &\leq \sum_{k=0}^{\infty} \mathbb{P}(S=k, S\neq P) + \sum_{k=0}^{\infty} \mathbb{P}(P=k, S\neq P) \\
&\leq 2\mathbb{P}(S\neq P).
\end{aligned}$$

Therefore, it is enough to find a coupling between S and P that gives the desired form to $\mathbb{P}(S \neq P)$. We construct the coupling as follows. Let (X_r, Y_r) for $1 \leq r \leq n$ be a sequence of independent pairs, where the pair (X_r, Y_r) takes values in $\{0,1\} \times \{0,1,2,\ldots\}$ with mass function

$$\mathbb{P}(X_r = x, Y_r = y) = \begin{cases} 1 - p_r & x = y = 0 \\ e^{-p_r} - 1 + p_r & x = 1, y = 0 \\ \frac{p_r^y}{y!} e^{-p_r} & x = 1, y > 0 \end{cases}.$$

We have

1. Clearly $1 - p_r \geq 0$, and $p_r^y / y! e^{-p_r} \geq 0$ also note that $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = e^{-x} - 1 + x$ is non-decreasing for $x \geq 0$ and $f(0) = 0$ so $e^{-p_r} - 1 + p_r \geq 0$.

2. $$1 - p_r + e^{-p_r} - 1 + p_r + \sum_{y=1}^{\infty} \frac{p_r^y}{y!} e^{-p_r} = e^{-p_r} + e^{-p_r}\left(e^{p_r} - 1\right) = 1.$$

3. $X_r \sim \text{Bernoulli}(p_r)$.

4. $Y_r \sim \text{Poisson}(p_r)$.

The first two points establish that the given function is indeed a legitimate probability mass function. Now we can let

$$S = \sum_{r=1}^{n} X_r \qquad P = \sum_{r=1}^{n} Y_r.$$

By additivity of independent Poisson random variables Y has a Poisson distribution with parameter $\lambda = \sum_{r=1}^{n} p_r$. We have

$$\begin{aligned} \mathbb{P}(S \neq P) \leq \mathbb{P}\left(\bigcup_{r=1}^{n} \{X_r \neq Y_r\}\right) &\leq \sum_{r=1}^{n} \mathbb{P}(X_r \neq Y_r) \\ &\leq \sum_{r=1}^{n} \left(e^{-p_r} - 1 + p_r + \mathbb{P}(Y_r \geq 2)\right) \\ &\leq \sum_{r=1}^{n} p_r(1 - e^{-p_r}) \leq \sum_{r=1}^{n} p_r^2. \end{aligned}$$

□

Definition 5 (Poisson Process)**.** A Poisson process with intensity λ is a process $N = \{N(t) : t \geq 0\}$ taking values in $S = \{0,1,2,\ldots\}$ such that:

1. $N(0) = 0$; if $s < t$ then $N(s) \leq N(t)$.

2.
$$\mathbb{P}\left(N(t+h)=n+m|N(t)=n\right)=\begin{cases}\lambda h+o(h) & m=1\\ o(h) & m>1\\ 1-\lambda h & m=0\end{cases}.$$

3. If $s<t$ the number $N(t)-N(s)$ of emissions in the interval $(s,t]$ is independent of the times if emission in $[0,s]$ and has a distribution that only depends on $t-s$.

Proposition 20. *$N(t)$ has the Poisson distribution with parameter λt; that is to say,*
$$\mathbb{P}\left(N(t)=j\right)=\frac{(\lambda t)^j}{j!}e^{-\lambda t}$$

Proof. Let $p_j(t)=\mathbb{P}\left(N(t)=j\right)$. By conditioning on $\mathbb{P}\left(N(t)=0\right)$ to see that
$$p_0(t+h)=(1-\lambda h+o(h))p_0(t)\implies\frac{p_0(t+h)-p_0(t)}{h}=-\lambda p_0(t)+\frac{o(h)}{h}p_0(t)$$
taking the limit as $h\to 0$
$$p_0'(t)=-\lambda p_0(t)$$
we also have $p_0(0)=1$ so we can conclude
$$p_0(t)=e^{-\lambda t}$$
which is consistent with the proposition. Assuming, for induction, that the proposition holds up to $j-1$ we can see that
$$\begin{aligned}p_j(t+h)=\mathbb{P}\left(N(t+h)=j\right)&=\sum_{k=0}^{j}\mathbb{P}\left(N(t+h)=j|N(t)=k\right)\mathbb{P}\left(N(t)=k\right)\\&=p_j(t)(1-\lambda h+o(h))+p_{j-1}(t)(\lambda h+o(h))+o(h).\end{aligned}$$
Or equivalently
$$\frac{p_j(t+h)-p_j(t)}{h}=-\lambda p_j(t)+\lambda p_{j-1}(t)+\frac{o(h)}{h}$$
when $h\to 0$ we get
$$p_j'(t)=-\lambda p_j(t)+\lambda p_{j-1}(t)\qquad p_j(0)=0.$$
Hence,
$$p_j(t)=\int_0^t\lambda p_{j-1}(\tau)e^{\lambda(\tau-t)}d\tau$$
using our induction hypothesis
$$p_j(t)=\frac{\lambda^j e^{-\lambda t}}{(j-1)!}\int_0^t\tau^{j-1}d\tau=\frac{(\lambda t)^j}{j!}e^{-\lambda t}.$$
□

Remark (Alternative Formulation of a Poisson Process)**.** Let $T_0, T_1, \ldots$ be given by

$$T_0 = 0, \qquad T_n = \inf\{t : N(t) = n\}.$$

Then T_n is the time of the nth arrival. The **inter-arrival times** are random variables $X_1, X_2, \ldots$ given by

$$X_n = T_n - T_{n-1}.$$

From knowledge of N we can find the values of $X_1, X_2, \ldots$ from above. Conversely, we can reconstruct N from the knowledge of the X_i by

$$T_n = \sum_{i=1}^{n} K_i, \qquad N(t) = \max\{n : T_n \leq t\}.$$

Proposition 21. *The random variables $X_1, X_2, \ldots$ are independent, each having the exponential distribution with mean $1/\lambda$.*

Proof. First consider $X_1 = T_1 - T_0 = T_1$. This means that

$$\mathbb{P}(X_1 > t) = \mathbb{P}(T_1 > t) = \mathbb{P}(N(t) = 0) = e^{-\lambda t}.$$

so

$$F_{X_1}(t) = \mathbb{P}(X_1 \leq t) = 1 - \mathbb{P}(X_1 > t) = 1 - e^{-\lambda t};$$

that is, X_1 is distributed as claimed. Now, suppose we know $X_i = t_i$ for $1 \leq i < n$. We have

$$\mathbb{P}(X_n > t | X_1 = t_1, \ldots, X_{n-1} = t_{n-1}) = \mathbb{P}\left(N(\sum_{i=1}^{n-1} t_i + t) - N(\sum_{i=1}^{n-1} t_i) = 0 | X_1 = t_1, \ldots, X_{n-1} = t_{n-1}\right);$$

that is, the conditional probability of the n-th inter-arrival time being greater than t is the same as the conditional probability of no arrival in in the interval

$$\left(\sum_{i=1}^{n-1} t_i, \sum_{i=1}^{n-1} t_i + t\right],$$

given the first $n-1$ inter-arrival times. By the third defining property of the Poisson process, this event is independent of all previous arrivals and has a distribution that only depends on the length of the interval - t. Hence, we have

$$\mathbb{P}(X_n > t | X_1 = t_1, \ldots, X_{n-1} = t_{n-1}) = e^{-\lambda t}.$$

and similar to the $n = 1$ case we can conclude that X_n is exponentially distribution with mean $1/\lambda$. □

Based on the properties of the Poisson process, if we take $A = \bigcup_{i=1}^{n}(a_i, b_i]$ where $(a_i, b_i]$ are disjoint we can see that the number of arrivals in A has a Poisson distribution with parameter

$$\sum_{i=1}^{n} \lambda(b_i - a_i) = \lambda \sum_{i=1}^{n}(b_i - a_i) = \int_A \lambda d\mu$$

where μ is the Lebesgue measure on $\mathbb{R}$. This result can be extended to any Borel set A. Moreover, the number of arrivals for disjoint sets are independent. This motivates the following definition.

Definition 6. Let $d \geq 1$ and let $\lambda : \mathbb{R}^d \to \mathbb{R}$ be a non-negative measurable function such that

$$\Lambda(A) = \int_A \lambda(x)dx < \infty$$

for all bounded sets $A \in \mathscr{B}(\mathbb{R}^d)$. The random countable subset Π of $\mathbb{R}^d$ is called a **non-homogeneous Poisson process with intensity function** λ if for all $A \in \mathscr{B}(\mathbb{R}^d)$, the random variables $N(A) = |\Pi \cap A|$ satisfy

1. $N(A)$ has the Poisson distribution with parameter $\Lambda(A)$, and
2. If $A_1, \ldots, A_n$ are disjoint sets in $\mathscr{B}(\mathbb{R}^d)$, then $N(A_1), \ldots, N(A_n)$ are independent random variables. We call the function $\Lambda(A)$ for $A \in \mathscr{B}(\mathbb{R}^d)$, the **mean measure** of the process Π.

Proposition 22 (Superposition Theorem)**.** *Let Π' and Π'' be independent Poisson processes on $\mathbb{R}^d$ with respective intensities λ' and λ''. The set $\Pi = \Pi' \cup \Pi''$ is a Poisson process with intensity function $\lambda = \lambda' + \lambda''$.*

Proposition 23 (Mapping Theorem)**.** *Let Π be a non-homogeneous Poisson process on $\mathbb{R}^d$ with intensity function λ, and let $f : \mathbb{R}^d \to \mathbb{R}^s$ be such that*

$$\Lambda\left(f^{-1}(y)\right) = 0, \qquad \text{for all } y \in \mathbb{R}^s.$$

Assume further that

$$\mu(B) = \Lambda\left(f^{-1}(y)\right) = \int_{f^{-1}B} \lambda(x)dx < \infty, \qquad \text{for all bounded } B \in \mathscr{B}(\mathbb{R}^d).$$

Then $f(\Pi)$ is a non-homogeneous Poisson process on $\mathbb{R}^s$ with mean measure μ.

Proposition 24 (Conditional Property)**.** *Let Π be a non-homogeneous Poisson process on $\mathbb{R}^d$ with intensity function λ and let A be a subset of $\mathbb{R}^d$ such that $0 < \Lambda(A) < \infty$. Conditional on the event that $|\Pi \cap A| = n$, then n points of the process lying in A have the same distribution as n points chosen independently at random in A according to the common probability measure*

$$\mathbb{Q}(B) = \frac{\Lambda(B)}{\Lambda(A)}, \qquad B \subseteq A.$$

The corresponding density function is $\lambda(x)/\Lambda(A)$ for $x \in A$, since

$$\mathbb{Q}(B) = \int_B \frac{\lambda(x)}{\Lambda(A)} dx.$$

When Π has constant intensity λ the theorem implies that, given $|\Pi \cap A| = n$, the n points in question are distributed uniformly and independently at random in A.

Proof. Let $A_1, \ldots, A_k$ be a partition of A. It is easy to see that for $n_1 + \ldots + n_k = n$, we have

$$\mathbb{P}\left(N(A_1) = n_1, \ldots, N(A_k) = n_k | N(A) = n\right) = \frac{\prod_{i=1}^k \mathbb{P}\left(N(A_i) = n_i\right)}{\mathbb{P}\left(N(A) = n\right)}$$

but we have

$$\prod_{i=1}^k \mathbb{P}\left(N(A_i) = n_i\right) = \prod_{i=1}^k \frac{\Lambda(A_i)^{n_i}}{n_i!} e^{-\Lambda(A_i)} = \frac{e^{-\Lambda(A_i) - \ldots - \Lambda(A_k)}}{n_1! \ldots n_k!} \prod_{i=1}^k \Lambda(A_i)^{n_i}$$
$$= \frac{e^{-\Lambda(A)}}{n_1! \ldots n_k!} \prod_{i=1}^k \Lambda(A_i)^{n_i}.$$

On the other hand

$$\mathbb{P}\left(N(A) = n\right) = \frac{e^{-\Lambda(A)}}{n!} \Lambda(A)^n.$$

Combinig these results and noting that $n_1 + \ldots + n_k = n$ we get

$$\therefore \mathbb{P}\left(N(A_1) = n_1, \ldots, N(A_k) = n_k | N(A) = n\right) = \frac{n!}{n_1! \ldots n_k!} \prod_{i=1}^k \left(\frac{\Lambda(A_i)}{\Lambda(A)}\right)^{n_i}$$

which matches the claimed distribution. □

Proposition 25 (Coloring Theorem)**.** *Let Π be a non-homogeneous Poisson process on $\mathbb{R}^d$ with intensity function λ. We color the points of Π in the following way. A point of Π at position x is colored green with probability $\gamma(x)$; otherwise it is colored scarlet (with probability $\sigma(x) = 1 - \gamma(x)$). Points are colored independently of one another. Let Γ and Σ be the sets of points colored green and scarlet respectively. Then Γ and Σ are independent Poisson processes with respective intensity functions $\gamma(x)\lambda(x)$ and $\sigma(x)\lambda(x)$.*

Proof. Add proof. □

Proposition 26 (Rényi's Theorem)**.** *Let Π be a random countable subset of $\mathbb{R}^d$, and let $\lambda : \mathbb{R}^d \to \mathbb{R}$ be a non-negative integrable function satisfying*

$$\Lambda(A) = \int_A \lambda(x) dx < \infty$$

for all bounded Borel sets A. If

$$\mathbb{P}(\Pi \cap A = \varnothing) = e^{-\Lambda(A)}$$

for any infinite union A of boxes, then Π is Poisson process with intensity function λ.

7 Generating Functions

Definition 7. Suppose X is a random variable that takes values in $\{0, 1, \ldots\}$ with probability mass function f. The **probability generating function** of the random variable X is given by

$$G(s) = \mathbb{E}\left[s^X\right] = \sum_{i=0}^{\infty} s^i \mathbb{P}(X = i) = \sum_{i=0}^{\infty} s^i f(i).$$

Remark. By comparing against a geometric series we can see that G always converges on $[0, 1]$. The region of convergence might be bigger depending on f.

Remark. We write G_X instead of G if we want to emphasize the role of X.

Example 5. The probability generating functions of some elementary discrete distributions are as follows

Constant If $\mathbb{P}(X = c) = 1$ then $G_X(s) = s^c$.

Bernoulli If $\mathbb{P}(X = 1) = p$ and $\mathbb{P}(X = 0) = 1 - p$ then $G_X(s) = 1 - p + ps$.

Geometric If $\mathbb{P}(X = k) = (1-p)^{k-1}p$ for some $0 < p < 1$ and any $k \geq 1$ then

$$G_X(s) = \sum_{k=1}^{\infty} (1-p)^{k-1} p s^k = ps \sum_{k=1}^{\infty} (s(1-p))^{k-1} = \frac{ps}{1 - s(1-p)}.$$

Poisson If X has the Poisson distribution with parameter λ then

$$G_X(s) = \sum_{k=1}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} s^k e^{-\lambda} \sum_{k=1}^{\infty} \frac{(\lambda s)^k}{k!} = e^{\lambda(s-1)}.$$

Proposition 27. *If X has a generating function $G(s)$ then*

1. $\mathbb{E}[X] = G'(1)$, *and more generally*
2. $\mathbb{E}[X(X-1)\ldots(X-k+1)] = G^{(k)}(1)$.

Proof. Suppose X has probability mass function f then

$$G^{(k)}(s) = \sum_{i=0}^{\infty} i(i-1)\ldots(i-k+1)s^{i-k} f(i)$$

and $s = 1$ we get

$$G^{(k)}(1) = \sum_{i=0}^{\infty} i(i-1)\dots(i-k+1)f(i) = \mathbb{E}\left[X(X-1)\dots(X-k+1)\right].$$

□

Definition 8 (Moment Generating Function)**.** If we are more interested in the moments of X then its moment generating function $M_X(t) := G_X(e^t)$ can be more convenient

$$\begin{aligned} M_X(t) &= \sum_{k=0}^{\infty} e^{tk}\mathbb{P}\left(X = k\right) \\ &= \sum_{k=0}^{\infty}\sum_{n=0}^{\infty} \frac{(tk)^n}{n!}\mathbb{P}\left(X = k\right) \\ &= \sum_{n=0}^{\infty} \frac{t^n}{n!}\left(\sum_{k=0}^{\infty} k^n \mathbb{P}\left(X = k\right)\right) \\ &= \sum_{n=0}^{\infty} \frac{t^n}{n!}\mathbb{E}\left[X^n\right]. \end{aligned}$$

Remark. Note that if X has an infinite moment, M_X will be infinite.

Proposition 28. *If X and Y are independent then $G_{X+Y}(s) = G_X(s)G_Y(s)$.*

Proof.

$$G_{X+Y}(s) = \sum_{i=0}^{\infty} \mathbb{P}\left(X + Y = i\right) s^i = \sum_{i=0}^{\infty}\left(\sum_{j=0}^{\infty} \mathbb{P}\left(X = i - j\right)\mathbb{P}\left(Y = j\right)\right) s^i$$

we recognize that the coefficient of the i-th term in the above series is identical to the coefficient of the i-th term in the Cauchy product $G_X(s)G_Y(s)$. Hence the two series are term-by-term equal. □

Example 6 (Binomial Distribution)**.** Let $X_1, \dots, X_n$ be independent Bernoulli variables with parameter p, and let $S = X_1 + \dots + X_n$. Each X_i has generating function $G_X(s) = q + ps$, where $q = 1 - p$. Applying the above theorem n times, we can find the generating function of the Binomial(n, p) random variable S:

$$G_S(s) = G_X(s)^n = (q + ps)^n.$$

Proposition 29. *Let $\{X_i; i \in \mathbb{Z}^{\geq 0}\}$ be a sequence of i.i.d. random variables with a common generating function $G_X(s)$, and let $N \geq 0$ be random variable which is independent of theX_i with generating function G_N. The random variable $S = X_1 + \dots + X_N$ has a generating function given by*

$$G_S(s) = G_N(G_X(s)).$$

Proof.

$$G_S(s) = \mathbb{E}\left[s^S\right] = \mathbb{E}\left[\mathbb{E}\left[s^S|N\right]\right] = \mathbb{E}\left[G_X(s)^N\right] = G_N(G_X(s)).$$

□

Example 7. A hen lays N eggs, where N is Poisson distributed with parameter λ. Each egg hatches with probability p, independently of all other eggs. Let K be the number of chicks. Then $K = X_1 + \ldots + X_N$ where X_is are independent Bernoulli variables with parameter p. How is K distributed?

Clearly

$$G_N(s) = e^{\lambda(s-1)} \qquad G_X(s) = q + ps$$

and so

$$G_K(s) = G_N(G_X(s)) = e^{\lambda(q+ps-1)} = e^{\lambda(1-p+ps-1)} = e^{p\lambda(s-1)}.$$

We conclude that $K \sim \text{Poisson}(p\lambda)$.

Definition 9. The **joint probability generating function** of variables X_1 and X_2 taking non-negative integral values is defined by

$$G_{X_1,X_2}(s_1, s_2) = \mathbb{E}\left[s_1^{X_1} s_2^{X_2}\right] = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \mathbb{P}\left(X_1 = i, X_2 = j\right) s_1^i s_2^j.$$

Proposition 30. *Random variables X_1 and X_2 are independent if and only if*

$$G_{X_1,X_2}(s_1, s_2) = G_{X_1}(s_1)G_{X_2}(s_2)$$

for all s_1, s_2.

Proof. Clearly, if X_1 and X_2 are independent $s_1^{X_1}$ and $s_2^{X_2}$ are independent for all s_1, s_2 so

$$G_{X_1,X_2}(s_1, s_2) = \mathbb{E}\left[s_1^{X_1} s_2^{X_2}\right] = \mathbb{E}\left[s_1^{X_1}\right]\mathbb{E}\left[s_2^{X_2}\right] = G_{X_1}(s_1)G_{X_2}(s_2).$$

Now assume that $G_{X_1,X_2}(s_1, s_2) = G_{X_1}(s_1)G_{X_2}(s_2)$. We have

$$G_{X_1,X_2}(s_1, s_2) = \mathbb{E}\left[s_1^{X_1} s_2^{X_2}\right] = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \mathbb{P}\left(X_1 = i, X_2 = j\right) s_1^i s_2^j$$

and the term $s_1^i s_2^j$ appears on the RHS only once with the coefficient $\mathbb{P}\left(X_1 = i\right)\mathbb{P}\left(X_2 = j\right)$. Equating the coefficients implies that the joint probability factorizes to a product of marginal probabilities, i.e. that X_1 and X_2 are independent. □

We now consider two important applications of generating functions.

7.1 Random Walk

Suppose $X_1, X_2, \ldots$ are independent random variables each taking value 1 with probability p and -1 with probability $1-p$. We write $S_n = \sum_{i=1}^{n} X_i$; the sequence $S = \{S_i; i \geq 0\}$ is a **simple random walk** starting at the origin. Imagine a particle at the origin at time $t = 0$. At each time step, the particle moves one unit to the right with probability p or moves one unit to the left with probability $1-p$. The random variable X_t can be thought of as indicating the position of the particle at time t.

Let $p_0(n) = \mathbb{P}(S_n = 0)$ be the probability of being at the origin after n steps, and let $f_0(n) = \mathbb{P}(S_1 \neq 0, \ldots, S_{n-1} \neq 0, S_n = 0)$ be the probability that the first return occurs after n steps. Then consider

$$P_0(s) = \sum_{n=0}^{\infty} p_0(n) s^n, \qquad F_0(s) = \sum_{n=0}^{\infty} f_0(n) s^n.$$

F_0 is the probability generating function of the random time T_0 until the particle makes its first return to the origin. That is, $F_0(s) = \mathbb{E}\left[s^{T_0}\right]$. Note that T_0 may be defective, and so it may be the case that

$$F_0(1) = \mathbb{P}(T_0 < \infty) < 1.$$

Proposition 31.

1. $P_0(s) = 1 + P_0(s)F_0(s)$.
2. $P_0(s) = (1 - 4pqs^2)^{-1/2}$.
3. $F_0(s) = 1 - (1 - 4pqs^2)^{1/2}$.

Proof.

1. Clearly $p_0(0) = 1$. For $n > 1$ we can express $p_0(n)$ in terms of the p_0 of smaller numbers by conditioning on f_0,

$$p_0(n) = \sum_{k=0}^{n} f_0(k) p_0(n-k) \qquad (f_0(0) := 0).$$

This allows us to write

$$P_0(s) = \sum_{n=0}^{\infty} p_0(n) s^n = 1 + \sum_{n=1}^{\infty} \sum_{k=0}^{n} f_0(k) p_0(n-k) s^n.$$

Recognizing the convolution, we can conclude that

$$P_0(s) = 1 + P_0(s) F_0(s).$$

2. To be at the origin after n steps we must take an equal number of steps to the left and to the right. Hence,

$$p_0(n) = \begin{cases} 0 & n \equiv 1 \pmod 2 \\ \binom{n}{\frac{n}{2}}(pq)^{\frac{n}{2}} & n \equiv 0 \pmod 2 \end{cases}.$$

So we have

$$P_0(s) = \sum_{n=0}^{\infty} \binom{2n}{n}(pq)^n s^{2n} = \sum_{n=0}^{\infty} \binom{-\frac{1}{2}}{n}(-4pqs^2)^n = (1-4pqs^2)^n.$$

Above we have used the identity

$$\binom{-\frac{1}{2}}{n} = \frac{(-1)^n}{2^n} \times \frac{1 \times 3 \times \ldots \times (2n-1)}{n!} = \frac{(-1)^n}{2^{2n}} \times \frac{(2n)!}{n!n!} = \left(-\frac{1}{4}\right)^n \binom{2n}{n}.$$

3. This follows immediately from the previous parts.

□

Corollary 1.

1. The probability that the particle ever returns to the origin is

$$\sum_{n=1}^{\infty} f_0(n) = F_0(1) = 1 - (1-4pq)^{1/2} = 1 - |p-q|.$$

2. If eventual return is certain, that is $F_0(1) = 1$ and $p = 1/2$, then the expected time to the first return is

$$\sum_{n=1}^{\infty} n f_0(n) = F_0'(1) = \infty.$$

We call the process **persistent** (or **recurrent**) if eventual return to the origin is (almost) certain; otherwise it is called **transient**. It is immediately obvious from (1) that the process is persistent if and only if $p = 1/2$.

Define $f_r(n) = \mathbb{P}(S_1 \neq r, \ldots, S_{n-1} \neq r, S_n = r)$ to be the probability that the first visit to point r occurs at the n-th step, with generating function

$$F_r(s) = \sum_{n=1}^{\infty} f_r(n) s^n.$$

Proposition 32.

1. $F_r(s) = [F_1(s)]^r$ *for* $r \geq 1$.

2. $F_1(s) = \left[1 - (1 - 4pqs^2)^{1/2}\right] / (2qs)$.

Proof.

1. Note that for $r \geq 1$ the only way we can get to r is to first get to $r-1$, so noting temporal and spatial homogeneity we can condition on the first time we get to $r-1$; i.e. f_{r-1}:

$$f_r(n) = \sum_{k=0}^{n} f_{r-1}(k) f_1(n-k).$$

So f_r is the convolution of f_{r-1} and f_1, and therefore $F_r(s) = F_{r-1}(s)F_1(s)$. Hence, $F_r(s) = F_1(s)^r$.

2. We employ first-step analysis:

$$f_1(n) = p \times 0 + q f_2(n-1) \qquad n \geq 2$$

Also, $f_1(1) = 1$ so we have

$$F_1(s) = \sum_{n=1}^{\infty} f_1(n) s^n = ps + qs \sum_{n=2}^{\infty} f_2(n-1) s^{n-1} = ps + qsF_2(s).$$

Using part 1 we get

$$F_1(s) = ps + qsF_1(s)^2 \implies qsF_1(s)^2 - F_1(s) + ps = 0$$

hence

$$F_1(s) = \frac{1 \pm \left(1 - 4pqs^2\right)^{1/2}}{2qs} = \frac{1 \pm \left(1 - 2pqs^2 + o(s^2)\right)^{1/2}}{2qs}.$$

Now as $s \to 0$ the limit of $F_1(s)$ must exist, and this is only the case if we have

$$F_1(s) = \frac{1 - \left(1 - 4pqs^2\right)^{1/2}}{2qs}.$$

□

Corollary 2. The probability that the walk ever visits the positive part of the real axis,

$$F_1(1) = \frac{1 - (1 - 4pq)^{1/2}}{2q} = \frac{1 - |p - q|}{2q} = \min\{1, \frac{p}{q}\}.$$

7.2 Branching Processes

Suppose that a population evolves in generations, and let Z_n be the number of members of the n-th generation. Each member of the n-th generation gives birth to a family, possibly empty, of members of the $(n+1)$-th generation; the size of the family is a random variable. We shall make the following assumptions about the family sizes:

1. the family sizes of the individuals of the branching process form a collection of independent random variables;
2. all family sizes have the same probability mass function f and a generating function G.

We are interested in the random sequence $Z_0, Z_1, \ldots$ of generation sizes. Let $G_n(s) = \mathbb{E}\left[X^{Z_n}\right]$.

Proposition 33. *It is the case that $G_{m+n}(s) = G_m(G_n(s)) = G_n(G_m(s))$, and thus $G_n(s) = G(G(G(\ldots(G(s))\ldots)))$ is the n-fold iterate of G.*

Proof. The population of generation $n+1$ can be obtained by summing N_n random variables each of which having generating function G. On the other hand, N_n is the population of the n-th generation that has generating function G_n. Hence, by proposition 29 we have $G_{n+1} = G_n \circ G$. Assuming that the first generation had population one; i.e. $G_1 = G$ we can easily see

$$G_n = \underbrace{G \circ G \circ \ldots \circ G}_{n \text{ times}}.$$

□

Proposition 34. *Let $\mu = \mathbb{E}[Z_1]$ and $\sigma^2 = Var(Z_1)$. Then*

$$\mathbb{E}[Z_n] = \mu^n \qquad Var(Z_n) = \begin{cases} n\sigma^2 & \mu = 1, \\ \frac{\sigma^2(\mu^n - 1)\mu^{n-1}}{\mu - 1} & \mu \neq 1. \end{cases}$$

Proof. When $n = 1$ the proposition holds by definition of μ and σ^2. Assume, for induction, that the proposition holds up to n. By proposition 8 we have

$$G_{n+1}(s) = G(G_n(s))$$

now

$$\mathbb{E}[Z_{n+1}] = G'_{n+1}(1) = G'_n(1)G'(G_n(1)) = \mu^n G'(1) = \mu^{n+1}.$$

Also,

$$\mathrm{Var}(Z_{n+1}) = \mathbb{E}\left[Z_{n+1}^2\right] - \mathbb{E}[Z_{n+1}]^2 = G''_{n+1}(1) + G'_{n+1}(1) - G'_{n+1}(1)^2, \quad (17)$$

and

$$G''_{n+1}(s) = (G'_n(s)G'(G_n(s)))' = G''_n(s)G'(G_n(s)) + G'_n(s)^2 G''(G_n(s))$$

so

$$G''_{n+1}(1) = G''_n(1)G'(G_n(1)) + G'_n(1)^2 G''(G_n(1)) = G''_n(1)\mu^n + \mu^{2n}(\sigma^2 - \mu + \mu^2)$$

substituting in (17) and using the induction hypothesis we conclude the inductive step. □

Example 8 (Geometric Branching)**.** Suppose that each family size has the mass function $f(k) = qp^k$, for $k \geq 0$, where $q = 1 - p$. Then

$$G(s) = \sum_{k=0}^{\infty} qp^k s^k = q(1 - ps)^{-1},$$

and each family size is one member less than a geometric variable. Using proposition and induction we can see that $G_n(s)$ has the form

$$G_n(s) = \frac{a_0(n) + a_1(n)s}{b_0(n) + b_1(n)s}$$

where

$$\underbrace{\begin{bmatrix} a_0(n) \\ a_1(n) \\ b_0(n) \\ b_1(n) \end{bmatrix}}_{x_n} = \underbrace{\begin{bmatrix} 0 & 0 & q & 0 \\ 0 & 0 & 0 & q \\ -p & 0 & 1 & 0 \\ 0 & -p & 0 & 1 \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} a_0(n) \\ a_1(n) \\ b_0(n) \\ b_1(n) \end{bmatrix}}_{x_{n-1}}.$$

Solving this linear recurrence we get

$$G_n(s) = \begin{cases} \frac{n-(n-1)s}{n+1-ns} & p = q = \frac{1}{2} \\ \frac{q[p^n - q^n - ps(p^{n-1} - q^{n-1})]}{p^{n+1} - q^{n+1} - ps(p^n - q^n)} & p \neq q. \end{cases}$$

This allows us to compute

$$\mathbb{P}(Z_n = 0) = G_n(0) = \begin{cases} \frac{n}{n+1} & p = q \\ \frac{q(p^n - q^n)}{p^{n+1} - q^{n+1}} & p \neq q. \end{cases}$$

so as $n \to \infty$

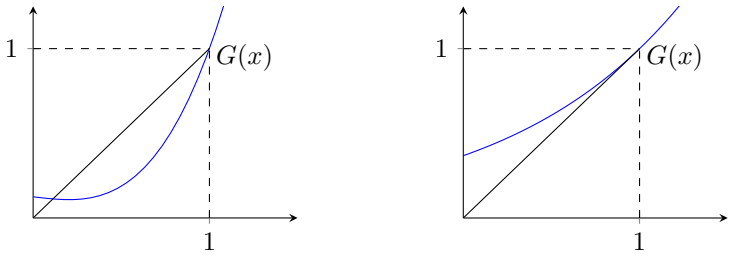
$$\mathbb{P}(Z_n = 0) \to \mathbb{P}(\text{ultimate extinction}) = \begin{cases} 1 & p \leq q \\ q/p & p > q. \end{cases}$$

Note that the expected family size is p/q.

Proposition 35. *As $n \to \infty$,*

$$\mathbb{P}(Z_n = 0) \to \mathbb{P}(\textit{ultimate extinction}) := \eta,$$

where η is the smallest non-negative root of the equation $G(s) = s$. Also, $\eta = 1$ if $\mu < 1$, and $\eta < 1$ if $\mu > 1$. If $\mu = 1$ then $\eta = 1$ so long as the family-size distribution has strictly positive variance.



(a) $\mu > 1$ (b) $\mu \leq 1$

Figure 1: Extinction Probabilities of a Branching Process

Proof. Note that

$$\mathbb{P}(Z_n = 0) = G_n(0) = G(G_{n-1}(0)).$$

also the sequence $\{G_n(0)\}$ is bounded above by 1 and is monotonic since G is increasing. Hence, since G is continuous over its region of convergence, the limit satisfies

$$\eta = G(\eta).$$

Now assume for some $\varphi \geq 0$ we have $\varphi = G(\varphi)$. Note that since G is increasing we must have

$$G_1(0) = G(0) \leq \varphi$$

also,

$$G_k(0) \leq \varphi \implies G(G_k(0)) \leq G(\varphi) \implies G_{k+1}(0) \leq \varphi.$$

So the limit of the sequence $\{G_n(0)\}$, the probability of extinction, has to be the smallest non-negative root of $G(s) = s$.

Now note that

$$G''(s) = \mathbb{E}\left[Z_1(Z_1 - 1)s^{Z_1 - 2}\right] \geq 0 \qquad \text{if } s \geq 0.$$

So G is convex on $[0, 1]$ with $G(1) = 1$. Using mean value theorem we can conclude that $G(s) = s$ has at most two roots if $G''(s) \neq 0$. These intersections coincide if $\mu = G'(1) \leq 1$. If $\mu > 1$ the intersections are distinct. Figure 1 illustrates the situation. □

7.3 Characteristic Functions

Recall

Definition 10. The **moment generating function** of the random variable X is defined to be the function $M : \mathbb{R} \to [0, \infty)$ given by $M(t) = \mathbb{E}\left[e^{tX}\right]$.

If $M(t) < \infty$ on some open interval containing the origin then:

1. $\mathbb{E}\left[X^k\right] = M^{(k)}(0)$;

2. $M(t) = \sum_{k=0}^{\infty} \frac{\mathbb{E}\left[X^k\right]}{k!} t^k$;

3. if X and Y are independent then $M_{X+Y}(t) = M_X(t) M_Y(t)$.

Notice that the moment generating function of X might be infinite for certain values of t. To remedy this potential flaw we define the characteristic function of X:

Definition 11. The **characteristic function** of a random variable X is the function $\phi : \mathbb{R} \to \mathbb{C}$ given by $\phi(t) = \mathbb{E}\left[e^{itX}\right]$ where $i = \sqrt{-1}$.

Proposition 36. *The characteristic function ϕ satisfies:*

1. *$\phi(0) = 1$, $|\phi(t)| \leq 1$ for all $t \in \mathbb{R}$.*

2. *ϕ is uniformly continuous on $\mathbb{R}$.*

3. *ϕ is non-negative definite, which is to say that*

$$\sum_{j,k} \phi(t_j - t_k) z_j \bar{z}_k \geq 0$$

for all real $t_1, \ldots, t_n$ and complex $z_1, \ldots, z_n$.

Proof. 1. By definition

$$\phi(0) = \mathbb{E}\left[e^{i0X}\right] = \mathbb{E}\left[1\right] = 1.$$

Also,

$$\begin{aligned} |\phi(t)| = \left|\mathbb{E}\left[e^{itX}\right]\right| &= \left| \int_{\mathbb{R}} i\sin(tx) + \cos(tx) dF_X(x) \right| \\ &\leq \int_{\mathbb{R}} \left| i\sin(tx) + \cos(tx) \right| dF_X(x) \\ &\leq \int_{\mathbb{R}} 1 dF_X(x) = 1. \end{aligned}$$

2.
$$|\phi(t_1) - \phi(t_2)| = \left| \mathbb{E}\left[e^{it_1X}(1 - e^{i(t_2 - t_1)X}) \right] \right| \leq \mathbb{E}\left[\left| 1 - e^{i(t_2 - t_1)X} \right| \right]$$

and as $t_2 \to t_1$ we can see that the right-hand side approaches zero. Hence, given any $x, y \in \mathbb{R}$ and $\varepsilon > 0$ we can pick $\delta > 0$ so that

$$\mathbb{E}\left[\left|1 - e^{i\delta X}\right|\right] < \varepsilon$$

which establishes

$$|\phi(x) - \phi(y)| < \varepsilon.$$

3. We have that

$$\sum_{j,k} \phi(t_j - t_k) z_j \bar{z}_k = \mathbb{E}\left[\left|\sum_j z_j \exp(it_j X)\right|^2\right] \geq 0$$

□

Proposition 36 characterizes the characteristic functions in the sense that ϕ is a characteristic function if and only if it satisfies all three conditions above. This result is called Bochner's theorem.

We also have the following

1. If $\phi^{(k)}(0)$ exists then $\begin{cases} \mathbb{E}\left[|X^k|\right] < \infty & k \equiv 0 \pmod 2 \\ \mathbb{E}\left[|X^{k-1}|\right] < \infty & k \equiv 1 \pmod 2 \end{cases}$

2. If $\mathbb{E}\left[|X^k|\right] < \infty$ then

$$\phi(t) = \sum_{j=0}^{k} \frac{\mathbb{E}\left[X^j\right]}{j!}(it)^j + o(t^k),$$

and so $\phi^{(k)}(0) = i^k \mathbb{E}\left[X^k\right]$.

Proposition 37. *If X and Y are independent then $\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t)$.*

Proof.

$$\phi_{X+Y}(t) = \mathbb{E}\left[e^{i(X+Y)t}\right] = \mathbb{E}\left[e^{iXt}e^{iYt}\right] = \mathbb{E}\left[e^{iXt}\right]\mathbb{E}\left[e^{iYt}\right] = \phi_X(t)\phi_Y(t).$$

□

Proposition 38. *If $a, b \in \mathbb{R}$ and $Y = aX + b$ then $\phi_Y(t) = e^{itb}\phi_X(at)$.*

Proof.

$$\phi_Y(t) = \mathbb{E}\left[e^{i(aX+b)t}\right] = \mathbb{E}\left[e^{ibt}e^{iaXt}\right] = e^{ibt}\mathbb{E}\left[e^{iaXt}\right] = e^{ibt}\phi_X(at).$$

□

Definition 12. The **joint characteristic function** of X and Y is the function $\phi_{X,Y} : \mathbb{R}^2 \to \mathbb{R}$ given by $\phi_{X,Y}(s,t) = \mathbb{E}\left[e^{isX}e^{itY}\right]$.

Proposition 39. *Random variables X and Y are independent if and only if $\phi_{X,Y}(s,t) = \phi_X(s)\phi_Y(t)$ for all s and t.*

Proposition 40. *Random variable X and Y have the same characteristic function if and only if they have the same distribution function.*

Remark. The last two propositions imply the Cramer-Wold device. Let

$$\mathbf{t} = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}$$

now

$$\mathbb{E}\left[\mathbf{t}'\mathbf{X}\right] = \phi_{t_1X_1+\ldots+t_nX_n}(1) = \mathbb{E}\left[e^{t_1X_1+\ldots+t_nX_n}\right] = \mathbb{E}\left[e^{t_1X_1}\ldots e^{t_nX_n}\right] = \phi_{\mathbf{X}}(\mathbf{t}).$$

Hence, knowing the mean of $\mathbf{t}'\mathbf{X}$ for all $\mathbf{t}$ is equivalent to knowing the characteristic function of $\mathbf{X}$ and, therefore, the distribution of $\mathbf{X}$.

Example 9.

1. **Exponential Distribution.** If $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$ then

$$\phi(t) = \mathbb{E}\left[e^{itX}\right] = \int_0^\infty e^{itx}\lambda e^{-\lambda x}dx = \frac{\lambda}{\lambda - it}.$$

2. **Normal Distribution.** If X is $N(0,1)$ then

$$\begin{aligned}
\phi(t) = \mathbb{E}\left[e^{itX}\right] &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp(itx - \frac{1}{2}x^2)dx \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}(x^2 - 2itx - t^2) - \frac{1}{2}t^2)dx \\
&= \exp(-\frac{1}{2}t^2)\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}(x-it)^2)dx \\
&= \exp(-\frac{1}{2}t^2).
\end{aligned}$$

   So $\phi(t) = e^{t^2/2}$. Then for $Y = \sigma X + \mu$, the characteristic function is:

$$\phi_Y(t) = e^{it\mu}\phi_X(\sigma t) = \exp(i\mu t - \frac{1}{2}\sigma^2 t^2).$$

3. **Multivariate Normal Distribution.** Let $\mathbf{X} \sim \mathcal{N}(\mu, \mathbf{\Sigma})$ and $\mathbf{x}' = [x_1 \ \ldots \ x_n]$. We have

   $$\phi_{\mathbf{X}}(\mathbf{x}) = \mathbb{E}\left[e^{iX_1x_1}\ldots e^{iX_nx_n}\right] = \mathbb{E}\left[e^{i(x_1X_1+\ldots+x_nX_n)}\right] = \mathbb{E}\left[e^{i(\mathbf{x}'\mathbf{X})}\right] = \phi_{\mathbf{x}'\mathbf{X}}(1)$$

   but $\mathbf{x}'\mathbf{X} \sim \mathcal{N}(\mathbf{x}'\mu, \mathbf{x}'\mathbf{\Sigma}\mathbf{x})$ so

$$\phi_{\mathbf{X}}(\mathbf{x}) = \exp(i\mathbf{x}'\mu - \frac{1}{2}\mathbf{x}'\mathbf{\Sigma}\mathbf{x}).$$

Definition 13. We say the sequence $F_1, F_2, \ldots$ of distribution functions **converges** to the distribution function F, written $F_n \to F$, if $F(x) = \lim_{n\to\infty} F_n(x)$ at each point x where F is continuous.

Proposition 41 (Lévy's Continuity Theorem)**.** *Suppose $\{F_n\}$ is a sequence of distribution functions with corresponding characteristic functions $\{\phi_n\}$.*

1. *If $F_n \to F$ for some distribution F with characteristic function ϕ, then $\phi_n(t) \to \phi(t)$ for all t.*
2. *Conversely, if $\phi(t) = \lim_{n\to\infty} \phi_n(t)$ exists and is continuous at $t = 0$, then phi is the characteristic function of some distribution function F, and $F_n \to F$.*

Definition 14 (Convergence in Distribution)**.** If $\{X_n\}$ is a sequence of random variables with respective distribution functions $\{F_n\}$ we say that X_n **converges in distribution** to X, written $X_n \xrightarrow{D} X$, if $F_n \to F$ as $n \to \infty$, where F is the distribution function of X.

8 Limit Theorems

8.1 A First Attempt at Law of Large Numbers

Proposition 42 (A Law of Large Numbers)**.** *If $\{X_n\}$ is a sequence of i.i.d. random variables with finite mean μ, their partial sums $S_n = X_1 + \ldots + X_n$ satisfy*

$$\frac{1}{n} S_n \xrightarrow{D} \mu.$$

Proof. Consider the characteristic function

$$\phi_{S_n/n}(t) = \phi_{S_n}(t/n) = \phi_{X_1}(t/n)^n.$$

Since $\mu = \mathbb{E}[X_1]$ is finite we can write

$$\phi_{X_1}(t/n) = \left(1 + \frac{it\mu}{n} + o(\frac{t}{n})\right)$$

so

$$\phi_{S_n/n}(t) = \left(1 + \frac{it\mu}{n} + o(\frac{t}{n})\right)^n$$

which approaches $\exp(it\mu)$, the characteristic function of the constant random variable μ, as $n \to \infty$. □

Proposition 43 (Central Limit Theorem)**.** *Let $\{X_n\}$ be a sequence of i.i.d. random variables with finite mean μ and finite non-zero variance σ^2, and let $S_n = X_1 + \ldots + X_n$. Then*

$$\frac{S_n - n\mu}{\sqrt{n\sigma^2}} \xrightarrow{D} \mathcal{N}(0,1) \qquad as \qquad n \to \infty.$$

Proof. Let $Y_i = \frac{X_i - \mu}{\sqrt{\sigma^2}}$ so that

$$\frac{S_n - n\mu}{\sqrt{n\sigma^2}} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i.$$

so the characteristic function of the distribution in question is given by

$$\phi_n(t) = \phi_{Y_1}(\frac{t}{\sqrt{n}})^n.$$

Note that $\mathbb{E}[Y_i] = 0$ and $\mathrm{Var}(Y_i) = 1$ so we can write

$$\phi_{Y_i}(t) = 1 - \frac{t^2}{2} + o(t^2)$$

hence,

$$\phi_n(t) = \left(1 - \frac{t^2}{2n} + o(\frac{t^2}{n})\right)^n$$

and we get $\phi_n(t) \to \exp(-t^2/2)$ as $n \to \infty$ which is the characteristic function of the standard normal distribution. □

9 Inequalities

Proposition 44 (Markov's Inequality)**.** *Let X be a random variable, and let g be a non-negative Borel-measurable function such that $\mathbb{E}[g(X)] < \infty$. Suppose that g is even and non-decreasing on $[0, \infty)$. Then, for every $\varepsilon > 0$,*

$$\mathbb{P}(|X| \geq \varepsilon) \leq \frac{\mathbb{E}[g(X)]}{g(\varepsilon)}.$$

Proof. Let

$$Y = g(\varepsilon)\mathbf{1}\{|X| \geq \varepsilon\}.$$

Note that our assumptions on imply that $g(X) \geq Y$, so

$$\mathbb{E}[g(X)] \geq \mathbb{E}[Y] = \mathbb{E}[g(\varepsilon)\mathbf{1}\{|X| \geq \varepsilon\}] = g(\varepsilon)\mathbb{E}[\mathbf{1}\{|X| \geq \varepsilon\}] = g(\varepsilon)\mathbb{P}(|X| \geq \varepsilon).$$

□

Proposition 45. *Let $X \geq 0$ a.s. and let F be its distribution function. Then*

$$\mathbb{E}[X] < \infty \iff \int_0^\infty [1 - F(x)]dx < \infty.$$

In this case the following relation holds:

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X > x)\, dx = \int_0^\infty [1 - F(x)]dx.$$

Proof.

$$\int_0^\infty \mathbb{P}\left(X > x\right) dx = \int_0^\infty \mathbb{E}\left[\mathbf{1}_{\{X>x\}}\right] dx = \mathbb{E}\left[\int_0^\infty \mathbf{1}_{\{X>x\}dx}\right] = \mathbb{E}\left[X\right].$$

□

Corollary 3. Let X be a random variable, and let $0 < p < \infty$, then

$$\mathbb{E}\left[|X|^p\right] = \int_0^\infty \mathbb{P}\left(|X|^p \geq x\right) dx$$

letting $x = u^p$ we get

$$\mathbb{E}\left[|X|^p\right] = \int_0^\infty \mathbb{P}\left(|X|^p \geq u^p\right) pu^{p-1} du = p\int_0^\infty x^{p-1}\mathbb{P}\left(|X| \geq x\right) dx$$

Proposition 46 (Jensen's Inequality)**.** *Suppose ϕ is convex, that is,*

$$\lambda\phi(x) + (1-\lambda)\phi(y) \geq \phi(\lambda x + (1-\lambda)y)$$

for all $\lambda \in (0,1)$ and $x, y \in \mathbb{R}$. Suppose X and $\phi(X)$ have finite expectations, then

$$\phi(\mathbb{E}\left[X\right]) \leq \mathbb{E}\left[\phi(X)\right].$$

Proof. Since ϕ is convex, it is differentiable. Consider the line touching the graph of ϕ at $\mathbb{E}\left[X\right]$,

$$l(x) = \phi(\mathbb{E}\left[X\right]) + (x - \mathbb{E}\left[X\right])m$$

Due to convexity we must have $l(x) \leq \phi(x)$ hence,

$$\mathbb{E}\left[l(x)\right] \leq \mathbb{E}\left[\phi(X)\right] \implies \mathbb{E}\left[\phi(\mathbb{E}\left[X\right])\right] + m\mathbb{E}\left[x - \mathbb{E}\left[X\right]\right] \leq \mathbb{E}\left[\phi(X)\right]$$

$$\therefore \phi(\mathbb{E}\left[X\right]) \leq \mathbb{E}\left[\phi(X)\right].$$

□

Proposition 47 (Young's Inequality)**.** *Let $g : \mathbb{R} \to \mathbb{R}$ be a differentiable, strictly increasing function such that $g(0) = 0$, then for all $a, b > 0$*

$$ab \leq \int_0^a g(x)dx + \int_0^b g^{-1}(x)dx.$$

Proof. Let

$$G(a) = ab - \int_0^a g(x)dx$$

we have

$$G'(a) = b - g(a)$$

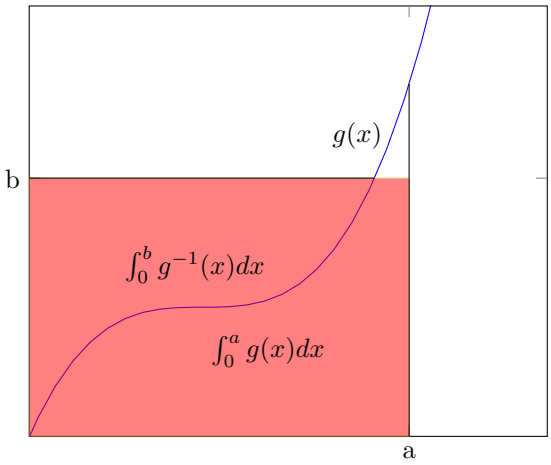


Figure 2: Graphical Illustration of Young's Inequality

and since g is strictly increasing using the first derivative test we can see that G attains its maximum at $a = g^{-1}(b)$. Now

$$G(g^{-1}(b)) = g^{-1}(b)b - \int_0^{g^{-1}(b)} g(x)dx.$$

Using the substitution $x = g^{-1}(u)$ and integration by parts we get

$$G(g^{-1}(b)) = \int_0^b g^{-1}(x)dx$$

which finishes the proof. Figure 2 is illuminating as well. □

Definition 15. Let $\|X\|_p = (\mathbb{E}[|X|^p])^{1/p}$ for $1 \leq p < \infty$, and notice $\|cX\|_p = |c|\|X\|_p$ for any real number c.

Proposition 48 (Hölder's Inequality)**.** *Let $p, q \in (1, \infty)$ so that $1/p + 1/q = 1$; then*

$$\mathbb{E}[|XY|] \leq \|X\|_q \|X\|_p$$

Proof. If either X or Y are a.s. 0, the inequality holds. Otherwise, we can divide both sides by $\|X\|_p$ and $\|Y\|_q$ to get normalized random variables X and Y. Consider the function $g(t) = t^{p-1}$. Note that

$$\frac{1}{p} + \frac{1}{q} = 1 \iff (p-1)(q-1) = 1.$$

We can apply Young's inequality to g to get

$$ab < \int_0^a t^{p-1}dt + \int_0^b t^{q-1}dt \implies ab < \frac{a^p}{p} + \frac{b^q}{q}.$$

Letting $a = |X|$, $b = |Y|$, and taking expectations we get

$$\mathbb{E}\left[|XY|\right] < \mathbb{E}\left[\frac{|X|^p}{p}\right] + \mathbb{E}\left[\frac{|Y|^q}{q}\right]$$

since X and Y are normalized on the right hand side we get

$$\mathbb{E}\left[\frac{|X|^p}{p}\right] + \mathbb{E}\left[\frac{|Y|^q}{q}\right] = \frac{1}{p} + \frac{1}{q} = 1 = \|X\|_p \|Y\|_q.$$

□

Remark. The special case where $p = q = 2$ is known as **Cauchy-Schwarz inequality**.

If $1 \leq p < q < \infty$ then if $|x| \leq 1$ we have $|x|^{-p} \leq 1$ and if $|x| \geq 1$ we have $|x|^{q-p} \geq 1$. Putting these together, we can conclude that

$$1 \leq |x|^{-p} + |x|^{q-p} \implies |x|^p \leq 1 + |x|^q.$$

Therefore, if $\|X\|_q$ is finite for some random variable X, then $\|X\|_p$ is also finite. A more refined result is the following:

Proposition 49 (Lyapunov's Inequality)**.** *If $1 \leq p < q < \infty$, then*

$$\|X\|_p \leq \|X\|_q.$$

Proof. Consider the convex function $x \mapsto x^{q/p}$ and the random variable $|X|^p$. By Jensen's inequality

$$\mathbb{E}\left[|X|^p\right]^{q/p} \leq \mathbb{E}\left[(|X|^p)^{q/p}\right] \implies \|X\|_p^q \leq \|X\|_q^q \implies \|X\|_p \leq \|X\|_q.$$

□

Proposition (Minkowski's Inequality)**.** *Suppose $1 \leq p < \infty$. If $\|X\|_p$ and $\|Y\|_p$ are finite, then $\|X + Y\|_p$ is finite, and furthermore*

$$\|X + Y\|_p \leq \|X\|_p \|Y\|_p.$$

Proof. First note that the function $f(x) = x^p$ is convex for $1 \leq < \infty$ on $\mathbb{R}^+$. So

$$\left(\frac{1}{2}x + \frac{1}{2}y\right)^p \leq \frac{1}{2}x^p + \frac{1}{2}y^p$$

for non-negative x and y. In particular

$$\left(\frac{1}{2}|X + Y|\right)^p \leq \left(\frac{1}{2}|X| + \frac{1}{2}|Y|\right)^p \leq \frac{1}{2}|X|^p + \frac{1}{2}|Y|^p$$

$$\therefore |X + Y|^p \leq 2^{p-1}(|X|^p + |Y|^p)$$

Hence, if $\|X\|_p$ and $\|Y\|_p$ are finite, $\|X+Y\|_p$ must be finite. Now

$$\|X+Y\|_p^p = \mathbb{E}\left[|X+Y|^p\right] \leq \mathbb{E}\left[|X||X+Y|^{p-1}\right] + \mathbb{E}\left[|Y||X+Y|^{p-1}\right]$$

by triangle inequality. Applying Hölders inequality we have

$$\mathbb{E}\left[|X||X+Y|^{p-1}\right] \leq \|X\|_p \|(X+Y)^{p-1}\|_{\frac{p}{p-1}} = \|X\|_p \|X+Y\|_p^{p-1}$$

and similarly

$$\mathbb{E}\left[|Y||X+Y|^{p-1}\right] \leq \|Y\|_p \|X+Y\|_p^{p-1}.$$

Hence

$$\|X+Y\|_p^p \leq \|X\|_p \|X+Y\|_p^{p-1} + \|Y\|_p \|X+Y\|_p^{p-1},$$

and therefore

$$\|X+Y\|_p \leq \|X\|_p + \|Y\|_p.$$

□

10 Convergence

So far we have seen various of instances of convergence in random sequences. As illustrated in the previous sections, there are multiple useful ways of thinking about convergence of a sequence of random variables.

Definition 16. Let $X, X_1, X_2, \ldots$ be random variables on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We say

1. $X_n \to X$ **almost surely**, written $X_n \xrightarrow{\text{a.s.}} X$, if

   $$\{\omega \in \Omega; X_n(\omega) \to X(\omega) \text{ as } n \to \infty\}$$

   is an event with probability one.

2. $X_n \to X$ **in r-th mean**, where $r \geq 1$, written $X_n \xrightarrow{r} X$, if the r-th moment of all X_i is finite and for all n

   $$\mathbb{E}\left[|X_n - X|^r\right] \to 0 \qquad \text{as} \qquad n \to \infty.$$

3. $X_n \to X$ **in probability**, written $X_n \xrightarrow{\text{P}} X$, if for any $\varepsilon > 0$ we have

   $$\mathbb{P}\left(|X_n - X| > \varepsilon\right) \to 0 \qquad \text{as} \qquad n \to \infty.$$

4. $X_n \to X$ **in distribution**, written $X_n \xrightarrow{\text{D}} X$, if

   $$\mathbb{P}\left(X_n \leq x\right) \to \mathbb{P}\left(X \leq x\right) \qquad \text{as} \qquad n \to \infty$$

   for all x at which the function $F_X(x) := \mathbb{P}\left(X \leq x\right)$ is continuous.

Proposition 50. *The following implications hold*

I. $X_n \xrightarrow{r} X$ *implies* $X_n \xrightarrow{\mathrm{P}} X$,

II. $X_n \xrightarrow{\text{a.s.}} X$ *implies* $X_n \xrightarrow{\mathrm{P}} X$,

III. $X_n \xrightarrow{\mathrm{P}} X$ *implies* $X_n \xrightarrow{\mathrm{D}} X$,

and $X_n \xrightarrow{r} X$ *implies* $X_n \xrightarrow{s} X$ *for all* $r > s \geq 1$. *No other implications hold in general.*

We give a proof for each implication separately.

Proposition 51. *Convergence in probability implies convergence in distribution. That is,* $X_n \xrightarrow{\mathrm{P}} X$ *implies* $X_n \xrightarrow{\mathrm{D}} X$.

Proof. We have

$$\mathbb{P}(X_n \leq x) = \mathbb{P}(X_n \leq x, X > x + \varepsilon) + \mathbb{P}(X_n \leq x, X \leq x + \varepsilon)$$

now

$$X_n \leq x, X > x + \varepsilon \implies X_n - X < -\varepsilon$$

and

$$\{X_n - X < -\varepsilon\} \subseteq \{|X_n - X| > \varepsilon\}$$

Also, clearly

$$\{X_n \leq x, X \leq x + \varepsilon\} \subseteq \{X \leq x + \varepsilon\}$$

so

$$\mathbb{P}(X_n \leq x) \leq \mathbb{P}(|X_n - X| > \varepsilon) + \mathbb{P}(X \leq x + \varepsilon)$$

therefore

$$\limsup_n \mathbb{P}(X_n \leq x) \leq \limsup_n (\mathbb{P}(|X_n - X| > \varepsilon) + \mathbb{P}(X \leq x + \varepsilon))$$

$$\therefore \limsup_n \mathbb{P}(X_n \leq x) \leq \mathbb{P}(X \leq x + \varepsilon) \tag{18}$$

Similarly

$$\mathbb{P}(X \leq x - \varepsilon) = \mathbb{P}(X \leq x - \varepsilon, X_n > x) + \mathbb{P}(X \leq x - \varepsilon, X_n \leq X)$$

which implies

$$\mathbb{P}(X \leq x - \varepsilon) \leq \mathbb{P}(|X - X_n| < \varepsilon) + \mathbb{P}(X_n \leq x).$$

and

$$\limsup_n (\mathbb{P}(X \leq x - \varepsilon) - \mathbb{P}(|X_n - X| > \varepsilon)) \leq \liminf_n \mathbb{P}(X_n \leq x)$$

$$\therefore \mathbb{P}(X \leq x - \varepsilon) \leq \liminf_n \mathbb{P}(X_n \leq x). \tag{19}$$

Combining (19) and (19) we get

$$\mathbb{P}(X \leq x - \varepsilon) \leq \liminf_n \mathbb{P}(X_n \leq x) \leq \limsup_n \mathbb{P}(X_n \leq x) \leq \mathbb{P}(X \leq x + \varepsilon)$$

for all $\varepsilon > 0$. Now, if $F_X(t) := \mathbb{P}(X \leq t)$ is continuous at x we can take $\varepsilon \downarrow 0$ and use the fact that F_X is non-decreasing to get

$$\mathbb{P}(X \leq x) \leq \liminf_n \mathbb{P}(X_n \leq x) \leq \limsup_n \mathbb{P}(X_n \leq x) \leq \mathbb{P}(X \leq x).$$

□

Remark. The converse assertion fails in general. For instance, let $X, X_1, X_2, \ldots$ be i.i.d. Bernoulli(p) variables. Clearly, $X_n \xrightarrow{\mathrm{D}} X$, while for any $\varepsilon > 0$

$$\begin{aligned}\mathbb{P}(|X_n - X| > \varepsilon) &= \mathbb{P}(X_n \neq X) \\ &= \mathbb{P}(X_n = 0, X = 1) + \mathbb{P}(X_n = 1, X = 0) = 2p(1-p) \not\to 0.\end{aligned}$$

Proposition 52. *For any $r > s \geq 1$ convergence in r-th mean implies convergence in s-th mean, moreover, convergence in first mean implies convergence in probability. That is,*

1. *If $r > s \geq 1$ and $X_n \xrightarrow{r} X$ then $X_n \xrightarrow{s} X$.*
2. *If $X_n \xrightarrow{1} X$ then $X_n \xrightarrow{\mathrm{P}} X$.*

Proof. 1. In the proof of Lyapunov's inequality, we showed that the existence of $\mathbb{E}[|X|^r]$ implies the existence of $\mathbb{E}[|X|^s]$. Also, by Lyapunov's inequality

$$0 \leq \|X_n - X\|_s \leq \|X_n - X\|_r = \mathbb{E}[|X_n - X|^r]^{\frac{1}{r}}.$$

since $X_n \xrightarrow{r} X$ and $g(x) = x^{1/r}$ is continuous on $\mathbb{R}^{\geq 0}$ we can conclude that $\|X_n - X\|_r \to 0$ and hence, by squeeze theorem, $\|X_n - X\|_s \to 0$ which implies $X_n \xrightarrow{s} X$ using a similar argument.

2. We can apply Markov's inequality to the random variable $X_n - X$ and the function $g(x) = |x|$:

$$0 \leq \mathbb{P}(|X_n - X| > \varepsilon) \leq \frac{\mathbb{E}[|X_n - X|]}{|\varepsilon|}$$

which implies $X_n \xrightarrow{\mathrm{P}} X$ by squeeze theorem.

□

Remark. The converse does not hold in general. Let $U \sim \text{Uniform}([0,1])$. Take any $1/r < p < 1/s$ and define

$$X_n = \begin{cases} n^p & 0 \leq u \leq 1/n \\ 0 & \text{o.w.} \end{cases} \qquad X = 0.$$

To see why the converse of 1 does not hold consider

$$\mathbb{E}\left[|X_n - X|^s\right] = \frac{1}{n} \cdot n^{sp} = n^{sp-1} \to 0 \qquad n \to \infty$$

since $sp - 1 < 0$ while

$$\mathbb{E}\left[|X_n - X|^r\right] = \frac{1}{n} \cdot n^{rp} = n^{rp-1} \to \infty \qquad n \to \infty$$

since $np - 1 > 0$. For an example where X_i converge to X in probability but not in first mean let $p = 1$ and note that $\mathbb{E}\left[|X_n - X|\right] = 1$ while

$$\mathbb{P}\left(|X_n - X| > \varepsilon\right) = \frac{1}{n} \to 0 \qquad n \to \infty.$$

Proposition 53. *Let $A_n(\varepsilon) = \{|X_n - X| > \varepsilon\}$ and $B_m(\varepsilon) = \cup_{n \geq m} A_n(\varepsilon)$. Then:*

1. *$X_n \xrightarrow{\text{a.s.}} X$ if and only if $\mathbb{P}\left(B_m(\varepsilon)\right) \to 0$ as $m \to \infty$, for all $\varepsilon > 0$,*
2. *$X_n \xrightarrow{\text{a.s.}} X$ if $\sum_n \mathbb{P}\left(A_n(\varepsilon)\right) < \infty$ for all $\varepsilon > 0$,*
3. *if $X_n \xrightarrow{\text{a.s.}} X$ then $X_n \xrightarrow{\text{P}} X$, but the converse fails in general.*

Proof. 1. Let $C = \{\omega \in \Omega; X_n(\omega) \to X(\omega)\}$ and $A(\varepsilon) = \cap_m B_m(\varepsilon)$. Suppose $\mathbb{P}(C) = 1$ and for any $\varepsilon > 0$ take $\omega_0 \in A(\varepsilon)$ we have

$$\begin{aligned}
\omega_0 \in A(\varepsilon) &\implies \omega_0 \in \cap_m B_m(\varepsilon) \\
&\implies \forall m \in \mathbb{N}; \exists n \geq m; \omega_0 \in A_n(\varepsilon) \\
&\implies \forall m \in \mathbb{N}; \exists n \geq m; |X_n(\omega_0) - X(\omega_0)| > \varepsilon \\
&\implies \omega_0 \notin C \\
&\therefore \omega_0 \in \Omega \setminus C.
\end{aligned}$$

That is, $A(\varepsilon) \subseteq \Omega \setminus C$. Hence $\mathbb{P}\left(A(\varepsilon)\right) = 0$. By definition of A and the continuity from above property of the probability measure $\mathbb{P}$ we conclude

$$B_m(\varepsilon) \downarrow A(\varepsilon), \qquad \mathbb{P}\left(A(\varepsilon)\right) = 0 \qquad \implies \qquad \mathbb{P}\left(B_m(\varepsilon)\right) \downarrow 0.$$

Now assume that $\mathbb{P}\left(B_m(\varepsilon)\right) \to 0$ for all $\varepsilon > 0$ and take $\omega_0 \in \Omega \setminus C$ this means that $X_n(\omega_0) \not\to X(\omega_0)$, that is, $\omega_0 \in A(\varepsilon)$ for some $\varepsilon > 0$. Therefore, we have

$$\mathbb{P}\left(\Omega \setminus C\right) \leq \mathbb{P}\left(\bigcup_{\varepsilon} A(\varepsilon)\right)$$

noting that $A(\varepsilon) \supseteq A(\varepsilon')$ for if $\varepsilon \geq \varepsilon'$ we can write the left-hand side as a countable union of As, for instance

$$\mathbb{P}\left(\bigcup_{\varepsilon} A(\varepsilon)\right) = \mathbb{P}\left(\bigcup_{m=1}^{\infty} A(\frac{1}{m}) \cup A(m)\right)$$

but

$$\mathbb{P}\left(\bigcup_{m=1}^{\infty} A(\frac{1}{m}) \cup A(m)\right) \leq \sum_{m=1}^{\infty}\left(\mathbb{P}\left(A(\frac{1}{m})\right) + \mathbb{P}\left(A(m)\right)\right) = 0.$$

Therefore $\mathbb{P}(\Omega \setminus C) \leq 0$ and hence $\mathbb{P}(C) = 1$.

2. Using part 1 we have

$$\mathbb{P}(B_m(\varepsilon)) \leq \sum_{n=m}^{\infty} \mathbb{P}(A_n(\varepsilon)) = \sum_{n=1}^{\infty} \mathbb{P}(A_n(\varepsilon)) - \sum_{n=m-1}^{\infty} \mathbb{P}(A_n(\varepsilon))$$

for any $\varepsilon > 0$. The last expression on the right hand side is well-defined since $\sum_{n=1}^{\infty} \mathbb{P}(A_n(\varepsilon)) < \infty$ and it converges to 0 from above. Hence, by squeeze theorem, $\mathbb{P}(B_m(\varepsilon))$ should converge to zero as $m \to \infty$.

3. Clearly for any $\varepsilon > 0$

$$\mathbb{P}(B_m(\varepsilon)) \geq \mathbb{P}(A_m(\varepsilon)) \geq 0$$

so if $X_n \xrightarrow{\text{a.s.}} X$ and, by part 1, $\mathbb{P}(B_m(\varepsilon)) \to 0$ by squeeze theorem we must have $\mathbb{P}(A_m(\varepsilon)) \to 0$ and hence $X_n \xrightarrow{\text{P}} X$.

□

Remark. A counterexample for the converse of part 3 of the above theorem is as follows. Take $(\Omega, \mathcal{F}, \mathbb{P}) = ([0,1], \mathscr{B}([0,1]), \mu)$ where μ is the Lebesgue measure on $\mathbb{R}$. For each $n \in \mathbb{N}$ let e_n be the unique non-negative integer so that

$$2^{e_n} \leq n < 2^{e_n+1}.$$

Define $X_n : \Omega \to \mathbb{R}$ as

$$X_n(\omega) = \begin{cases} 1 & \frac{n-2^{e_n}}{2^{e_n}} < \omega < \frac{n-2^{e_n}+1}{2^{e_n}} \\ 0 & \text{o.w.} \end{cases} \tag{20}$$

and let $X = 0$. The first six X_i are depicted in figure 3. Clearly,

$$\mathbb{P}(|X_n - X| > \varepsilon) = \mathbb{P}(X_n = 1) = \frac{1}{2^{e_n}}$$

and $e_n \to \infty$ as $n \to \infty$ so $X_n \xrightarrow{\text{P}} X$. On the other hand, for each $\omega \in \Omega$ the sequence $\{X_n(\omega)\}$ takes on 1 infinitely many times, hence

$$\mathbb{P}(\{\omega \in \Omega; X_n(\omega) \to X(\omega)\}) = \mathbb{P}(\varnothing) = 0.$$

Proposition 54. *If $X_n \xrightarrow{\text{P}} X$, there exists a non-random increasing sequence of integers $\{n_k\}$ so that $X_{n_k} \xrightarrow{\text{a.s.}} X$ as $k \to \infty$.*

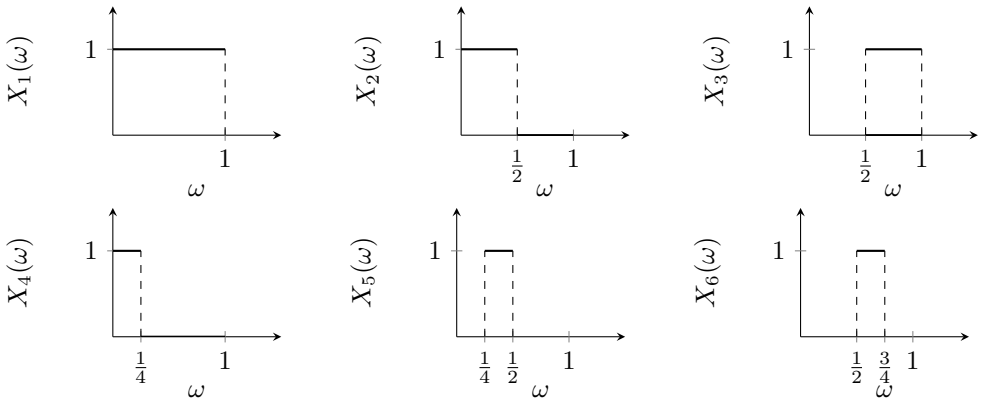


Figure 3: Graphs of $X_1, \ldots, X_6$ as defined in (20).

Proof. Since $X_n \xrightarrow{\mathrm{P}} X$, we have

$$\mathbb{P}\left(|X_n - X| > \varepsilon\right) \to 0 \qquad \text{as} n \to \infty.$$

Hence, for each $k \in \mathbb{N}$ we can find $n_k \in \mathbb{N}$ so that

$$\mathbb{P}\left(|X_{n_k} - X| > \frac{1}{k}\right) < \frac{1}{k^2}.$$

For any $\varepsilon > 0$

$$\sum_{1/k<\varepsilon} \mathbb{P}\left(|X_n - X| > \varepsilon\right) \leq \sum_{1/k<\varepsilon} \mathbb{P}\left(|X_n - X| > \frac{1}{k}\right) \leq \infty$$

since $\sum_k 1/k^2$ is convergent. By part 2 of proposition 53 we have $X_{n_k} \xrightarrow{\text{a.s.}} X$. □

Proposition 55 (Skorokhod's Representation Theorem)**.** *If $\{X_n\}$ and X, with distribution functions $\{F_n\}$ and F, are such that*

$$X_n \xrightarrow{\mathrm{D}} X \text{ (or equivalently } F_n \to F) \text{ as } n \to \infty$$

then there exists a probability space $(\Omega', \mathcal{F}', \mathbb{P}')$ and random variables $\{Y_n\}$ and Y mapping Ω' into $\mathbb{R}$, such that

1. *$\{Y_n\}$ and Y have distribution functions $\{F_n\}$ and F,*
2. *$Y_n \xrightarrow{\text{a.s.}} Y$ as $n \to \infty$.*

Proof. Add proof! □

Proposition 56. *If $X_n \xrightarrow{\text{D}} X$ and $g : \mathbb{R} \to \mathbb{R}$ is continuous, then $g(X_n) \xrightarrow{\text{D}} g(X)$.*

Proof. Take $(\Omega', \mathcal{F}', \mathbb{P}')$ and Y_n as in proposition 55, then since g is continuous we have

$$Y_n(\omega) \to Y(\omega) \implies g(Y_n(\omega)) \to g(Y(\omega))$$

hence

$$\{\omega \in \Omega'; Y_n(\omega) \to Y(\omega)\} \subseteq \{\omega \in \Omega'; g(Y_n(\omega)) \to g(Y(\omega))\}$$

so we get

$$\mathbb{P}'(\{\omega \in \Omega'; Y_n(\omega) \to Y(\omega)\}) = 1 \implies \mathbb{P}'(\{\omega \in \Omega'; g(Y_n(\omega)) \to g(Y(\omega))\}) = 1$$

and therefore $g(Y_n) \xrightarrow{\text{a.s.}} g(Y)$ on Ω'. This implies $g(Y_n) \xrightarrow{\text{D}} g(Y)$, but Y_n and X_n as well as Y and X are identically distributed, so we must have $g(X_n) \xrightarrow{\text{D}} g(X)$. □

Proposition 57. *The following three statemnets are equivalent.*

1. $X_n \xrightarrow{\text{D}} X$.
2. $\mathbb{E}[g(X_n)] \to \mathbb{E}[g(X)]$ *for all bounded continuous functions* g.
3. $\mathbb{E}[g(X_n)] \to \mathbb{E}[g(X)]$ *for all bounded functions* g *of the form* $g(x) = f(x)\mathbf{1}_{[a,b]}(x)$ *where* f *is continuous on* $[a, b]$ *and* a *and* b *are points of continuity of the distribution functio of the random variable* X.

Proof. Add if you can! □

11 Borel-Cantelli Lemmas and the Strong Law of Large Numbers

Given a sequence of events $\{A_n\}$, we can consider the probability of the following events

$$\limsup A_n = \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} A_m = \{\omega \in \Omega; \omega \in A_n \text{ for infinitely many } n\}$$

$$\liminf A_n = \bigcup_{n=1}^{\infty} \bigcap_{m=n}^{\infty} A_m = \{\omega \in \Omega; \omega \in A_n \text{ for all but finitely many } n\}.$$

The Borel-Cantelli lemmas relate the probabilities of these events to the series $\sum \mathbb{P}(A_n)$. The names lim sup and lim inf can be explained by noting that

$$\limsup_n \mathbf{1}_{A_n} = \mathbf{1}_{\limsup A_n} \qquad \liminf_n \mathbf{1}_{A_n} = \mathbf{1}_{\liminf A_n}$$

It is common to write $\limsup A_n = \{\omega \in \Omega; \omega \in A_n \text{ i.o.}\}$ where i.o. stands for infinitely often.

Proposition 58 (Borel-Cantelli Lemma)**.** *Given a sequence of events* $\{A_n\}$*,*

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty \implies \mathbb{P}(\limsup A_n) = 0.$$

That is, if the series $\sum \mathbb{P}(A_n)$ *converges to a real number,* $\{A_n\}$ *will occur finitely many times, almost surely.*

Proof. By sub-additivity of probability measure $\mathbb{P}$ we have

$$0 \leq \mathbb{P}\left(\bigcup_{m=n}^{\infty} A_m\right) \leq \sum_{m=n}^{\infty} \mathbb{P}(A_m).$$

Since the series $\sum \mathbb{P}(A_n)$ is convergent we can write the right-hand-side as

$$\sum_{m=n}^{\infty} \mathbb{P}(A_m) = \sum_{m=1}^{\infty} \mathbb{P}(A_m) - \sum_{m=1}^{n-1} \mathbb{P}(A_m)$$

which converges to zero as $n \to \infty$. Hence, by squeeze theorem we must have

$$\mathbb{P}\left(\bigcup_{m=n}^{\infty} A_m\right) \to 0 \qquad \text{as } n \to \infty,$$

and since

$$\bigcup_{m=n}^{\infty} A_m \downarrow \limsup A_n$$

by continuity from above of the probability measure $\mathbb{P}$ we have

$$\mathbb{P}(\limsup A_n) = \lim_{n\to\infty} \mathbb{P}\left(\bigcup_{m=n}^{\infty} A_m\right) = 0.$$

□

Remark. A shorter proof that using Fubini-Tonelli theorem is as follows: Let N be the number of events A_n that occur, given by

$$N = \sum_{n=1}^{\infty} \mathbf{1}_{A_n}.$$

Now consider

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

where we can switch the sum and expectation using Fubini-Tonelli theorem, noting that the sum is taken over non-negative random variables. Since N has finite expectation, it must be finite almost surely.

The second Borel-Cantelli lemma, provides a partial converse:

Proposition 59 (Second Borel-Cantelli Lemma)**.** *For independent events* $\{A_n\}$*,*

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty \implies \mathbb{P}(\limsup A_n) = 1.$$

Proof. Let B be the complement of $\limsup A_n$. By DeMorgan's laws we have

$$B = \left(\bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} A_m \right)^c = \bigcup_{n=1}^{\infty} \bigcap_{m=n}^{\infty} A_m^c.$$

Since the events $\{A_n\}$ are independent,

$$0 \leq \mathbb{P}\left(\bigcap_{m=n}^{\infty} A_m^c \right) = \prod_{m=n}^{\infty} \mathbb{P}(A_m^c) = \prod_{m=n}^{\infty} (1 - \mathbb{P}(A_m)).$$

Since $1 - x \leq e^{-x}$ for all x we can conclude

$$\prod_{m=n}^{\infty} (1 - \mathbb{P}(A_m)) \leq \prod_{m=n}^{\infty} e^{-\mathbb{P}(A_m)} = e^{-\sum m=n^{\infty} \mathbb{P}(A_m)}$$

which converges to zero as $n \to \infty$. Hence

$$\mathbb{P}\left(\bigcap_{m=n}^{\infty} A_m^c \right) \to 0.$$

Since

$$\bigcap_{m=n}^{\infty} A_m^c \uparrow B$$

by the continuity from below of the probability measure $\mathbb{P}$ we conclude that

$$\mathbb{P}(B) = \lim_n \bigcap_{m=n}^{\infty} A_m^c = 0 \implies \mathbb{P}(\limsup A_n) = 1.$$

□

Proposition 60.

Proposition 61 (Strong Law of Large Numbers)**.**

Example 10 (Renewal Theory)**.**

Proposition 62.

Proposition 63 (Glivenko-Cantelli Theorem)**.**

12 Information Theory

Proposition 64 (Shannon's Theorem)**.** *Let $X_1, X_2, \ldots, \in \{1, \ldots, r\}$ be independent with $\mathbb{P}(X_i = k) = p(k) > 0$ for $1 \leq k \leq r$. Here we are thinking of $1, \ldots, r$ as the letters of an alphabet, and $X_1, X_2, \ldots$ are successive letters produced by an information source (in this i.i.d. case the proverbial monkey at a typewriter). Let $\pi_n(\omega) = p(X_1(\omega)) \ldots p(X_n(\omega))$ be the probability of the realization we observed in the first n trials. Since $\log \pi_n(\omega)$ is a sum of independent random variables it follows from the strong law of large numbers that*

$$-n^{-1} \log \pi_n \xrightarrow{\text{a.s.}} H := -\sum_{k=1}^{r} p(k) \log p(k)$$

*The constant H is called the **entropy** of the source and is a measure of how random it is. The last result is known as the **asymptotic equipartition property**: if $\varepsilon > 0$ then as $n \to \infty$*

$$\mathbb{P}\left\{\exp(-n(H+\varepsilon)) \leq \pi_n(\omega) \leq \exp(-n(H-\varepsilon))\right\} \to 1.$$

The rest of this section is devoted to the study of entropies and related concept. We only include discrete random variables in our analysis and adopt the following notation: given a discrete random variable X, we use $\mathscr{X}$ to denote its support and $p_X(x) = \mathbb{P}(X = x)$ to show its probability mass function.

Definition 17. The **entropy** $H(X)$ of a discrete random variable X is defined by

$$H(X) = -\sum_{x \in \mathscr{X}} p(x) \log p(x).$$

Remark. Since $H(X)$ only depends on the distribution of X, we also write $H(p)$ for the above quantity. The logarithm is to the base 2 and the entropy is expressed in "bits." For example, the entropy of a fair coin toss is 1 bit.

Remark. We will use the convention that $0 \log 0 = 0$, which is easily justified by continuity since $x \log x \to 0$ as $x \to 0$.

Remark. The entropy of X can also be interpreted as the expected value of $\log \frac{1}{p(X)}$. Thus

$$H(X) = \mathbb{E}\left[\log \frac{1}{p(X)}\right].$$

We can think of $-\log \frac{1}{\mathbb{P}(()A)}$ as the "information content" of event A. From this perspective, entropy is the average information content of the realizations of the random variable X.

1. The information content of an event is a decreasing

Lemma 1.

$$H(X) \geq 0$$

and equality holds if and only if X is a constant.

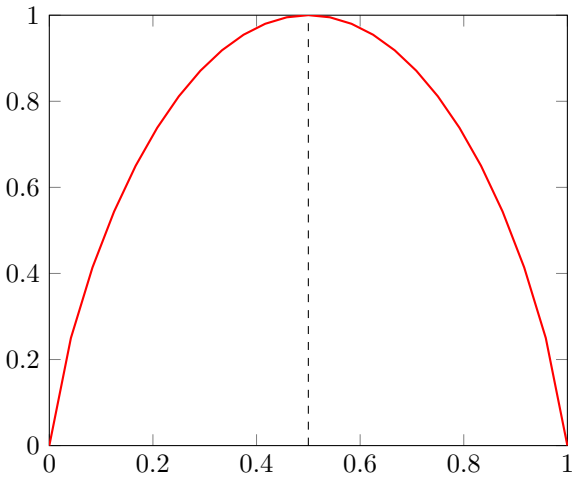


Figure 4: The binary entropy function.

Proof. $0 \leq p(x) \leq 1$ implies $-\log p(x) \geq 0$. Clearly, if $X = x$ with probability one we get $H(x) = 1 \log 1 = 0$. On the other hand, if $H(X) = 0$ we must have $p(x) \log p(x) = 0$ for all $x \in \mathscr{X}$. Since by definition of $\mathscr{X}$, we must have $p(x) > 0$ for $x \in \mathscr{X}$, we conclude that we must have $\log p(x) = 0$, i.e. $p(x) = 1$. So $\mathscr{X}$ is a singleton and X is constant. □

Example 11 (Binary Entropy Function)**.** Let $X \sim \text{Bernoulli}(p)$, then

$$H(X) = -p \log p - (1-p) \log(1-p) := H(p).$$

In particular, H attains it maximum value of 1 at $p = 1/2$. Refer to figure, 4 for the graph of H.

Example 12. Let

$$X = \begin{cases} a & \text{with probability } 1/2, \\ b & \text{with probability } 1/4, \\ c & \text{with probability } 1/8, \\ d & \text{with probability } 1/8, \end{cases}$$

The entropy of X is

$$H(X) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{4}\log\frac{1}{4} - \frac{1}{8}\log\frac{1}{8} - \frac{1}{8}\log\frac{1}{8} = \frac{7}{4}\text{bits}.$$

Remark. Suppose we wish to determine the value of X with the minimum number of binary questions. An efficient first question is "Is $X = a$?" This splits the probability in half. If the answer to the first question is no, then the second question can be "Is $X = b$?" The third question can be "Is $X = c$?" The

resulting expected number of binary questions required is 1.75. This turns out to be the minimum expected number of binary questions required to determine the value of X.

Definition 18. The **joint entropy** $H(X,Y)$ of a pair of discrete random variables (X,Y) with a joint distribution $p(x,y)$ is defined as

$$H(X,Y) = -\sum_{x\in\mathscr{X}}\sum_{y\in\mathscr{Y}} p(x,y)\log p(x,y),$$

which can also be expressed as

$$H(X,Y) = -\mathbb{E}\left[\log p(X,Y)\right].$$

Definition 19. The **conditional entropy** $H(Y|X)$ of a pair of discrete random variables (X,Y) with a joint distribution $p(x,y)$ is defined as

$$\begin{aligned} H(Y|X) &= \sum_{x\in\mathscr{X}} p_X(x) H(Y|X=x) \\ &= -\sum_{x\in\mathscr{X}} p_X(x) \sum_{y\in\mathscr{Y}} p_{Y|X}(y|x)\log p_{Y|X}(y|x) \\ &= \sum_{x\in\mathscr{X}}\sum_{y\in\mathscr{Y}} p_{X,Y}(x,y)\log p_{Y|X}(y|x) \\ &= -\mathbb{E}\left[\log p(Y|X)\right] \end{aligned}$$

Remark. Note that $H(Y|X) \geq 0$ and equality holds if and only if given X, Y takes on a single value with probability 1. That is, there exists a function f so that $Y = f(X)$.

Proposition 65 (Chain Rule)**.**

$$H(X,Y) = H(X) + H(Y|X).$$

Proof.

$$\begin{aligned} H(X,Y) &= \sum_{x\in\mathscr{X}}\sum_{y\in\mathscr{Y}} p(x,y)\log p(x,y) \\ &= \sum_{x\in\mathscr{X}}\sum_{y\in\mathscr{Y}} p(x,y)\log\left(p(x)p(y|x)\right) \\ &= \sum_{x\in\mathscr{X}}\sum_{y\in\mathscr{Y}} p(x,y)\log p(x) + \sum_{x\in\mathscr{X}}\sum_{y\in\mathscr{Y}} p(x,y)\log p(y|x) \\ &= \sum_{x\in\mathscr{X}} p(x)\log p(x) + \sum_{x\in\mathscr{X}}\sum_{y\in\mathscr{Y}} p(x,y)\log p(y|x) \\ &= H(X) + H(Y|X). \end{aligned}$$

Corollary 4.

$$H(X,Y|Z) = H(X|Z) + H(Y|X,Z).$$

Remark. Note that $H(X|Y) \neq H(Y|X)$ but

$$H(X) - H(X|Y) = H(Y) - H(Y|X).$$

□

Definition 20. The **relative entropy** or **Kullback-Leibler distance** between two probability mass functions $p(x)$ and $q(x)$ is defined as

$$D(p||q) = \sum_{x \in \mathscr{X}} p(x) \log \frac{p(x)}{q(x)} = \mathbb{E}_p \left[\log \frac{p(x)}{q(x)} \right].$$

Remark. Note that even though we use the term "distance" to refer to relative entropy, it is not a metric on the space of probability mass functions. The reason is that it is not symmetric, i.e. in general $D(p||q) \neq D(q||p)$. For this reason, the relative entropy is sometimes referred to as the **Kullback-Leibler divergence**.

Definition 21. Consider two random variables X and Y with a joint probability mass function $p_{X,Y}(x,y)$ and marginal probability mass functions $p_X(x)$ and $p_Y(y)$. The **mutual information** $I(X;Y)$ is the relative entropy between the

joint distribution and the product distribution $p_X(x)p_Y(y)$, i.e.

$$\begin{aligned} I(X;Y) &= D(p_{X,Y}(x,y)||p_X(x)p_Y(y)) \\ &= \sum_{x \in \mathscr{X}} \sum_{y \in \mathscr{Y}} p_{X,Y}(x,y) \log \frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)} \\ &= \mathbb{E}_{p_{X,Y}(x,y)} \left[\log \frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)} \right]. \end{aligned}$$

Remark. We can relate mutual information to entropy as

$$\begin{aligned} I(X;Y) &= D(p_{X,Y}(x,y)||p_X(x)p_Y(y)) \\ &= \sum_{x \in \mathscr{X}} \sum_{y \in \mathscr{Y}} p_{X,Y}(x,y) \log \frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)} \\ &= \sum_{x \in \mathscr{X}} \sum_{y \in \mathscr{Y}} p_{X,Y}(x,y) \log \frac{p_{X|Y}(x|y)}{p_X(x)} \\ &= \sum_{x \in \mathscr{X}} \sum_{y \in \mathscr{Y}} p_{X,Y}(x,y) \left(\log p_{X|Y}(x|y) - p_{X,Y}(x,y) p_X(x) \right) \\ &= H(X) - H(X|Y), \end{aligned}$$

or similarly

$$I(X;Y) = H(Y) - H(Y|X).$$

Another way to relate mutual information to entropy is

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

since

$$\begin{aligned}
H(Y|X) &= -\sum_{x\in\mathscr{X}}\sum_{y\in\mathscr{Y}} p_{X,Y}(x,y)\log p_{Y|X}(y|x) \\
&= -\sum_{x\in\mathscr{X}}\sum_{y\in\mathscr{Y}} p_{X,Y}(x,y)\log p_{X,Y}(x,y) + \sum_{x\in\mathscr{X}}\sum_{y\in\mathscr{Y}} p_{X,Y}(x,y)\log p_X(x) \\
&= H(X,Y) - H(X).
\end{aligned}$$

Remark. Note that

$$I(X;X) = H(X) - H(X|X) = H(X)$$

so entropy is sometimes called self information.

To summarize

Proposition 66.

$$\begin{aligned}
I(X;Y) &= H(X) - H(X|Y) \\
I(X;Y) &= H(Y) - H(Y|X) \\
I(X;Y) &= H(X) + H(Y) - H(X,Y) \\
I(X;Y) &= I(Y;X) \\
I(X;X) &= H(X)
\end{aligned}$$

Proposition 67 (Chain Rule for Entropy)**.** *Let $X_1, X_2, \ldots, X_n$ be drawn according to $p(x_1, \ldots, x_n)$. Then,*

$$H(X_1, X_2, \ldots, X_n) = H(X_1) + \sum_{i=2}^{n} H(X_i|X_1, \ldots, X_{i-1}).$$

Proof. We can use induction on n. We have already established that the proposition holds for $n = 2$. Suppose the proposition holds for $n < k$ for $k \geq 2$ and consider k random variables $X_1, X_2, \ldots, X_k$ drawn according to $p(x_1, \ldots, x_k)$. Then,

$$\begin{aligned}
H(X_1, \ldots, X_k) &= -\sum_{x_1,\ldots,x_k\in\mathscr{X}^k} p(x_1, \ldots, x_k)\log p(x_1, \ldots, x_k) \\
&= -\sum_{x_1\in\mathscr{X}}\sum_{x_2,\ldots,x_k\in\mathscr{X}^{k-1}} p(x_1)p(x_2, \ldots, x_k|x_1)\log[p(x_1)p(x_2, \ldots, x_k|x_1)] \\
&= -\sum_{x_1\in\mathscr{X}} p(x_1)\log p(x_1) \sum_{x_2,\ldots,x_k\in\mathscr{X}^{k-1}} p(x_2, \ldots, x_k|x_1) \\
&\quad -\sum_{x_1\in\mathscr{X}} p(x_1) \sum_{x_2,\ldots,x_k\in\mathscr{X}^{k-1}} p(x_2, \ldots, x_k|x_1)\log p(x_2, \ldots, x_k|x_1) \\
&= H(X_1) + H(X_2, \ldots, X_k|X_1).
\end{aligned}$$

Noting the definition of conditional entropy and applying the induction hypothesis to $H(X_2, \ldots, X_k|X_1)$ finishes the proof. □

A similar result holds for mutual information, which uses the concept of conditional mutual information.

Definition 22. The **conditional mutual information** of random variables X and Y given Z is defined by

$$I(X,Y;Z) = H(X|Z) - H(X|Y,Z) = \mathbb{E}_{p(x,y,z)} \log \frac{p(X,Y|Z}{p(X|Z)p(Y|Z)}.$$

Proposition 68 (Chain Rule for Mutual Information)**.**

$$I(X_1,\ldots,X_n;Y) = \sum_{i=1}^{n} I(X_i;Y|X_1,\ldots,X_{i-1})$$

Proof.

$$\begin{aligned} I(X_1,\ldots,X_n;Y) &= H(X_1,\ldots,X_n) - H(X_1,\ldots,X_n|Y) \\ &= \sum_{i=1}^{n} H(X_i|X_1,\ldots,X_{i-1}) - \sum_{i=1}^{n} H(X_i|X_1,\ldots,X_{i-1},Y) \\ &= \sum_{i=1}^{n} [H(X_i|X_1,\ldots,X_{i-1}) - H(X_i|X_1,\ldots,X_{i-1},Y)] \\ &= \sum_{i=1}^{n} I(X_i;Y|X_1,\ldots,X_{i-1}). \end{aligned}$$

□

Proposition 69 (Information Inequality)**.** *Let $p(x)$ and $q(x)$ be two probability mass functions for $x \in \mathscr{X}$. Then*

$$D(p||q) \geq 0$$

with equality if and only if

$$\forall x \in \mathscr{X}; p(x) = q(x).$$

Proof. Letting $X \sim p$ and noting that $f(x) = -\log x$ is a convex function, we can apply Jensen's inequality to random variables $q(X)/p(X)$ to get

$$-\log \mathbb{E}\left[\frac{q(X)}{p(X)}\right] \leq \mathbb{E}\left[-\log \frac{q(X)}{p(X)}\right] = -\sum_{x\in\mathscr{X}} p(x) \log \frac{q(x)}{p(x)} = D(p||q),$$

but

$$\mathbb{E}\left[\frac{q(X)}{p(X)}\right] = \sum_{x\in\mathscr{X}} p(x)\frac{q(x)}{p(x)} = \sum_{x\in\mathscr{X}} q(x) = 1.$$

So,

$$0 = -\log 1 = -\log \mathbb{E}\left[\frac{q(X)}{p(X)}\right] \leq= -\sum_{x\in\mathscr{X}} p(x) \log \frac{q(x)}{p(x)} = D(p||q).$$

Equality only holds when $q(X)/p(X) = 1$, due to strict convexity of f. □

Corollary 5 (Non-negativity of Mutual Information)**.** For any two random variables X and Y,

$$I(X;Y) \geq 0,$$

with equality if and only if X and Y are independent

Proof. $I(X;Y) = D(p_{X,Y}(x,y)||p_X(x)p_Y(y)) \geq 0$ with equality if and only if $p_{X,Y}(x,y) = p_X(x)p_Y(y)$; i.e. X and Y being independent. □

13 Markov Processes

Let $\{X_i; i \in \mathbb{Z}^{\geq 0}\}$ be a sequence of discrete random variables that take values in some countable set S called the **state space**.

Definition 23. The process X is a **Markov chain** if is satisfies the **Markov condition**:

$$\mathbb{P}\left(X_{n+1} = s | X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n\right) = \mathbb{P}\left(X_{n+1} = s | X_n = x_n\right).$$

for all $n \geq 0$ and $s \in S$.

Remark. By conditioning on $X_{n+1}, \ldots, X_{n+m-1}$ we can see that the Markov property is equivalent to

$$\mathbb{P}\left(X_{m+n} = s | X_0 = x_0, \ldots, X_m = x_m\right) = \mathbb{P}\left(X_{m+n} = s | X_m = x_m\right)$$

for any $m, n \geq 0$. Using this property we can show that for any increasing sequence of integers $n_1 < n_2 < \ldots < n_k \leq n$ we have

$$\begin{aligned}
&\mathbb{P}\left(X_n = s | X_{n_1} = x_1, \ldots, X_{n_k} = x_k\right) \\
&= \sum_{x_j; j \in \mathcal{J}} \mathbb{P}\left(X_n = s | X_0 = x_0, \ldots, X_{n_k} = x_k\right) \mathbb{P}\left(\cup_{j \in \mathcal{J}} X_j = x_j | X_{n_1} = x_1, \ldots, X_{n_k} = x_k\right) \\
&= \mathbb{P}\left(X_n = s | X_{n_k} = x_k\right) \sum_{x_j; j \in \mathcal{J}} \mathbb{P}\left(\cup_{j \in \mathcal{J}} X_j = x_j | X_{n_1} = x_1, \ldots, X_{n_k} = x_k\right) \\
&= \mathbb{P}\left(X_n = s | X_{n_k} = x_k\right).
\end{aligned}$$

where $\mathcal{J}$ is the set of all non-negative integers less than n_k that don't appear in the sequence $\{n_k\}$.

Definition 24. The chain X is called **homogeneous** if for any $n \in \mathbb{Z}^{\geq 0}$ and $i, j \in S$

$$\mathbb{P}\left(X_{n+1} = j | X_n = i\right) = \mathbb{P}\left(X_1 = j | X_0 = i\right).$$

The **transition matrix** $\mathbf{P} = (p_{ij})$ is the $|S| \times |S|$ matrix of **transition probabilities**

$$p_{ij} = \mathbb{P}\left(X_{n+1} = j | X_n = i\right).$$

Throughout the rest of this section, Markov chains are assumed to be homogeneous unless otherwise specified.

Remark. Given the above definition, we observe that the i-th row of the transition matrix specifies the conditional probability distribution of X_{n+1} given $X_n = i$, hence

1. Every element of the i-th row, and therefore every element of $\mathbf{P}$ must be non-negative.
2. The elements of the i-th row should sum up to one; i.e. $\sum_j p_{ij} = 1$ for all i.

Definition 25. The **n-step transition matrix** $\mathbf{P}(m, m+n) = (p_{ij}(m, m+n))$ is the matrix of n-step transition probabilities $(p_{ij}(m, m+n)) = \mathbb{P}(X_{m+n} = j | X_m = i)$.

Proposition 70 (Chapman-Kolmogorov Equations)**.**

$$p_{ij}(m, m+n+r) = \sum_k p_{ik}(m, m+r) p_{kj}(m+r, m+n+r)/$$

Therefore, $\mathbf{P}(m, m+n+r) = \mathbf{P}(m, m+r)\mathbf{P}(m+r, m+n+r)$*, and* $\mathbf{P}(m, m+n) = \mathbf{P}^n$*, the n-th power of* $\mathbf{P}$*.*

Proof. Conditioning on X_{m+r} when when calculating $\mathbb{P}(X_{m+n+r} = j | X_m = i)$ we get

$$\begin{aligned}\mathbb{P}(X_{m+n+r} = j | X_m = i) &= \sum_k \mathbb{P}(X_{m+n+r} = j | X_{m+r} = k, X_m = i)\, \mathbb{P}(X_{m+r} = j | X_m = i) \\ &= \sum_k \mathbb{P}(X_{m+n+r} = j | X_{m+r} = k)\, \mathbb{P}(X_{m+r} = j | X_m = i)\end{aligned}$$

where to get the last equality we have used the stronger form of Markov condition as discussed in the remarks after definition 23. □

Lemma 2. *Let* $\mu_i^{(n)} = \mathbb{P}(X_n = i)$ *be the mass function of* X_n*, and write* $\mu^{(n)}$ *for the row vector with entries* $(\mu_i^{(n)} : i \in S)$*. Then,*

$$\mu^{(m+n)} = \mu^{(m)}\mathbf{P}^n \qquad \mu^{(n)} = \mu^{(0)}\mathbf{P}^n$$

Proof.

$$\begin{aligned}\mu_i^{(m+n)} = \mathbb{P}(X_{m+n} = i) &= \sum_j \mathbb{P}(X_{m+n} = i | X_m = j)\, \mathbb{P}(X_m = j) \\ &= \sum_j \mu_j^{(m)} \mathbf{P}_{ji} = (\mu^{(m)}\mathbf{P})_i.\end{aligned}$$

□

Example 13 (Simple Random Walk)**.** The simple random walk on integers has state space $S = \mathbb{Z}$ and transition probabilities

$$p_{ij} = \begin{cases} p & j = i+1 \\ q = 1-p & j = i-1 \\ 0 & \text{o.w.} \end{cases}$$

If we start at state i and after n steps end up at state i having taken x steps forward and y steps backward, we must have

$$\begin{cases} x+y & = n \\ i+x-y & = j \end{cases}$$

and hence

$$x = \frac{n+j-i}{2}.$$

Therefore, the n-step transition probabilities are given by

$$p_{ij}(n) = \begin{cases} \binom{n}{x} p^x q^{n-x} & 2|n+j-i \\ 0 & \text{o.w.} \end{cases} \quad \text{where} \quad x = \frac{n+j-i}{2}.$$

Example 14 (Branching Process)**.** We can model a branching process where G is the probability generating function of the offspring distribution as a Markov chain with state space being $S = \mathbb{Z}^{\geq 0}$ and the transition probability p_{ij} being the coefficient of s^j in $G(s)^i$. Also, since we know that the probability generating function of the offspring distribution after n generations is G_n, which is G composed with itself for n times, we can conclude that the n-step transition probability $p_{ij}(n)$ is the coefficient of s^j in $G_n(s)^i$.

13.1 Classification of States

Definition 26. State i is called **persistent** (or **recurrent**) if

$$\mathbb{P}\left(X_n = i \text{for some } n \geq 1 | X_0 = i\right) = 1,$$

which is to say that the probability of eventual return to i, having started from i, is 1. If this probability is strictly less than 1, the state is called **transient**.

Similar to our analysis of the simple random walk, we can define $f_{ij}(n)$ to be the probability of returning to j after n steps for the first time, starting from i; that is,

$$f_{ij}(n) = \mathbb{P}\left(X_1 \neq j, X_2 \neq j, \ldots, X_{n-1} \neq j, X_n = j | X_0 = i\right).$$

We can express the probability of ever visiting state j starting from state i, f_{ij}, as

$$f_{ij} = \sum_{n=1}^{\infty} f_{ij}(n).$$

So j is persistent if and only if $p_{jj} = 1$. Our objective is to relate persistence to the transition probabilities. Consider the probability generating functions

$$P_{ij}(s) = \sum_{n=0}^{\infty} p_{ij}(n) s^n \qquad F_{ij}(s) = \sum_{n=0}^{\infty} f_{ij}(n) s^n$$

where we have adopted the convention that $f_{ij}(0) = 0$ and $p_{ij}(0) = \delta_{ij}$. We have the following result.

Proposition 71.

$$P_{ij}(s) = \delta_{ij} + F_{ij}(s)P_{jj}(s)$$

Proof. Conditioning on the first time we visit state j in $p_{ij}(n)$ for $n \geq 1$ we have

$$p_{ij}(n) = \sum_{k=1}^{n} f_{ij}(k)p_{jj}(n-k).$$

Given our convention that $f_{ij}(0) = 0$ we can start the sum from $k = 0$ and observe that

$$p_{ij}(n)s^n = \sum_{k=0}^{n} f_{ij}(k)s^k p_{jj}(n-k)s^{n-k}$$

so

$$\begin{aligned}\sum_{n=0}^{\infty} p_{ij}(n)s^n &= \delta_{ij}s^0 + \sum_{n=1}^{\infty}\sum_{k=0}^{n} f_{ij}(k)s^k p_{jj}(n-k)s^{n-k} \\ &= \delta_{ij} + F_{ij}(s)P_{jj}(s) - f_{ij}(0)p_{jj}(0) \\ &= \delta_{ij} + F_{ij}(s)P_{jj}(s).\end{aligned}$$

□

Proposition 72. *1. State j is persistent if $\sum_n p_{jj}(n) = \infty$, and if this holds $\sum_n p_{ij}(n) = \infty$ for all i such that $f_{ij} > 0$.*

2. State j is transient if $\sum_n p_{jj}(n) < \infty$, and if this holds then $\sum_n p_{ij}(n) < \infty$ for all i.

Proof. 1. From the proposition we have

$$P_{jj}(s) = 1 + P_{jj}(s)F_{jj}(s) \implies P_{jj}(s) = \frac{1}{1 - F_{jj}(s)}.$$

Note that

$$\lim_{s\to 1} P_{jj}(s) = \infty \iff F_{jj}(s) = 1.$$

Since P_{jj} is a probability generating function, we must have

$$P_{jj}(1) = \lim_{s\to 1} P_{jj}(s)$$

Note that $F_{jj}(s) = 1$ if and only if state j is persistent. Also, note that if state j is persistent

$$\lim_{s\to 1} P_{ij}(s) = \lim_{s\to 1} \delta_{ij} + F_{ij}(s)P_{jj}(s) = \infty.$$

so $\sum_n p_{ij}(n) = \infty$.

2. Similar to part 1, state j is transient if and only if $F_{jj}(1) < 1$ which is equivalent to

$$P_{jj}(1) = \lim_{s \to 1} P_{jj}(s) < \infty.$$

which also implies that

$$\lim_{s \to 1} P_{ij}(s) = \lim_{s \to 1} \delta_{ij} + F_{ij}(s) P_{jj}(s) < \infty.$$

□

Remark. Use Borel-Cantelli lemmas to prove.