ECON2125/8013 Maths Notes

John Stachurski

January 26, 2015

Contents

Chapter 1

Linear Algebra

1.1 Vectors Space

[roadmap]

1.1.1 Vectors

Let's quickly review the basics. An important set for us will be, for arbitrary $N \in \mathbb{N}$, the set of all N-vectors, or vectors of length N. This set is denoted by $\mathbb{R}^N$, and a typical element is of the form

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix} \quad \text{where } x_n \in \mathbb{R} \text{ for each } n$$

(As usual, $\mathbb{R} = \mathbb{R}^1$ represents the set of all real numbers, which is, in essence, the union of the rational and irrational numbers.) Here $\mathbf{x}$ has been written vertically, as a column of numbers, but we could also write it horizontally, like so: $\mathbf{x} = (x_1, \ldots, x_N)$. At this stage, we are viewing vectors just as sequences of numbers, so it makes no difference whether they are written vertically or horizontally. It's only when we get to matrix multiplication that we have to concern ourselves with distinguishing between column (vertical) and row (horizontal) vectors.

The vector of ones will be denoted $\mathbf{1}$, while the vector of zeros will be denoted $\mathbf{0}$:

$$\mathbf{1} := \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \qquad \mathbf{0} := \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$$

For elements of $\mathbb{R}^N$ there are two fundamental algebraic operations: addition and scalar multiplication. If $\mathbf{x} \in \mathbb{R}^N$ and $\mathbf{y} \in \mathbb{R}^N$, then the **vector sum** is defined by

$$\mathbf{x} + \mathbf{y} :=: \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix} + \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} := \begin{pmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \vdots \\ x_N + y_N \end{pmatrix}$$

If $\alpha \in \mathbb{R}$, then the **scalar product** of α and $\mathbf{x}$ is defined to be

$$\alpha\mathbf{x} := \begin{pmatrix} \alpha x_1 \\ \alpha x_2 \\ \vdots \\ \alpha x_N \end{pmatrix}$$

Thus, addition and scalar multiplication are defined in terms of ordinary addition and multiplication in $\mathbb{R}$, and computed element-by-element, by adding and multiplying respectively. Figures 1.1 and 1.1 show examples of vector addition and scalar multiplication in the case $N = 2$. In the figure, vectors are represented as arrows, starting at the origin and ending at the location in $\mathbb{R}^2$ defined by the vector.

Remark: In some instances, the notion of scalar multiplication includes multiplication of vectors by complex numbers. In what follows we will work almost entirely with real scalars, and hence scalar multiplication means real scalar multiplication unless otherwise stated.

We have defined addition and scalar multiplication of vectors, but not subtraction. Subtraction is performed element by element, analogous to addition. The definition can be given in terms of addition and scalar multiplication. $\mathbf{x} - \mathbf{y} := \mathbf{x} + (-1)\mathbf{y}$. An illustration of this operation is given in figure 1.3. One way to remember this is to draw a line from $\mathbf{y}$ to $\mathbf{x}$, and then shift it to the origin.

The **inner product** of two vectors $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^N$ is denoted by $\mathbf{x}'\mathbf{y}$, and defined as the sum of the products of their elements:

$$\mathbf{x}'\mathbf{y} := \sum_{n=1}^{N} x_n y_n$$

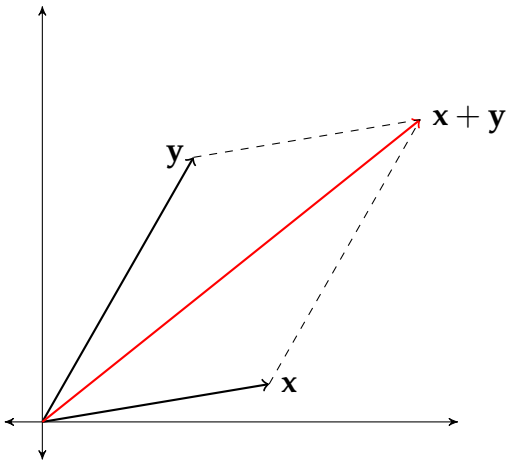


Figure 1.1: Vector addition

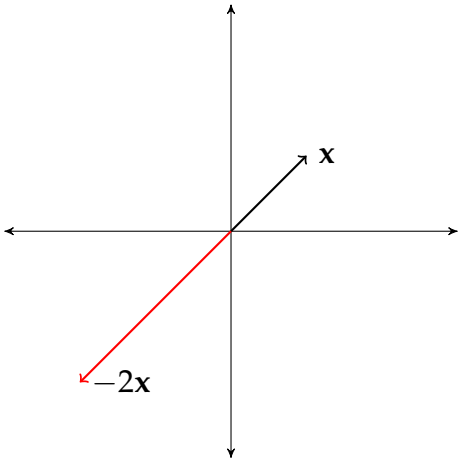


Figure 1.2: Scalar multiplication

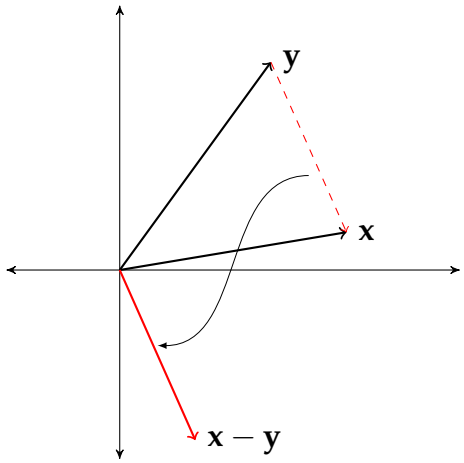


Figure 1.3: Difference between vectors

Fact 1.1.1. For any $\alpha, \beta \in \mathbb{R}$ and any $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^N$, the following statements are true:

1. $\mathbf{x}'\mathbf{y} = \mathbf{y}'\mathbf{x}$
2. $(\alpha\mathbf{x})'(\beta\mathbf{y}) = \alpha\beta(\mathbf{x}'\mathbf{y})$
3. $\mathbf{x}'(\alpha\mathbf{y} + \beta\mathbf{z}) = \alpha(\mathbf{x}'\mathbf{y}) + \beta(\mathbf{x}'\mathbf{z})$

These properties are easy to check from the definition. However, since this is our first formal claim let's step through one of the arguments carefully, just as an exercise: Consider the second equality, which is $(\alpha\mathbf{x})'(\beta\mathbf{y}) = \alpha\beta(\mathbf{x}'\mathbf{y})$. The claim is that the stated equality holds for any $\alpha, \beta \in \mathbb{R}$ and any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$. To verify a "for any" claim we have to show that the stated property holds for *arbitrary* choices of these objects. Thus we start by saying *pick any* $\alpha, \beta \in \mathbb{R}$ and *any* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$. By the definitions of scalar multiplication and inner product respectively, we have

$$(\alpha\mathbf{x})'(\beta\mathbf{y}) = \sum_{n=1}^{N} \alpha x_n \beta y_n$$

Passing the scalars out of the sum, we see that this equals $\alpha\beta\sum_{n=1}^{N} x_n y_n$, which, using the definition of inner product again, is $\alpha\beta(\mathbf{x}'\mathbf{y})$. This verifies the claim.

Perhaps this style of argument seemed like overkill for such a simple statement. But getting into the habit of constructing careful arguments will help you a great deal

as we work through more complex material. One thing you'll find in particular if you're still new to formal reasoning is that getting the first sentence of a proof right (establishing exactly what it is you're going to show) is critical. If you get this right the rest should flow more easily.

The (Euclidean) **norm** of a vector $\mathbf{x} \in \mathbb{R}^N$ is defined as

$$\|\mathbf{x}\| := \sqrt{\mathbf{x}'\mathbf{x}} := \left(\sum_{n=1}^{N} x_n^2 \right)^{1/2} \tag{1.1}$$

and represents the length of the vector $\mathbf{x}$. (In the arrow representation of vectors in figures 1.1–1.3, the norm of the vector is equal to the length of the arrow.)

Fact 1.1.2. For any $\alpha \in \mathbb{R}$ and any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$, the following statements are true:

1. $\|\mathbf{x}\| \geq 0$ and $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} = \mathbf{0}$
2. $\|\alpha\mathbf{x}\| = |\alpha|\|\mathbf{x}\|$
3. $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$
4. $|\mathbf{x}'\mathbf{y}| \leq \|\mathbf{x}\|\|\mathbf{y}\|$

The first two properties you can verify yourself without difficulty. Proofs for the second two are a bit harder. The third property is called the **triangle inequality**, while the fourth is called the **Cauchy-Schwarz inequality**. The proof of the Cauchy-Schwarz inequality is given as a solved exercise after we've built up some more tools (see exercise 3.6.13 below). If you're prepared to accept the Cauchy-Schwarz inequality for now, then the triangle inequality follows, because, by the properties of the inner product given in fact 1.1.1,

$$\|\mathbf{x} + \mathbf{y}\|^2 = (\mathbf{x} + \mathbf{y})'(\mathbf{x} + \mathbf{y}) = \mathbf{x}'\mathbf{x} + 2\mathbf{x}'\mathbf{y} + \mathbf{y}'\mathbf{y} \leq \mathbf{x}'\mathbf{x} + 2|\mathbf{x}'\mathbf{y}| + \mathbf{y}'\mathbf{y}$$

Applying the Cauchy-Schwarz inequality leads to $\|\mathbf{x} + \mathbf{y}\|^2 \leq (\|\mathbf{x}\| + \|\mathbf{y}\|)^2$. Taking the square root gives the triangle inequality.

Given two vectors $\mathbf{x}$ and $\mathbf{y}$, the value $\|\mathbf{x} - \mathbf{y}\|$ has the interpretation of being the "distance" between these points. To see why, consult figure 1.3 again.

1.1.2 Spans and Linear Subspaces

One of the most elementary ways to work with vectors is to combine them using linear operations. Given K vectors $\mathbf{x}_1, \ldots, \mathbf{x}_K$ in $\mathbb{R}^N$, a **linear combination** of these vectors is a new vector of the form

$$\mathbf{y} = \sum_{k=1}^{K} \alpha_k \mathbf{x}_k = \alpha_1 \mathbf{x}_1 + \cdots + \alpha_K \mathbf{x}_K$$

for some collection of scalars $\alpha_1, \ldots, \alpha_K$ (i.e., with $\alpha_k \in \mathbb{R}$ for all k).

Fact 1.1.3. Inner products of linear combinations satisfy the following rule:

$$\left(\sum_{k=1}^{K} \alpha_k \mathbf{x}_k \right)' \left(\sum_{j=1}^{J} \beta_j \mathbf{y}_j \right) = \sum_{k=1}^{K} \sum_{j=1}^{J} \alpha_k \beta_j \mathbf{x}_k' \mathbf{y}_j$$

Given any nonempty $X \subset \mathbb{R}^N$, the set of all vectors that can be made by (finite) linear combinations of elements of X is called the **span** of X, and denoted by $\text{span}(X)$. For example, the set of all linear combinations of $X := \{\mathbf{x}_1, \ldots, \mathbf{x}_K\}$ is

$$\text{span}(X) := \left\{ \text{all vectors } \sum_{k=1}^{K} \alpha_k \mathbf{x}_k \text{ such that } \boldsymbol{\alpha} := (\alpha_1, \ldots, \alpha_K) \in \mathbb{R}^K \right\}$$

Example 1.1.1. Let $X = \{\mathbf{1}\} = \{(1,1)\} \subset \mathbb{R}^2$. The span of X is all vectors of the form $\alpha \mathbf{1} = (\alpha, \alpha)$ with $\alpha \in \mathbb{R}$. This constitutes a line in the plane. Since we can take $\alpha = 0$, it follows that the origin $\mathbf{0}$ is in $\text{span}(X)$. In fact $\text{span}(X)$ is the unique line in the plane that passes through both $\mathbf{0}$ and the vector $\mathbf{1} = (1,1)$.

Example 1.1.2. Let $\mathbf{x}_1 = (3,4,2)$ and let $\mathbf{x}_2 = (3,-4,0.4)$. This pair of vectors is shown on the left-hand side of figure 1.4. The span of $\{\mathbf{x}_1, \mathbf{x}_2\}$ is a plane in $\mathbb{R}^3$ that passes through both of these vectors and the origin. The plane is shown on the right-hand side of figure 1.4.

Let Y be any subset of $\mathbb{R}^N$, and let $X := \{\mathbf{x}_1, \ldots, \mathbf{x}_K\}$. If $Y \subset \text{span}(X)$, we say that the vectors in X **span the set** Y, or that X is a **spanning set** for Y. This is a particularly nice situation when Y is large but X is small, because it means that all the vectors in the large set Y are "described" by the small number of vectors in X. We'll have a lot more to say about this idea.

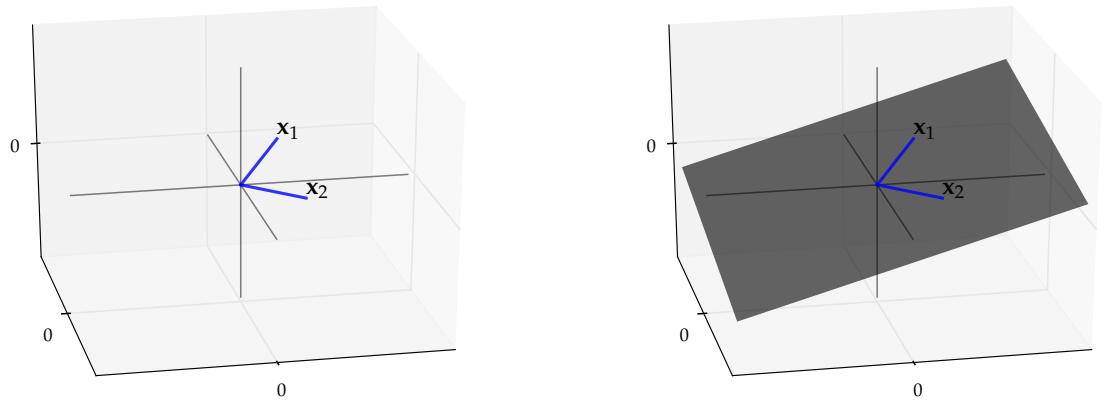


Figure 1.4: Span of $\mathbf{x}_1, \mathbf{x}_2$

Example 1.1.3. Consider the vectors $\{\mathbf{e}_1, \ldots, \mathbf{e}_N\} \subset \mathbb{R}^N$, where $\mathbf{e}_n$ has all zeros except for a 1 as the n-th element:

$$\mathbf{e}_1 := \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \mathbf{e}_2 := \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}, \cdots, \mathbf{e}_N := \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}$$

The case of $\mathbb{R}^2$ is illustrated in figure 1.5. The vectors $\mathbf{e}_1, \ldots, \mathbf{e}_N$ are called the **canonical basis vectors** of $\mathbb{R}^N$—we'll see why later on. One reason is that $\{\mathbf{e}_1, \ldots, \mathbf{e}_N\}$ spans all of $\mathbb{R}^N$. To see this in the case of $N = 2$ (check general N yourself), observe that for any $\mathbf{y} \in \mathbb{R}^2$, we have

$$\mathbf{y} := \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} y_1 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ y_1 \end{pmatrix} = y_1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + y_2 \begin{pmatrix} 0 \\ 1 \end{pmatrix} = y_1 \mathbf{e}_1 + y_2 \mathbf{e}_2$$

Thus, $\mathbf{y} \in \text{span}\{\mathbf{e}_1, \mathbf{e}_2\}$ as claimed. Since $\mathbf{y}$ is just an arbitrary vector in $\mathbb{R}^2$, we have shown that $\{\mathbf{e}_1, \mathbf{e}_2\}$ spans $\mathbb{R}^2$.

Example 1.1.4. Consider the set $P := \{(x_1, x_2, 0) \in \mathbb{R}^3 : x_1, x_2 \in \mathbb{R}\}$. Graphically, P corresponds to the flat plane in $\mathbb{R}^3$, where the height coordinate is always zero. If we take $\mathbf{e}_1 = (1, 0, 0)$ and $\mathbf{e}_2 = (0, 1, 0)$, then given $\mathbf{y} = (y_1, y_2, 0) \in P$ we have $\mathbf{y} = y_1 \mathbf{e}_1 + y_2 \mathbf{e}_2$. In other words, any $\mathbf{y} \in P$ can be expressed as a linear combination of $\mathbf{e}_1$ and $\mathbf{e}_2$, and $\{\mathbf{e}_1, \mathbf{e}_2\}$ is a spanning set for P.

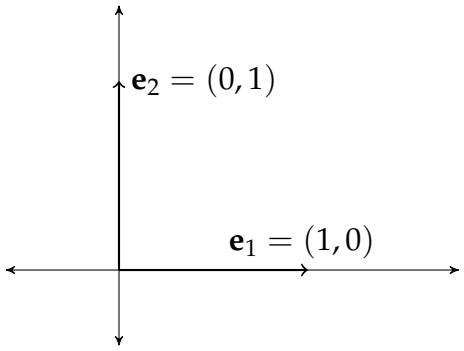


Figure 1.5: Canonical basis vectors in $\mathbb{R}^2$

Fact 1.1.4. Let X and Y be any two finite subsets of $\mathbb{R}^N$. If $X \subset Y$, then we have $\text{span}(X) \subset \text{span}(Y)$.

One of the key features of the span of a set X is that it is "closed" under the linear operations of vector addition and scalar multiplication, in the sense that if we take elements of the span and combine them using these operations, the resulting vectors are still in the span.

To check that the $\text{span}(X)$ is closed under vector addition for arbitrary nonempty $X \subset \mathbb{R}^N$, just observe that if $\mathbf{y}, \mathbf{z} \in \text{span}(X)$, then we can express them as finite linear combinations of elements of X, like so:

$$\mathbf{y} = \sum_{k=1}^{K} \alpha_k \mathbf{x}_k \quad \text{and} \quad \mathbf{z} = \sum_{k=1}^{K'} \alpha'_k \mathbf{x}'_k$$

Here each $\mathbf{x}_k$ and $\mathbf{x}'_k$ is an element of X, and each α_k and α'_k is a scalar. Adding $\mathbf{y}$ and $\mathbf{z}$ gives

$$\mathbf{y} + \mathbf{z} = \sum_{k=1}^{K} \alpha_k \mathbf{x}_k + \sum_{k=1}^{K'} \alpha'_k \mathbf{x}'_k$$

which is yet another finite linear combination of elements of X. Hence $\text{span}(X)$ is closed under vector addition as claimed. Another easy argument shows that $\text{span}(X)$ is closed under scalar multiplication.

The notion of a set being closed under scalar multiplication and vector addition is important enough to have its own name: A nonempty subset S of $\mathbb{R}^N$ is called a **linear subspace** (or just **subspace**) of $\mathbb{R}^N$ if, for any $\mathbf{x}$ and $\mathbf{y}$ in S, and any α and β in $\mathbb{R}$, the linear combination $\alpha\mathbf{x} + \beta\mathbf{y}$ is also in S.

Example 1.1.5. It follows immediately from the proceeding discussion that if X is any nonempty subset of $\mathbb{R}^N$, then $\text{span}(X)$ is a linear subspace of $\mathbb{R}^N$. For this reason, $\text{span}(X)$ is often called the **linear subspace spanned by X.**

In $\mathbb{R}^3$, lines and planes that pass through the origin are linear subspaces. You will get more of a feel for this as we go along. Other linear subspaces of $\mathbb{R}^3$ are the singleton set containing the zero element $\mathbf{0}$, and the set $\mathbb{R}^3$ itself.

Fact 1.1.5. Let S be a linear subspace of $\mathbb{R}^N$. The following statements are true:

1. The origin $\mathbf{0}$ is an element of S.
2. If $X \subset S$, then $\text{span}(X) \subset S$.
3. $\text{span}(S) = S$.

1.1.3 Linear Independence

In some sense, the span of $X := \{\mathbf{x}_1, \ldots, \mathbf{x}_K\}$ is a measure of the "diversity" of the vectors in X—the more diverse are the elements of X, the greater is the set of vectors that can be represented as linear combinations of its elements. In particular, if X is not very "diverse," then some "similar" elements may be redundant, in the sense that one can remove an element $\mathbf{x}_i$ from the collection X without reducing its span.

Let's consider two extremes. First consider the vectors $\mathbf{e}_1 := (1, 0)$ and $\mathbf{e}_2 := (0, 1)$ in $\mathbb{R}^2$ (figure 1.5). As we saw in example 1.1.3, the span $\{\mathbf{e}_1, \mathbf{e}_2\}$ is all of $\mathbb{R}^2$. With just these two vectors, we can span the whole plane. In algebraic terms, these vectors are relatively diverse. We can also see their diversity in the fact that if we remove one of the vectors from $\{\mathbf{e}_1, \mathbf{e}_2\}$, the span is no longer all of $\mathbb{R}^2$. In fact it is just a line in $\mathbb{R}^2$. Hence both vectors have their own role to play in forming the span.

Now consider the pair $\mathbf{e}_1$ and $\mathbf{x} := -2\mathbf{e}_1 = (-2, 0)$, as shown in figure 1.6. This pair is not very diverse. In fact, if $\mathbf{y} \in \text{span}\{\mathbf{e}_1, \mathbf{x}\}$, then, for some α_1 and α_2,

$$\mathbf{y} = \alpha_1\mathbf{e}_1 + \alpha_2\mathbf{x} = \alpha_1\mathbf{e}_1 + \alpha_2(-2)\mathbf{e}_1 = (\alpha_1 - 2\alpha_2)\mathbf{e}_1 \in \text{span}\{\mathbf{e}_1\}$$

In other words, any element of $\text{span}\{\mathbf{e}_1, \mathbf{x}\}$ is also an element of $\text{span}\{\mathbf{e}_1\}$. We can kick $\mathbf{x}$ out of the set $\{\mathbf{e}_1, \mathbf{x}\}$ without reducing the span.

Let's translate these ideas into formal definitions. In general, the set of vectors $X := \{\mathbf{x}_1, \ldots, \mathbf{x}_K\}$ in $\mathbb{R}^N$ is called **linearly dependent** if one (or more) vector(s) can be

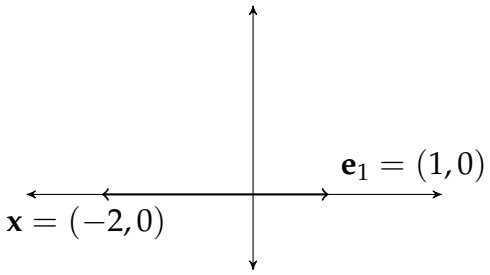


Figure 1.6: The vectors $\mathbf{e}_1$ and $\mathbf{x}$

removed without changing $\text{span}(X)$. We call X **linearly independent** if it is not linearly dependent.

To see the definition of independence in a slightly different light, suppose that $X := \{\mathbf{x}_1, \ldots, \mathbf{x}_K\}$ is linearly dependent, with

$$\text{span}\{\mathbf{x}_1, \ldots, \mathbf{x}_K\} = \text{span}\{\mathbf{x}_2, \ldots, \mathbf{x}_K\}$$

Since $\mathbf{x}_1 \in \text{span}\{\mathbf{x}_1, \ldots, \mathbf{x}_K\}$ certainly holds, this equality implies that

$$\mathbf{x}_1 \in \text{span}\{\mathbf{x}_2, \ldots, \mathbf{x}_K\}$$

Hence, there exist constants $\alpha_2, \ldots, \alpha_K$ with

$$\mathbf{x}_1 = \alpha_2 \mathbf{x}_2 + \cdots \alpha_K \mathbf{x}_K$$

In other words, $\mathbf{x}_1$ can be expressed as a linear combination of the other elements in X. This is a general rule: Linear dependence means that at least one vector in the set can be written as a linear combination of the others. Linear independence means the opposite is true.

There is yet a third way to express linear independence, which is very succinct and often the most useful condition to aim for when checking linear independence. It is given as the first part of the next fact, which clarifies the various relationships.

Fact 1.1.6 (Definitions of linear independence)**.** The following statements are all equivalent definitions of linear independence of $X := \{\mathbf{x}_1, \ldots, \mathbf{x}_K\} \subset \mathbb{R}^N$:

1. If $\alpha_1 \mathbf{x}_1 + \cdots + \alpha_K \mathbf{x}_K = \mathbf{0}$ then $\alpha_1 = \cdots = \alpha_K = 0$.
2. If X_0 is a proper subset of X, then $\text{span}(X_0)$ is a proper subset of $\text{span}(X)$.[1]

[1] A is a proper subset of B if $A \subset B$ and $A \neq B$.

3. No vector in X can be written as a linear combination of the others.

Exercise 1.6.11 asks you to check all of these equivalences.

Perhaps the most important example of linearly independent vectors in $\mathbb{R}^N$ is the canonical basis vectors in example 1.1.3. Indeed, if $\alpha_j \neq 0$ for some j, then $\sum_{k=1}^K \alpha_k \mathbf{e}_k = (\alpha_1, \ldots, \alpha_K) \neq \mathbf{0}$. More examples are given in the exercises.

Fact 1.1.7. If $X := \{\mathbf{x}_1, \ldots, \mathbf{x}_K\}$ is linearly independent, then

1. Every subset of X is linearly independent.
2. X does not contain $\mathbf{0}$.
3. $X \cup \{\mathbf{x}\}$ is linearly independent for all $\mathbf{x} \in \mathbb{R}^N$ such that $\mathbf{x} \notin \operatorname{span}(X)$.

The proof is a solved exercise (exercise 1.6.12 on page 39).

One reason for our interest in the concept of linear independence lies in the following problem: We know when a point in $\mathbb{R}^N$ can be expressed as a linear combination of some fixed set of vectors X. This is true precisely when that point is in the span of X. What we do not know is when that representation is unique. It turns out that the relevant condition is independence:

Theorem 1.1.1. *Let $X := \{\mathbf{x}_1, \ldots, \mathbf{x}_K\}$ be any collection of vectors in $\mathbb{R}^N$. The following statements are equivalent:*

1. *X is linearly independent.*
2. *For each $\mathbf{y} \in \operatorname{span}(X)$ there exists exactly one set of scalars $\alpha_1, \ldots, \alpha_K$ such that*

$$\mathbf{y} = \sum_{k=1}^K \alpha_k \mathbf{x}_k \tag{1.2}$$

Proof. Let X be linearly independent and pick any $\mathbf{y} \in \operatorname{span}(X)$. Since $\mathbf{y}$ is in the span of X, we know that there exists at least one set of scalars such that (1.2) holds. Suppose now that there are two. In particular, suppose that $\mathbf{y} = \sum_{k=1}^K \alpha_k \mathbf{x}_k = \sum_{k=1}^K \beta_k \mathbf{x}_k$. It follows from the second equality that $\sum_{k=1}^K (\alpha_k - \beta_k)\mathbf{x}_k = \mathbf{0}$. Using fact 1.1.6, we conclude that $\alpha_k = \beta_k$ for all k. In other words, the representation is unique.

On the other hand, suppose that for each $\mathbf{y} \in \text{span}(X)$ there exists exactly one set of scalars $\alpha_1, \ldots, \alpha_K$ such that $\mathbf{y} = \sum_{k=1}^K \alpha_k \mathbf{x}_k$. Since $\mathbf{0} \in \text{span}(X)$ must be true (why?), this implies that there exists only one set of scalars such that $\mathbf{0} = \sum_{k=1}^K \alpha_k \mathbf{x}_k$. Since $\alpha_1 = \cdots = \alpha_k = 0$ has this property, we conclude that no nonzero scalars yield $\mathbf{0} = \sum_{k=1}^K \alpha_k \mathbf{x}_k$. In other words, X is linearly independent. □

1.1.4 Bases and Dimension

In essence, the dimension of a linear subspace is the minimum number of vectors needed to span it. For example, consider the plane

$$P := \{(x_1, x_2, 0) \in \mathbb{R}^3 : x_1, x_2 \in \mathbb{R}\} \tag{1.3}$$

from example 1.1.4. Intuitively, this plane is a "two-dimensional" subset of $\mathbb{R}^3$. This intuition agrees with the definition above. Indeed, P cannot be spanned by one vector, for if we take a single vector in $\mathbb{R}^3$, then the span created by that singleton is only a line in $\mathbb{R}^3$, not a plane. On the other hand, P *can* be spanned by two vectors, as we saw in example 1.1.4. Finally, while P can also be spanned by three or more vectors, it turns out that one of the vectors will always be redundant—it does not change the span. In other words, any collection of 3 or more vectors from P will be linearly dependent.

Let's discuss this more formally. We begin with the following deep result:

Theorem 1.1.2. *If S is a linear subspace of $\mathbb{R}^N$ and S is spanned by K vectors, then every linearly independent subset of S has at most K vectors.*

Put differently, if S is spanned by K vectors, then any subset of S with more than K vectors will be linearly dependent. Many practical results lean on this theorem. The proof is not particularly hard, but it is a little long. You can find it in most texts on linear algebra (see, e.g., theorem 1.1 of [21]). Here's an example of what we can do with theorem 1.1.2.

Theorem 1.1.3. *Let $X := \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ be any N vectors in $\mathbb{R}^N$. The following statements are equivalent:*

1. $\text{span}(X) = \mathbb{R}^N$.
2. *X is linearly independent.*

Proof. Suppose first that X is linearly independent. Seeking a contradiction, suppose in addition that there exists an $\mathbf{x} \in \mathbb{R}^N$ with $\mathbf{x} \notin \text{span}(X)$. By fact 1.1.7 it then follows that the $N+1$ element set $X \cup \{\mathbf{x}\} \subset \mathbb{R}^N$ is linearly independent. Since $\mathbb{R}^N$ is spanned by the N canonical basis vectors, this statement stands in contradiction to theorem 1.1.2.

Conversely, suppose that $\text{span}(X) = \mathbb{R}^N$ but X is not linearly independent. Then, by fact 1.1.6, there exists a proper subset X_0 of X with $\text{span}(X_0) = \text{span}(X)$. Since X_0 is a proper subset of X it contains $K < N$ elements. We now have K vectors spanning $\mathbb{R}^N$. In particular, the span of K vectors contains the N linearly independent vectors $\mathbf{e}_1, \ldots, \mathbf{e}_N$. Once again, this statement stands in contradiction to theorem 1.1.2. □

We now come to a key definition. If S is a linear subspace of $\mathbb{R}^N$ and B is some finite subset of $\mathbb{R}^N$, then B is called a **basis** of S if B spans S and is linearly independent.

Example 1.1.6. The pair $\{\mathbf{e}_1, \mathbf{e}_2\}$ is a basis for the set P defined in (1.3).

Example 1.1.7. The set of canonical basis vectors $\{\mathbf{e}_1, \ldots, \mathbf{e}_N\} \subset \mathbb{R}^N$ described in example 1.1.3 is linearly independent and its span is equal to all of $\mathbb{R}^N$ (see page 7). As a result, $\{\mathbf{e}_1, \ldots, \mathbf{e}_N\}$ is a basis for $\mathbb{R}^N$—as anticipated by the name.

As stated above, the dimension of a linear subspace is the minimum number of vectors needed to span it. To formalize this idea, we will use the following fundamental result about bases:

Theorem 1.1.4. *If S is a linear subspace of $\mathbb{R}^N$ distinct from $\{\mathbf{0}\}$, then*

1. *S has at least one basis, and*
2. *every basis of S has the same number of elements.*

The proof of part 1 can be found in any good text on linear algebra. See, for example, theorem 1.1 of [21]. Part 2 follows from theorem 1.1.2 and is left as an exercise (exercise 1.6.13).

If S is a linear subspace of $\mathbb{R}^N$, then common number identified in theorem 1.1.4 is called the **dimension** of S, and written as $\dim(S)$. For example,

1. $\dim(P) = 2$ for the plane in (1.3), because $\{\mathbf{e}_1, \mathbf{e}_2\} \subset \mathbb{R}^3$ is a basis.
2. $\dim(\mathbb{R}^N) = N$, because $\{\mathbf{e}_1, \ldots, \mathbf{e}_N\} \subset \mathbb{R}^N$ is a basis.

In $\mathbb{R}^N$ the singleton subspace $\{\mathbf{0}\}$ is said to have zero dimension. A line $\{\alpha\mathbf{x} \in \mathbb{R}^N : \alpha \in \mathbb{R}\}$ through the origin is obviously one dimensional.

If we take a set of K vectors, then how large will its span be in terms of dimension? The next lemma answers this question.

Lemma 1.1.1. *Let $X := \{\mathbf{x}_1, \ldots, \mathbf{x}_K\} \subset \mathbb{R}^N$. Then*

1. $\dim(\text{span}(X)) \leq K$.
2. $\dim(\text{span}(X)) = K$ *if and only if X is linearly independent.*

Proof. Regarding part 1, let B be a basis of $\text{span}(X)$. By definition, B is a linearly independent subset of $\text{span}(X)$. Since $\text{span}(X)$ is spanned by K vectors, theorem 1.1.2 implies that B has no more than K elements. Hence, $\dim(\text{span}(X)) \leq K$.

Regarding part 2, suppose first that X is linearly independent. Then X is a basis for $\text{span}(X)$. Since X has K elements, we conclude that $\dim(\text{span}(X)) = K$. Conversely, if $\dim(\text{span}(X)) = K$ then X must be linearly independent. For if X is not linearly independent, then exists a proper subset X_0 of X such that $\text{span}(X_0) = \text{span}(X)$. By part 1 of this theorem, we then have $\dim(\text{span}(X_0)) \leq \#X_0 \leq K - 1$. Therefore, $\dim(\text{span}(X)) \leq K - 1$. Contradiction. □

Part 2 of lemma 1.1.1 is important in what follows, and also rather intuitive. It says that the span of a set will be large when it's elements are algebraically diverse.

Let's finish this section with facts that can be deduced from the preceding results.

Fact 1.1.8. The following statements are true:

1. Let S and T be K-dimensional linear subspaces of $\mathbb{R}^N$. If $S \subset T$, then $S = T$.
2. If S is an M-dimensional linear subspace of $\mathbb{R}^N$ where $M < N$, then $S \neq \mathbb{R}^N$.

The first part of fact 1.1.8 has many useful implications, one of which is that the only N-dimensional linear subspace of $\mathbb{R}^N$ is $\mathbb{R}^N$.

1.2 Linear Maps

[Roadmap]

1.2.1 Linearity

There are many different classes of functions and perhaps the most important of these is the class of linear functions. In high school one learns about linear functions as those whose graph is a straight line. We need a more formal (and accurate) definition: A function $T\colon \mathbb{R}^K \to \mathbb{R}^N$ is called **linear** if

$$T(\alpha\mathbf{x} + \beta\mathbf{y}) = \alpha T\mathbf{x} + \beta T\mathbf{y} \text{ for any } \mathbf{x}, \mathbf{y} \in \mathbb{R}^K \text{ and } \alpha, \beta \in \mathbb{R}$$

Remark: Mathematicians usually write linear functions with upper case letters and omit the parenthesis around the argument where no confusion arises. It is a custom of their tribe.

Example 1.2.1. The function $T\colon \mathbb{R} \to \mathbb{R}$ defined by $Tx = 2x$ is linear because for any α, β, x, y in $\mathbb{R}$, we have

$$T(\alpha x + \beta y) = 2(\alpha x + \beta y) = \alpha 2x + \beta 2y = \alpha Tx + \beta Ty$$

Example 1.2.2. Given constants $c_1, \ldots, c_K$, the function $T\colon \mathbb{R}^K \to \mathbb{R}$ defined by

$$T\mathbf{x} = T(x_1, \ldots, x_K) = \sum_{k=1}^{K} c_k x_k$$

is linear, because if we take any α, β in $\mathbb{R}$ and $\mathbf{x}, \mathbf{y}$ in $\mathbb{R}^K$, then

$$T(\alpha\mathbf{x} + \beta\mathbf{y}) = \sum_{k=1}^{K} c_k[\alpha x_k + \beta y_k] = \alpha \sum_{k=1}^{K} c_k x_k + \beta \sum_{k=1}^{K} c_k y_k = \alpha T\mathbf{x} + \beta T\mathbf{y}$$

Example 1.2.3. The function $f\colon \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x^2$ is *non*linear, because if we take $\alpha = \beta = x = y = 1$, then

$$f(\alpha x + \beta y) = f(2) = 4 \quad \text{while} \quad \alpha f(x) + \beta f(y) = 1 + 1 = 2$$

Thinking of linear functions as those whose graph is a straight line is incorrect. For example, the function $f\colon \mathbb{R} \to \mathbb{R}$ defined by $f(x) = 1 + 2x$ is *non*linear, because if we take $\alpha = \beta = x = y = 1$, then $f(\alpha x + \beta y) = f(2) = 5$, while $\alpha f(x) + \beta f(y) = 3 + 3 = 6$. This kind of function is actually called an **affine** function.

The range of a linear map is the span of the image of the canonical basis functions:

Lemma 1.2.1. *If $T\colon \mathbb{R}^K \to \mathbb{R}^N$ is a linear map, then*

$$\operatorname{rng}(T) = \operatorname{span}(V) \quad \textit{where} \quad V := \{T\mathbf{e}_1, \ldots, T\mathbf{e}_K\}$$

Proof. Any $\mathbf{x} \in \mathbb{R}^K$ can be expressed in terms of the basis vectors as $\sum_{k=1}^{K} \alpha_k \mathbf{e}_k$, for some suitable choice of scalars. Hence $\text{rng}(T)$ is the set of all points of the form

$$T\mathbf{x} = T\left[\sum_{k=1}^{K} \alpha_k \mathbf{e}_k\right] = \sum_{k=1}^{K} \alpha_k T\mathbf{e}_k$$

as we vary $\alpha_1, \ldots, \alpha_K$ over all scalar combinations. This coincides with the definition of $\text{span}(V)$. □

For linear map $T\colon \mathbb{R}^K \to \mathbb{R}^N$, the **null space** or **kernel** of T is defined as

$$\ker(T) := \{\mathbf{x} \in \mathbb{R}^K : T\mathbf{x} = \mathbf{0}\}$$

Fact 1.2.1. If $T\colon \mathbb{R}^K \to \mathbb{R}^N$ is linear, then

1. $\ker(T)$ is a linear subspace of $\mathbb{R}^K$.
2. $\ker(T) = \{\mathbf{0}\}$ if and only if T is one-to-one.

The proofs are straightforward. For example, suppose that and $T\mathbf{x} = T\mathbf{y}$ for arbitrary $\mathbf{x}, \mathbf{y} \in \mathbb{R}^K$. Then $\mathbf{x} - \mathbf{y} \in \ker(T)$. If $\ker(T) = \{\mathbf{0}\}$ then it must be that $\mathbf{x} = \mathbf{y}$. Since $\mathbf{x}$ and $\mathbf{y}$ were arbitrary we conclude that T is one-to-one.

1.2.2 Linear Maps from $\mathbb{R}^N$ to $\mathbb{R}^N$

Linear functions from $\mathbb{R}^N$ to itself (linear self-mappings) have an extremely neat and useful property: They are onto if and only if they are one-to-one. In other words, for linear self-mappings (maps from a space back into itself), the properties of being one-to-one, onto and a bijection are all entirely equivalent. The next theorem gives details.

Theorem 1.2.1. *If T is a linear function from $\mathbb{R}^N$ to $\mathbb{R}^N$ then all of the following are equivalent:*

1. *T is a bijection.*
2. *T is onto.*
3. *T is one-to-one.*

4. $\ker(T) = \{\mathbf{0}\}$.

5. *The set of vectors* $V := \{T\mathbf{e}_1, \ldots, T\mathbf{e}_N\}$ *is linearly independent.*

If any one of these equivalent conditions is true, then T is called **nonsingular**. (In other words, a function from $\mathbb{R}^N$ to itself is called a nonsingular function if it is a linear bijection.) Otherwise T is called **singular**. The proof of theorem 1.2.1 is not overly difficult given all the work we've already done. It is left as a (solved) exercise (exercise 1.6.15).

If T is nonsingular, then, being a bijection, it must have an inverse function T^{-1} that is also a bijection (fact 4.2.1 on page 149). It turns out that this inverse function inherits the linearity of T. In summary,

Fact 1.2.2. If $T\colon \mathbb{R}^N \to \mathbb{R}^N$ is nonsingular then so is T^{-1}.

1.2.3 Linear Maps Across Dimensions

Theorem 1.2.1 only applies to linear maps between spaces of the *same* dimension. Here are fundamental results for the other two cases (smaller to bigger and bigger to smaller).

Theorem 1.2.2. *For a linear map* T *from* $\mathbb{R}^K \to \mathbb{R}^N$*, the following statements are true:*

1. *If* $K < N$ *then* T *is not onto.*

2. *If* $K > N$ *then* T *is not one-to-one.*

The most important implication: If $N \neq K$, then we can forget about bijections.

Proof. Regarding part 1, let $K < N$ and let $T\colon \mathbb{R}^K \to \mathbb{R}^N$ be linear. T cannot be onto, for if T were onto then we would have $\text{rng}(T) = \mathbb{R}^N$, in which case the vectors in $V = \{T\mathbf{e}_1, \ldots, T\mathbf{e}_K\}$ in lemma 1.2.1 would span $\mathbb{R}^N$, despite having only $K < N$ elements. This is impossible. (Why?)

The proof of part 2 is left as an exercise (exercise 1.6.16). □

1.3 Matrices and Linear Equations

[Roadmap]

1.3.1 Basic Definitions

A $N \times K$ **matrix** is a rectangular array $\mathbf{A}$ of real numbers with N rows and K columns, written in the following way:

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1K} \\ a_{21} & a_{22} & \cdots & a_{2K} \\ \vdots & \vdots & & \vdots \\ a_{N1} & a_{N2} & \cdots & a_{NK} \end{pmatrix}$$

Often, the values a_{nk} in the matrix represent coefficients in a system of linear equations, such as

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1K}x_K &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2K}x_K &= b_2 \\ &\vdots \\ a_{N1}x_1 + a_{N2}x_2 + \cdots + a_{NK}x_K &= b_N \end{aligned} \tag{1.4}$$

We'll explore this relationship further after some more definitions.

In matrix $\mathbf{A}$, the symbol a_{nk} stands for the element in the n-th row of the k-th column. For obvious reasons, the matrix $\mathbf{A}$ is also called a vector if either $N = 1$ or $K = 1$. In the former case, $\mathbf{A}$ is called a **row vector**, while in the latter case it is called a **column vector**. If $\mathbf{A}$ is $N \times K$ and $N = K$, then $\mathbf{A}$ is called **square**. When convenient, we will use the notation $\text{row}_n(\mathbf{A})$ to refer to the n-th row of $\mathbf{A}$, and $\text{col}_k(\mathbf{A})$ to refer to it's k-th column.

The square matrix $\mathbf{I}$ where n,k-th element is 1 if $n = k$ and zero otherwise is called the **identity matrix**, and denoted by $\mathbf{I}$:

$$\mathbf{I} := \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

Just as was the case for vectors, a number of algebraic operations are defined for matrices. The first two, scalar multiplication and addition, are immediate generalizations of the vector case: For $\gamma \in \mathbb{R}$, we let

$$\gamma \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1K} \\ a_{21} & a_{22} & \cdots & a_{2K} \\ \vdots & \vdots & & \vdots \\ a_{N1} & a_{N2} & \cdots & a_{NK} \end{pmatrix} := \begin{pmatrix} \gamma a_{11} & \gamma a_{12} & \cdots & \gamma a_{1K} \\ \gamma a_{21} & \gamma a_{22} & \cdots & \gamma a_{2K} \\ \vdots & \vdots & & \vdots \\ \gamma a_{N1} & \gamma a_{N2} & \cdots & \gamma a_{NK} \end{pmatrix}$$

while

$$\begin{pmatrix} a_{11} & \cdots & a_{1K} \\ a_{21} & \cdots & a_{2K} \\ \vdots & \vdots & \vdots \\ a_{N1} & \cdots & a_{NK} \end{pmatrix} + \begin{pmatrix} b_{11} & \cdots & b_{1K} \\ b_{21} & \cdots & b_{2K} \\ \vdots & \vdots & \vdots \\ b_{N1} & \cdots & b_{NK} \end{pmatrix} := \begin{pmatrix} a_{11} + b_{11} & \cdots & a_{1K} + b_{1K} \\ a_{21} + b_{21} & \cdots & a_{2K} + b_{2K} \\ \vdots & \vdots & \vdots \\ a_{N1} + b_{N1} & \cdots & a_{NK} + b_{NK} \end{pmatrix}$$

In the latter case, the matrices have to have the same number of rows and columns in order for the definition to make sense.

Now let's look at multiplication of matrices. If $\mathbf{A}$ and $\mathbf{B}$ are two matrices, then their product $\mathbf{C} := \mathbf{AB}$ is formed by taking as it's i, j-th element the inner product of the i-th row of $\mathbf{A}$ and the j-th column of $\mathbf{B}$. That is,

$$c_{ij} = \text{row}_i(\mathbf{A})' \, \text{col}_j(\mathbf{B}) = \sum_{k=1}^{K} a_{ik} b_{kj}$$

Here's the picture for $i = j = 1$:

$$\begin{pmatrix} a_{11} & \cdots & a_{1K} \\ a_{21} & \cdots & a_{2K} \\ \vdots & \vdots & \vdots \\ a_{N1} & \cdots & a_{NK} \end{pmatrix} \begin{pmatrix} b_{11} & \cdots & b_{1J} \\ b_{21} & \cdots & b_{2J} \\ \vdots & \vdots & \vdots \\ b_{K1} & \cdots & b_{KJ} \end{pmatrix} = \begin{pmatrix} c_{11} & \cdots & c_{1J} \\ c_{21} & \cdots & c_{2J} \\ \vdots & \vdots & \vdots \\ c_{N1} & \cdots & c_{NJ} \end{pmatrix}$$

Since inner products are only defined for vectors of equal length, this requires that the length of the rows of $\mathbf{A}$ is equal to the length of the columns of $\mathbf{B}$. In other words, if $\mathbf{A}$ is $N \times K$ and $\mathbf{B}$ is $J \times M$, then we require $K = J$. The resulting matrix $\mathbf{AB}$ is $N \times M$. Here's the rule to remember:

$$\text{product of } N \times K \text{ and } K \times M \text{ is } N \times M$$

From the definition, it is clear that multiplication is not commutative, in that $\mathbf{AB}$ and $\mathbf{BA}$ are not generally the same thing. Indeed $\mathbf{BA}$ is not well-defined unless $N = M$ also holds. Even in this case, the two are not generally equal. Other than that, multiplication behaves pretty much as we'd expect:

Fact 1.3.1. For conformable matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ and scalar α we have

1. $\mathbf{A}(\mathbf{BC}) = (\mathbf{AB})\mathbf{C}$
2. $\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$

3. $(\mathbf{A}+\mathbf{B})\mathbf{C} = \mathbf{AC}+\mathbf{BC}$
4. $\mathbf{A}\alpha\mathbf{B} = \alpha\mathbf{AB}$

Here, we are using the word "conformable" to indicate dimensions are such that the operation in question makes sense. For example, we'll say "for two conformable matrices $\mathbf{A}$ and $\mathbf{B}$, the product $\mathbf{AB}$ satisfies xyz" if the dimensions of $\mathbf{A}$ and $\mathbf{B}$ are such that the product is well defined; and similarly for addition, etc.

You can verify that, assuming conformability, we always have

$$\mathbf{AI} = \mathbf{IA} = \mathbf{A}$$

For this reason the identity matrix is sometimes called the "multiplicative unit".

1.3.2 Matrices as Maps

The single most useful way to think of a matrix is as a mapping over vector space. In particular, an $N \times K$ matrix $\mathbf{A}$ can be thought of as a map sending a vector $\mathbf{x} \in \mathbb{R}^K$ into a new vector $\mathbf{y} = \mathbf{Ax}$ in $\mathbb{R}^N$. Among the collection of all functions from $\mathbb{R}^K$ to $\mathbb{R}^N$, these functions defined by matrices have a special property: they are all linear.

To see that this is so, take a fixed $N \times K$ matrix $\mathbf{A}$ and consider the function $T\colon \mathbb{R}^K \to \mathbb{R}^N$ defined by $T\mathbf{x} = \mathbf{Ax}$. To see that T is a linear function, pick any $\mathbf{x}$, $\mathbf{y}$ in $\mathbb{R}^K$, and any scalars α and β. The rules of matrix arithmetic (see fact 1.3.1) tell us that

$$T(\alpha\mathbf{x}+\beta\mathbf{y}) := \mathbf{A}(\alpha\mathbf{x}+\beta\mathbf{y}) = \mathbf{A}\alpha\mathbf{x}+\mathbf{A}\beta\mathbf{y} = \alpha\mathbf{Ax}+\beta\mathbf{Ay} =: \alpha T\mathbf{x}+\beta T\mathbf{y}$$

In other words, T is linear.

So matrices make linear functions. How about some examples of a linear functions that don't involve matrices? Actually we can't give any such examples because there are none:

Theorem 1.3.1. *Let T be a function from $\mathbb{R}^K$ to $\mathbb{R}^N$. The following are equivalent:*

1. *T is linear.*
2. *There exists an $N \times K$ matrix $\mathbf{A}$ such that $T\mathbf{x} = \mathbf{Ax}$ for all $\mathbf{x} \in \mathbb{R}^K$.*

In other words, the set of linear functions from $\mathbb{R}^K$ to $\mathbb{R}^N$ and the set of $N \times K$ matrices are in one-to-one correspondence.

Proof. We've already proved one implication. Regarding the second, let $T\colon \mathbb{R}^K \to \mathbb{R}^N$ be linear. We aim to construct a matrix $\mathbf{A}$ such that $T\mathbf{x} = \mathbf{A}\mathbf{x}$ for all $\mathbf{x} \in \mathbb{R}^K$. As usual, let $\mathbf{e}_k$ be the k-th canonical basis vector in $\mathbb{R}^K$. Define an $N \times K$ matrix $\mathbf{A}$ by $\operatorname{col}_k(\mathbf{A}) = T\mathbf{e}_k$. Pick any $\mathbf{x} = (x_1, \ldots, x_K) \in \mathbb{R}^K$. We can also write $\mathbf{x} = \sum_{k=1}^K x_k \mathbf{e}_k$. By linearity we have

$$T\mathbf{x} = \sum_{k=1}^K x_k T\mathbf{e}_k = \sum_{k=1}^K x_k \operatorname{col}_k(\mathbf{A})$$

This is just $\mathbf{A}\mathbf{x}$, and we are done. □

Let $\mathbf{A}$ be an $N \times K$ matrix and consider the corresponding linear map T defined by $T\mathbf{x} = \mathbf{A}\mathbf{x}$. The range of this map is typically write $\operatorname{span}(\mathbf{A})$ instead of $\operatorname{rng}(T)$. That is,

$$\operatorname{span}(\mathbf{A}) := \{\mathbf{A}\mathbf{x} : \mathbf{x} \in \mathbb{R}^K\}$$

The reason for this notation is that the range of T, or the set of all $\mathbf{A}\mathbf{x}$ where $\mathbf{x} \in \mathbb{R}^K$, is precisely the span of $\mathbf{A}$'s columns:

$$\operatorname{span}(\mathbf{A}) = \operatorname{span}\{\operatorname{col}_1(\mathbf{A}), \ldots, \operatorname{col}_K(\mathbf{A})\} \tag{1.5}$$

(This is an important point. Please work through it using the definition of matrix multiplication if you are not fully convinced that it's true.) For obvious reasons, this set is also called the **column space** of $\mathbf{A}$. It is a linear subspace of $\mathbb{R}^N$. (Why?)

How large is the column space of a given matrix? To answer that question we have to say what "large" means. In the context of linear algebra, size of subspaces is usually measured by dimension. The dimension of $\operatorname{span}(\mathbf{A})$ is known as the **rank** of $\mathbf{A}$. That is,

$$\operatorname{rank}(\mathbf{A}) := \dim(\operatorname{span}(\mathbf{A}))$$

$\mathbf{A}$ is said to have **full column rank** if $\operatorname{rank}(\mathbf{A})$ is equal to K, the number of its columns. The reason we say "full" rank here is that, by definition, $\operatorname{span}(\mathbf{A})$ is the span of K vectors. Hence, by part 1 of lemma 1.1.1 on page 14, we must have $\dim(\operatorname{span}(\mathbf{A})) \leq K$. In other words, the rank of $\mathbf{A}$ is less than or equal to K. $\mathbf{A}$ is said to have full column rank when this maximum is achieved.

When is this maximum achieved? By part 2 of lemma 1.1.1, this while be the case precisely when the columns of $\mathbf{A}$ are linearly independent. Let's state this as a fact:

Fact 1.3.2. Let $\mathbf{A}$ be an $N \times K$ matrix. The following statements are equivalent:

1. $\mathbf{A}$ is of full column rank.

2. The columns of $\mathbf{A}$ are linearly independent.

3. If $\mathbf{Ax} = \mathbf{0}$, then $\mathbf{x} = \mathbf{0}$.

The last equivalence follows from fact 1.1.6 on page 10.

1.3.3 Square Matrices and Invertibility

The underlying problem driving much of this discussion is the need to solve systems of equations of the form (1.4), which can be written more conveniently in matrix notation as $\mathbf{Ax} = \mathbf{b}$. Ideally we want general conditions on $\mathbf{A}$ such that a solution to this equation will always exist, plus a way to compute the solution.

There are a variety of scenarios depending on the properties of $\mathbf{A}$. To narrow things down, let's concentrate for now on perhaps the most important case, where $\mathbf{A}$ is a square $N \times N$ matrix. Let's also aim for the best case: A set of conditions on $\mathbf{A}$ such that, for every single $\mathbf{b} \in \mathbb{R}^N$, there exists exactly one $\mathbf{x} \in \mathbb{R}^N$ such that $\mathbf{Ax} = \mathbf{b}$.

The best way to understand this problem is to recall the discussion in §1.2.1, where we started to think about matrices as maps. Letting T be the linear map $T\mathbf{x} = \mathbf{Ax}$, the question we are asking here is when does each point in $\mathbb{R}^N$ have one and only one preimage under T. In other words, *when is T a bijection*?

Recall from §1.2.2 that linear bijections have a special name—they are called non-singular functions. Moreover, theorem 1.2.1 on page 1.2.1 gives us all sorts of nice equivalences for this property. For example, it's enough to know that $T\mathbf{x} = \mathbf{Ax}$ is one-to-one, or onto, or that the image of the set of canonical basis vectors is linearly independent. The next theorem simply replicates these equivalences in the language of matrices.

Theorem 1.3.2. *For $N \times N$ matrix $\mathbf{A}$, the following are equivalent:*

1. *The columns of $\mathbf{A}$ are linearly independent.*

2. $\text{rank}(\mathbf{A}) = N$.

3. $\text{span}(\mathbf{A}) = \mathbb{R}^N$.

4. *If $\mathbf{Ax} = \mathbf{Ay}$, then $\mathbf{x} = \mathbf{y}$.*

5. *If $\mathbf{Ax} = \mathbf{0}$, then $\mathbf{x} = \mathbf{0}$.*

6. *For each* $\mathbf{b} \in \mathbb{R}^N$*, the equation* $\mathbf{Ax} = \mathbf{b}$ *has a solution.*

7. *For each* $\mathbf{b} \in \mathbb{R}^N$*, the equation* $\mathbf{Ax} = \mathbf{b}$ *has a unique solution.*

Obtaining a proof of theorem 1.3.2 is just a matter of going back to theorem 1.2.1 on page 16 and checking the implications for $T\mathbf{x} = \mathbf{Ax}$. For example, with a bit of algebra you should be able to convince yourself that linear independence of $\{T\mathbf{e}_1, \ldots, T\mathbf{e}_N\}$ is equivalent to linear independence of the columns of $\mathbf{A}$.

Following common usage, if any of the equivalent conditions in theorem 1.3.2 are true we will call not just the map T but also the underlying matrix $\mathbf{A}$ **nonsingular**. If any one and hence all of these conditions fail, then $\mathbf{A}$ is called **singular**. A nonsingular matrix sometimes also referred to as **invertible**.

Theorem 1.3.3. *For nonsingular* $\mathbf{A}$ *the following statements are true:*

1. *There exists a square matrix* $\mathbf{B}$ *such that* $\mathbf{AB} = \mathbf{BA} = \mathbf{I}$*, where where* $\mathbf{I}$ *is the identity matrix. The matrix* $\mathbf{B}$ *is called the* **inverse** *of* $\mathbf{A}$*, and written as* $\mathbf{A}^{-1}$*.*

2. *For each* $\mathbf{b} \in \mathbb{R}^N$*, the unique solution to* $\mathbf{Ax} = \mathbf{b}$ *is given by*

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b} \tag{1.6}$$

Proof. Let T be the linear map associated with $\mathbf{A}$ via $T\mathbf{x} = \mathbf{Ax}$. Since A is nonsingular, T is also, by definition, nonsingular, and hence, by fact 1.2.2 on page 17, has a nonsingular inverse T^{-1}. Being nonsingular, T^{-1} is necessarily linear, and hence, by theorem 1.3.1 on page 20, there exists a matrix B such that $T^{-1}\mathbf{x} = \mathbf{Bx}$ for all $\mathbf{x}$. By the definition of the inverse we have

$$\mathbf{ABx} = T(T^{-1}(\mathbf{x})) = \mathbf{x} = \mathbf{Ix}$$

Since this holds for any $\mathbf{x}$ we have $\mathbf{AB} = \mathbf{I}$ (see exercise 1.6.27). A similar argument shows that $\mathbf{BA} = \mathbf{I}$.

Regarding the second claim, $\mathbf{A}^{-1}\mathbf{b}$ is a solution to $\mathbf{Ax} = \mathbf{b}$, since $\mathbf{AA}^{-1}\mathbf{b} = \mathbf{Ib} = \mathbf{b}$. Uniqueness follows from theorem 1.3.2 (and in particular the fact that nonsingularity of $\mathbf{A}$ includes the implication that the map $\mathbf{x} \mapsto \mathbf{Ax}$ is one-to-one). □

The next handy fact tells us that to to prove that $\mathbf{B}$ is an inverse of $\mathbf{A}$, it suffices to show either that $\mathbf{B}$ is a "left inverse" or that $\mathbf{B}$ is a "right inverse". A short proof is given in §1.3.4.

Fact 1.3.3. Let $\mathbf{A}$ be an $N \times N$ matrix. If there exists a $N \times N$ matrix $\mathbf{B}$ such that either $\mathbf{AB} = \mathbf{I}$ or $\mathbf{BA} = \mathbf{I}$, then $\mathbf{A}$ is nonsingular and $\mathbf{A}^{-1} = \mathbf{B}$.

The next fact collects more useful results about inverse matrices.

Fact 1.3.4. If $\mathbf{A}$ and $\mathbf{B}$ are nonsingular and $\alpha \neq 0$, then

1. $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$,
2. $(\alpha\mathbf{A})^{-1} = \alpha^{-1}\mathbf{A}^{-1}$, and
3. $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$.

The relation $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$ is just a special case of the analogous rule for inversion of bijections—see fact 4.2.1 on page 149.

1.3.4 Determinants

To each square matrix $\mathbf{A}$, we can associate a unique number $\det(\mathbf{A})$ called the determinant of $\mathbf{A}$. The determinant is a bit fiddly to describe, but it turns out to give a neat one-number summary of certain useful properties.

To begin, let $S(N)$ be the set of all bijections from $\{1, \ldots, N\}$ to itself, also called the set of **permutations** on $\{1, \ldots, N\}$. For $\pi \in S(N)$ we define the **signature** of π as

$$\operatorname{sgn}(\pi) := \prod_{m<n} \frac{\pi(m) - \pi(n)}{m - n}$$

The **determinant** of $N \times N$ matrix $\mathbf{A}$ is then given as

$$\det(\mathbf{A}) := \sum_{\pi \in S(N)} \operatorname{sgn}(\pi) \prod_{n=1}^{N} a_{\pi(n)n}$$

In the 2×2 case it can be show that this definition reduces to

$$\det \begin{pmatrix} a & b \\ c & d \end{pmatrix} = ad - bc \tag{1.7}$$

Solving for determinants of most larger matrices is a fiddly task, best left for computers. However, the determinant has many neat properties that can be used for proving various results.

Fact 1.3.5. If $\mathbf{I}$ is the $N \times N$ identity, $\mathbf{A}$ and $\mathbf{B}$ are $N \times N$ matrices and $\alpha \in \mathbb{R}$, then

1. $\det(\mathbf{I}) = 1$
2. $\mathbf{A}$ is nonsingular if and only if $\det(\mathbf{A}) \neq 0$
3. $\det(\mathbf{AB}) = \det(\mathbf{A}) \det(\mathbf{B})$
4. $\det(\alpha \mathbf{A}) = \alpha^N \det(\mathbf{A})$
5. $\det(\mathbf{A}^{-1}) = (\det(\mathbf{A}))^{-1}$,

As an example of how these facts can be useful, let's establish fact 1.3.3. Fix square matrix $\mathbf{A}$ and suppose that a right inverse $\mathbf{B}$ exists, in the sense that $\mathbf{AB} = \mathbf{I}$. This equality immediately implies that both $\mathbf{A}$ and $\mathbf{B}$ are nonsingular. Indeed, if we apply det to both sides of $\mathbf{AB} = \mathbf{I}$ and use the rules in fact 1.3.5 we get $\det(\mathbf{A}) \det(\mathbf{B}) = 1$. It follows that both $\det(\mathbf{A})$ and $\det(\mathbf{B})$ are nonzero, and hence both matrices are nonsingular.

The rest is just a mopping up operation. To show that $\mathbf{B}$ is the inverse of $\mathbf{A}$, we just need to check that, in addition to $\mathbf{AB} = \mathbf{I}$ we also have $\mathbf{BA} = \mathbf{I}$. To obtain the latter equality from the former, premultiply the former by $\mathbf{B}$ to get $\mathbf{BAB} = \mathbf{B}$, and then postmultiply by $\mathbf{B}^{-1}$ to get $\mathbf{BA} = \mathbf{I}$. The proof for the left inverse case is similar.

1.3.5 Solving Equations with Tall Matrices

In §1.3.3 we talked about solving equations of the form $\mathbf{Ax} = \mathbf{b}$ when $\mathbf{A}$ is square. Now let's turn to the case where $\mathbf{A}$ is $N \times K$ and $K < N$. Such a system of equations is said to be **overdetermined**. This corresponds to the situation where the number of equations (equal to N) is larger than the number of unknowns (the K elements of $\mathbf{x}$). Intuitively, in such a situation, we may not be able find a $\mathbf{x}$ that satisfies all N equations.

To repeat, we seek a solution $\mathbf{x} \in \mathbb{R}^K$ to $\mathbf{Ax} = \mathbf{b}$ with $K < N$ and $\mathbf{b} \in \mathbb{R}^N$ given. As discussed in §1.3.2, $\mathbf{A}$ can be viewed as a mapping from $\mathbb{R}^K$ to $\mathbb{R}^N$. A solution to $\mathbf{Ax} = \mathbf{b}$ exists precisely when $\mathbf{b}$ lies in the range of this map. As discussed on page 21, its range is the linear subspace of $\mathbb{R}^N$ spanned by the columns of $\mathbf{A}$:

$$\text{span}(\mathbf{A}) := \{\text{all vectors } \mathbf{Ax} \text{ with } \mathbf{x} \in \mathbb{R}^K\} =: \text{column space of } \mathbf{A}$$

Given our assumption that $K < N$, the scenario $\mathbf{b} \in \text{span}(\mathbf{A})$ is very rare. The reason is that the dimension of $\text{span}(\mathbf{A})$, which is precisely the rank of $\mathbf{A}$, is less than or equal to K (see §1.3.2) and hence strictly less than N. We known from fact 1.1.8 on page 14 that such a space is not equal to $\mathbb{R}^N$, where $\mathbf{b}$ lives. In fact we can say more: All K-dimensional subspaces of $\mathbb{R}^N$ are "negligible," and hence the "chance" of $\mathbf{b}$ happening to lie in this subspace is likewise small. For example, consider the case where $N = 3$ and $K = 2$. Then the column space of $\mathbf{A}$ forms at most a 2 dimensional plane in $\mathbb{R}^3$. Intuitively, this set has *no volume* because planes have no "thickness," and hence the chance of a randomly chosen $\mathbf{b}$ lying in this plane is essentially zero.[2]

As a result, the standard approach is to admit that an exact solution may not exist, and instead focus on finding a $\mathbf{x} \in \mathbb{R}^K$ such that $\mathbf{Ax}$ is as close to $\mathbf{b}$ as possible. This problem is taken up in §3.3.2, after we have developed tools sufficient to tackle it.

1.3.6 Solving Equations with Wide Matrices

Now let's turn to the case where $\mathbf{A}$ is $N \times K$ and $K > N$. In this setting the system of equations $\mathbf{Ax} = \mathbf{b}$ is said to be **underdetermined**, meaning that the number of equations (equal to N) is strictly smaller than the number of unknowns (the K elements of $\mathbf{x}$). Intuitively, in such a situation, we may not have enough information to pin down a unique solution $\mathbf{x}$. Indeed, by theorem 1.2.2 on page 17, the map $\mathbf{x} \mapsto \mathbf{Ax}$ *cannot* be one-to-one in this setting. In fact the following is true.

Fact 1.3.6. If $\mathbf{A}$ is $N \times K$, the equation $\mathbf{Ax} = \mathbf{b}$ has a solution and $K > N$, then the same equation has an infinity of solutions.

Unlike overdetermined systems, working with underdetermined systems is relatively rare in econometrics and economics more generally. It usually means that you have insufficient theory to pin down the endogenous variables in your model—you need to find another equation (an arbitrage condition, a feasibility constraint, etc.) before the model can be solved. In other words, the problem is not how to find a clever method to solve underdetermined systems. Rather, it is to better understand the underlying economic problem, and write down a system that is fully determined.

[2] More formally, if $K < N$ then any K dimensional subspace of $\mathbb{R}^N$ has *Lebesgue measure zero* in $\mathbb{R}^N$. This result is available in most texts on measure theory.

1.4 Other Matrix Operations

[Roadmap]

1.4.1 Types of Matrices

For a square $N \times N$ matrix $\mathbf{A}$, the N elements of the form a_{nn} for $n = 1, \ldots, N$ are called the **principal diagonal**:

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & a_{2N} \\ \vdots & \vdots & & \vdots \\ a_{N1} & a_{N2} & \cdots & a_{NN} \end{pmatrix}$$

A square matrix $\mathbf{A}$ is called **diagonal** if all entries off the principal diagonal are zero. For example, the identity matrix is diagonal. The following notation is often used to define diagonal matrices:

$$\operatorname{diag}(d_1, \ldots, d_N) := \begin{pmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & d_N \end{pmatrix}$$

The **k-th power** of a square matrix $\mathbf{A}$ is written $\mathbf{A}^k$ and indicates $\mathbf{A} \cdots \mathbf{A}$ with k terms. If it exists, the **square root** of $\mathbf{A}$ is written $\mathbf{A}^{1/2}$ and defined as the matrix such that $\mathbf{A}^{1/2}\mathbf{A}^{1/2}$ is $\mathbf{A}$.

With diagonal matrices we can compute powers, roots, inverses and products very easily:

Fact 1.4.1. Let $\mathbf{C} = \operatorname{diag}(c_1, \ldots, c_N)$ and $\mathbf{D} = \operatorname{diag}(d_1, \ldots, d_N)$. The following statements are true:

1. $\mathbf{CD} = \operatorname{diag}(c_1 d_1, \ldots, c_N d_N)$.
2. $\mathbf{D}^k = \operatorname{diag}(d_1^k, \ldots, d_N^k)$ for any $k \in \mathbb{N}$.
3. If $d_n \geq 0$ for all n, then $\mathbf{D}^{1/2}$ exists and equals $\operatorname{diag}(\sqrt{d_1}, \ldots, \sqrt{d_N})$.
4. If $d_n \neq 0$ for all n, then $\mathbf{D}$ is nonsingular and $\mathbf{D}^{-1} = \operatorname{diag}(d_1^{-1}, \ldots, d_N^{-1})$.

You can check part 1 from the definition of matrix multiplication. The other parts follow directly.

A square matrix is called **lower triangular** if every element strictly above the principle diagonal is zero. It is called **upper triangular** if every element strictly below the principle diagonal is zero. For example, in

$$\mathbf{L} := \begin{pmatrix} 1 & 0 & 0 \\ 2 & 5 & 0 \\ 3 & 6 & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{U} := \begin{pmatrix} 1 & 2 & 3 \\ 0 & 5 & 6 \\ 0 & 0 & 1 \end{pmatrix}$$

the matrices $\mathbf{L}$ and $\mathbf{U}$ are lower and upper triangular respectively. The great advantage of triangular matrices is that the associated linear equations are trivial to solve using either forward or backward substitution. For example, with the system

$$\begin{pmatrix} 1 & 0 & 0 \\ 2 & 5 & 0 \\ 3 & 6 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} x_1 \\ 2x_1 + 5x_2 \\ 3x_1 + 6x_2 + x_3 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}$$

the top equation involves only x_1, so we can solve for its value directly. Plugging that value into the second equation, we can solve out for x_2 and so on.

Fact 1.4.2. If $\mathbf{A} = (a_{mn})$ is triangular, then $\det(\mathbf{A}) = \prod_{n=1}^{N} a_{nn}$.

1.4.2 Transpose and Trace

The **transpose** of $N \times K$ matrix $\mathbf{A}$ is a $K \times N$ matrix $\mathbf{A}'$ such that $\text{col}_n(\mathbf{A}') = \text{row}_n(\mathbf{A})$. For example, given

$$\mathbf{A} := \begin{pmatrix} 10 & 40 \\ 20 & 50 \\ 30 & 60 \end{pmatrix} \qquad \mathbf{B} := \begin{pmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{pmatrix} \tag{1.8}$$

the transposes are

$$\mathbf{A}' = \begin{pmatrix} 10 & 20 & 30 \\ 40 & 50 & 60 \end{pmatrix} \qquad \mathbf{B}' := \begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{pmatrix}$$

Fact 1.4.3. For conformable matrices $\mathbf{A}$ and $\mathbf{B}$, transposition satisfies

1. $(\mathbf{A}')' = \mathbf{A}$.
2. $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$.
3. $(\mathbf{A} + \mathbf{B})' = \mathbf{A}' + \mathbf{B}'$.
4. $(c\mathbf{A})' = c\mathbf{A}'$ for any constant c.

Fact 1.4.4. For each square matrix $\mathbf{A}$, we have

1. $\det(\mathbf{A}') = \det(\mathbf{A})$, and
2. If $\mathbf{A}$ is nonsingular then so is $\mathbf{A}'$, and $(\mathbf{A}')^{-1} = (\mathbf{A}^{-1})'$.

A square matrix $\mathbf{A}$ is called **symmetric** if $\mathbf{A}' = \mathbf{A}$. This is equivalent to the statement that $a_{nk} = a_{kn}$ for every k and n. Note that $\mathbf{A}'\mathbf{A}$ and $\mathbf{A}\mathbf{A}'$ are always well-defined and symmetric.

The **trace** of a square matrix is the sum of the elements on its principal diagonal:

$$\operatorname{trace}\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & a_{2N} \\ \vdots & \vdots & & \vdots \\ a_{N1} & a_{N2} & \cdots & a_{NN} \end{pmatrix} = \sum_{n=1}^{N} a_{nn}$$

Fact 1.4.5. Transposition does not alter trace: $\operatorname{trace}(\mathbf{A}) = \operatorname{trace}(\mathbf{A}')$.

Fact 1.4.6. If $\mathbf{A}$ and $\mathbf{B}$ are $N \times N$ matrices and α and β are two scalars, then

$$\operatorname{trace}(\alpha\mathbf{A} + \beta\mathbf{B}) = \alpha \operatorname{trace}(\mathbf{A}) + \beta \operatorname{trace}(\mathbf{B})$$

Moreover, if $\mathbf{A}$ is $N \times M$ and $\mathbf{B}$ is $M \times N$, then $\operatorname{trace}(\mathbf{AB}) = \operatorname{trace}(\mathbf{BA})$.

The rank of a matrix can be difficult to determine. One case where it is easy is where the matrix is idempotent. A square matrix $\mathbf{A}$ is called **idempotent** if $\mathbf{AA} = \mathbf{A}$.

Fact 1.4.7. If $\mathbf{A}$ is idempotent, then $\operatorname{rank}(\mathbf{A}) = \operatorname{trace}(\mathbf{A})$.

1.4.3 Eigenvalues and Eigenvectors

Let $\mathbf{A}$ be $N \times N$. As in § 1.3.2, think of $\mathbf{A}$ as a linear map, so that $\mathbf{Ax}$ is the image of $\mathbf{x}$ under $\mathbf{A}$. In general $\mathbf{A}$ will map $\mathbf{x}$ to some arbitrary new location but sometimes $\mathbf{x}$ will only be *scaled*. That is,

$$\mathbf{Ax} = \lambda \mathbf{x} \tag{1.9}$$

for some scalar λ. If $\mathbf{x}$ and λ satisfy (1.9) and $\mathbf{x}$ is nonzero, then $\mathbf{x}$ is called an **eigenvector** of $\mathbf{A}$, λ is called an **eigenvalue** and $(\mathbf{x}, \lambda)$ is called an **eigenpair**. For example, if $\mathbf{I}$ is the $N \times N$ identity then $(1, \mathbf{x})$ is an eigenpair of $\mathbf{I}$ for every nonzero $\mathbf{x} \in \mathbb{R}^N$. Evidently any scalar multiple of an eigenvector of $\mathbf{A}$ is also an eigenvector of $\mathbf{A}$.

We can given many other examples of matrices with eigenpairs. But what about the matrix

$$\mathbf{R} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

This matrix induces counter-clockwise rotation on any point by 90°. For any *real* number λ and $\mathbf{x} \in \mathbb{R}^2$ the scaling in (1.9) clearly fails. However, if we admit the possibility that λ and the elements of $\mathbf{x}$ can be complex we find that (1.9) can hold. Thus, for this rotation matrix, you will be able to confirm that $\lambda = i$ and $\mathbf{x} = (1, -i)'$ is an eigenpair for $\mathbf{R}$. This example shows that contemplation of complex eigenpairs is useful. Hence an eigenpair is always taken to be complex valued unless explicitly stated to be real.

Fact 1.4.8. If $\mathbf{A}$ is $N \times N$ and $\mathbf{I}$ is the $N \times N$ identity, then λ is an eigenvalue of $\mathbf{A}$ if and only if

$$\det(\mathbf{A} - \lambda \mathbf{I}) = 0$$

For the 2×2 matrix in (1.7), one can use the rule for the 2×2 determinant in (1.7), fact 1.4.8 and a little bit of algebra to show that its eigenvalues are given by the two roots of the polynomial expression

$$\lambda^2 - (a + d)\lambda + (ad - bc) = 0$$

More generally, given any $N \times N$ matrix $\mathbf{A}$, it can be shown via the fundamental theorem of algebra that there exist complex numbers $\lambda_1, \ldots, \lambda_N$, not necessarily distinct, such that

$$\det(\mathbf{A} - \lambda \mathbf{I}) = \prod_{n=1}^{N} (\lambda_n - \lambda) \tag{1.10}$$

It is clear that each λ_n satisfies $\det(\mathbf{A} - \lambda_n \mathbf{I}) = 0$ and hence is an eigenvalue of $\mathbf{A}$. In particular, $\lambda_1, \ldots, \lambda_N$ is the set of eigenvalues of $\mathbf{A}$, although we stress again that not all are necessarily distinct.

Fact 1.4.9. Let $\mathbf{A}$ be $N \times N$ and let $\lambda_1, \ldots, \lambda_N$ be the eigenvalues defined in (1.10). The following statements are true:

1. $\det(\mathbf{A}) = \prod_{n=1}^N \lambda_n$.
2. $\text{trace}(\mathbf{A}) = \sum_{n=1}^N \lambda_n$.
3. If $\mathbf{A}$ is symmetric, then $\lambda_n \in \mathbb{R}$ for all n.
4. If $\mathbf{A}$ is nonsingular, then the eigenvalues of $\mathbf{A}^{-1}$ are $1/\lambda_1, \ldots, 1/\lambda_N$.
5. If $\mathbf{A} = \text{diag}(d_1, \ldots, d_N)$, then $\lambda_n = d_n$ for all n.

It follows immediately from item 1 of fact 1.4.9 that $\mathbf{A}$ is nonsingular if and only if all its eigenvalues are nonzero.

1.4.4 Similar Matrices

An important concept in the field of dynamic systems is *conjugacy*. For example, let $f \colon A \to A$ where A is any set. We are often interested in the evolution of sequences defined recursively by such maps:

$$a_{t+1} = f(a_t), \quad a_0 = \text{ a given point in } A$$

Evidently this sequence also satisfies $a_t = f^t(a_0)$, where f^t is the t-th composition of f with itself.

Even when f is known, the properties of f^t are not always easy to discern. In this situation it can help to have another map $g \colon B \to B$ that is **conjugate** to f. That is, there exists some bijection $\tau \colon B \to A$ such that

$$f = \tau \circ g \circ \tau^{-1}$$

This means that $f(a) = \tau(g(\tau^{-1}(a)))$ for all $a \in A$. Observe that, under this conjugacy,

$$f^2 = f \circ f = \tau \circ g \circ \tau^{-1} \circ \tau \circ g \circ \tau^{-1} = \tau \circ g^2 \circ \tau^{-1}$$

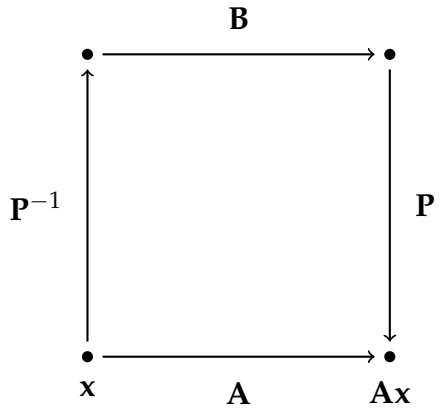


Figure 1.7: **A** is similar to **B**

In other words, f^2 is likewise conjugate to g^2. An inductive argument shows that the same is true for all t, so $f^t = \tau \circ g^t \circ \tau^{-1}$. This is handy if computing g^t is in some sense easier than computing f^t. We'll see an example of this in just a moment.

In the case of linear maps—that is, matrices—it is natural to study conjugacy in a setting where the bijection is also required to be linear. In most texts this is called *similarity*. In particular, matrix **A** is said to be **similar** to a matrix **B** if there exists an invertible matrix **P** such that $\mathbf{A} = \mathbf{P}\mathbf{B}\mathbf{P}^{-1}$. Figure 1.7 shows the conjugate relationship of the two matrices when thought of as maps.

As a special case of the reasoning for f and g given above, we get

Fact 1.4.10. If **A** is similar to **B**, then $\mathbf{A}^t$ is similar to $\mathbf{B}^t$ for all $t \in \mathbb{N}$.

As discussed above, similarity of **A** to a given matrix **B** is most useful when **B** is somehow simpler than **A**, or more amenable to a given operation. About the simplest kind of matrices we work with are diagonal matrices. Hence similarity to a diagonal matrix is particularly desirable. If **A** is similar to a diagonal matrix, then **A** is called **diagonalizable**.

One instance where this scenario is useful is in studying $\mathbf{A}^t$ for some given $t \in \mathbb{N}$. If **A** is diagonalizable with $\mathbf{A} = \mathbf{P}\mathbf{D}\mathbf{P}^{-1}$ for some $\mathbf{D} = \operatorname{diag}(\lambda_1, \ldots, \lambda_N)$, then, in view of fact 1.4.10 and fact 1.4.1 on page 27, we have

$$\mathbf{A}^t = \mathbf{P} \operatorname{diag}(\lambda_1^t, \ldots, \lambda_N^t)\mathbf{P}^{-1}$$

In this expression we've used the symbol λ_n for the scalars, which is reminiscent of our notation for eigenvalues. There is a reason for this:

Theorem 1.4.1. *Let* $\mathbf{A}$ *be an* $N \times N$ *matrix. The following statements are true:*

1. *If* $\mathbf{A} = \mathbf{PDP}^{-1}$ *for some* $\mathbf{D} = \operatorname{diag}(\lambda_1, \ldots, \lambda_N)$*, then* $(\operatorname{col}_n(\mathbf{P}), \lambda_n)$ *is an eigenpair of* $\mathbf{A}$ *for each* n*.*
2. *Conversely, if* $\mathbf{A}$ *had* N *linearly independent eigenvectors, then* $\mathbf{A}$ *can be diagonalized via* $\mathbf{A} = \mathbf{PDP}^{-1}$*, where* $(\operatorname{col}_n(\mathbf{P}), \lambda_n)$ *is an eigenpair of* $\mathbf{A}$ *for each* n*.*

Proof. If $\mathbf{A} = \mathbf{PDP}^{-1}$ then $\mathbf{AP} = \mathbf{PD}$. Equating the n-th column on each side gives $\mathbf{A}\mathbf{p}_n = \lambda_n \mathbf{p}_n$, where $\mathbf{p}_n := \operatorname{col}_n(\mathbf{P})$. Thus, to show that $(\mathbf{p}_n, \lambda_n)$ is an eigenpair, we need only check that $\mathbf{p}_n$ is not the zero vector. In fact this is immediate, because if it was the zero vector then $\mathbf{P}$ would not be invertible. (Why?)

Conversely, if $\mathbf{A}$ has N linearly independent eigenvectors $\mathbf{p}_1, \ldots, \mathbf{p}_N$, then by forming $\mathbf{P}$ via $\operatorname{col}_n(\mathbf{P}) = \mathbf{p}_n$ and taking $\mathbf{D} = \operatorname{diag}(\lambda_1, \ldots, \lambda_N)$ where λ_n is the eigenvalue associated with $\mathbf{p}_n$, we can stack the individual vector equations $\mathbf{A}\mathbf{p}_n = \lambda_n \mathbf{p}_n$ into the matrix form $\mathbf{AP} = \mathbf{PD}$. Using the invertibility implied by linear independence, we then have $\mathbf{A} = \mathbf{PDP}^{-1}$. In particular, $\mathbf{A}$ is diagonalizable. □

1.4.5 Matrix Norm and Neumann Series

Consider the vector difference equation $\mathbf{x}_{t+1} = \mathbf{A}\mathbf{x}_t + \mathbf{b}$, where $\mathbf{x}_t \in \mathbb{R}^N$ represents the values of some variables of interest (e.g., consumption, investment, etc.) at time t, and $\mathbf{A}$ and $\mathbf{b}$ form the parameters in the law of motion for $\mathbf{x}_t$. One question of interest for such systems is whether or not there is a vector $\mathbf{x} \in \mathbb{R}^N$ such that $\mathbf{x}_t = \mathbf{x}$ implies $\mathbf{x}_{t+1} = \mathbf{x}$. In other words, we seek a $\mathbf{x} \in \mathbb{R}^N$ that solves the system of equations

$$\mathbf{x} = \mathbf{A}\mathbf{x} + \mathbf{b}, \quad \text{where} \quad \mathbf{A} \text{ is } N \times N \text{ and } \mathbf{b} \text{ is } N \times 1$$

In considering this problem we can get some insight from the scalar case $x = ax + b$. Here we know that if $|a| < 1$, then this equation has the solution

$$\bar{x} = \frac{b}{1 - a} = b \sum_{k=0}^{\infty} a^k$$

The second equality follows from elementary results on geometric series.

It turns out a very similar result is true in $\mathbb{R}^N$ if we replace the condition $|a| < 1$ with $\|\mathbf{A}\| < 1$ where $\|\mathbf{A}\|$ is the **matrix norm** or **operator norm** of $\mathbf{A}$. This is defined

for any $N \times K$ matrix $\mathbf{A}$ as

$$\|\mathbf{A}\| := \max \left\{ \frac{\|\mathbf{A}\mathbf{x}\|}{\|\mathbf{x}\|} \ : \ \mathbf{x} \in \mathbb{R}^K, \ \mathbf{x} \neq \mathbf{0} \right\} \tag{1.11}$$

Note that in this (standard) notation there are two different norms in play. The left hand side is a matrix norm. The norm expressions on the right hand side are ordinary Euclidean vector norms.

The matrix norm behaves very much like the Euclidean norm. For example,

Fact 1.4.11. For any $N \times K$ matrices $\mathbf{A}$ and $\mathbf{B}$, the matrix norm satisfies

1. $\|\mathbf{A}\| = \mathbf{0}$ if and only if all entries of $\mathbf{A}$ are zero.
2. $\|\alpha \mathbf{A}\| = |\alpha| \|\mathbf{A}\|$ for any scalar α.
3. $\|\mathbf{A} + \mathbf{B}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\|$.
4. If $\mathbf{A}$ is a square matrix, then

$$\|\mathbf{A}\| = \max\{\lambda^{1/2} : \lambda \text{ is an eigenvalue of } \mathbf{A}'\mathbf{A}\} \tag{1.12}$$

Returning to the problem at hand, it's clear from (1.11) that if $\|\mathbf{A}\| < 1$, then any nonzero point is "contracted" by $\mathbf{A}$, in the sense of being pulled closer to the origin (i.e., $\|\mathbf{A}\mathbf{x}\| < \|\mathbf{x}\|$). In this sense its implications are similar to $|a| < 1$ in the scalar case. In particular, we have the following parallel result:

Theorem 1.4.2. *Let $\mathbf{b}$ be any vector in $\mathbb{R}^N$ and let $\mathbf{I}$ be the $N \times N$ identity. If $\mathbf{A}$ is an $N \times N$ matrix with $\|\mathbf{A}^j\| < 1$ for some $j \in \mathbb{N}$, then $\mathbf{I} - \mathbf{A}$ is invertible, and the system of equations $\mathbf{x} = \mathbf{A}\mathbf{x} + \mathbf{b}$ has the unique solution*

$$\bar{\mathbf{x}} = (\mathbf{I} - \mathbf{A})^{-1}\mathbf{b} = \mathbf{b} \sum_{i=0}^{\infty} \mathbf{A}^i$$

The second equality means that $\mathbf{b} \sum_{i=0}^{k} \mathbf{A}^i$ converges to $(\mathbf{I} - \mathbf{A})^{-1}\mathbf{b}$ as $k \to \infty$. The infinite sum is called the **Neumann series** associated with $\mathbf{A}$.

One way to test the conditions of theorem 1.4.2 is to use (1.12). Another is to use the following fact, which restricts the eigenvalues of $\mathbf{A}$ directly, rather than $\mathbf{A}'\mathbf{A}$. It is based on the **spectral radius**, which is defined for square $\mathbf{A}$ as

$$\varrho(\mathbf{A}) := \max\{|\lambda| : \lambda \text{ is an eigenvalue of } \mathbf{A}\} \tag{1.13}$$

Here $|\lambda|$ is the **modulus** of the possibly complex number λ. In particular, if $\lambda = a + ib$, then $|\lambda| = (a^2 + b^2)^{1/2}$. If $\lambda \in \mathbb{R}$ then this reduces to the usual notion of absolute value.

Fact 1.4.12. If $\varrho(\mathbf{A}) < 1$, then $\|\mathbf{A}^j\| < 1$ for some $j \in \mathbb{N}$.

1.4.6 Quadratic Forms

We've spent a lot of time discussing linear maps, one class of which is the linear real-valued maps. By theorem 1.3.1 we know that any linear map from $\mathbb{R}^N$ to $\mathbb{R}$ takes the form $\mathbf{x} \mapsto \mathbf{x}'\mathbf{a}$ for some vector $\mathbf{a} \in \mathbb{R}^N$. The next level of complexity is quadratic real-valued maps. To describe them, let $\mathbf{A}$ be $N \times N$ and symmetric, and let $\mathbf{x}$ be $N \times 1$. The **quadratic function** or **quadratic form** on $\mathbb{R}^N$ associated with $\mathbf{A}$ is the map Q defined by

$$Q(\mathbf{x}) := \mathbf{x}'\mathbf{A}\mathbf{x} = \sum_{j=1}^{N} \sum_{i=1}^{N} a_{ij} x_i x_j$$

To give a simple illustration, let $N = 2$ and let $\mathbf{A}$ be the identity matrix $\mathbf{I}$. In this case,

$$Q(\mathbf{x}) = \|\mathbf{x}\|^2 = x_1^2 + x_2^2$$

A 3D graph of this function is shown in figure 1.8.

One thing you'll notice about this function is that its graph lies everywhere above zero, or $Q(\mathbf{x}) \geq 0$. In fact we know that $\|\mathbf{x}\|^2$ is nonnegative and will be zero only when $\mathbf{x} = \mathbf{0}$. Hence the graph touches zero only at the point $\mathbf{x} = \mathbf{0}$. Many other choices of $\mathbf{A}$ yield a quadratic form with this property. Such $\mathbf{A}$ are said to be positive definite. More generally, an $N \times N$ symmetric matrix $\mathbf{A}$ is called

- **nonnegative definite** if $\mathbf{x}'\mathbf{A}\mathbf{x} \geq 0$ for all $\mathbf{x} \in \mathbb{R}^N$,
- **positive definite** if $\mathbf{x}'\mathbf{A}\mathbf{x} > 0$ for all $\mathbf{x} \in \mathbb{R}^N$ with $\mathbf{x} \neq \mathbf{0}$,
- **nonpositive definite** if $\mathbf{x}'\mathbf{A}\mathbf{x} \leq 0$ for all $\mathbf{x} \in \mathbb{R}^N$, and
- **negative definite** if $\mathbf{x}'\mathbf{A}\mathbf{x} < 0$ for all $\mathbf{x} \in \mathbb{R}^N$ with $\mathbf{x} \neq \mathbf{0}$.

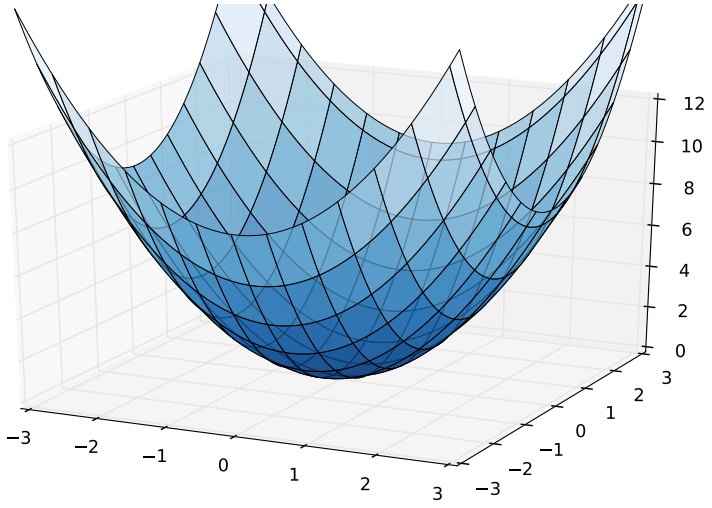


Figure 1.8: The quadratic function $Q(\mathbf{x}) = x_1^2 + x_2^2$

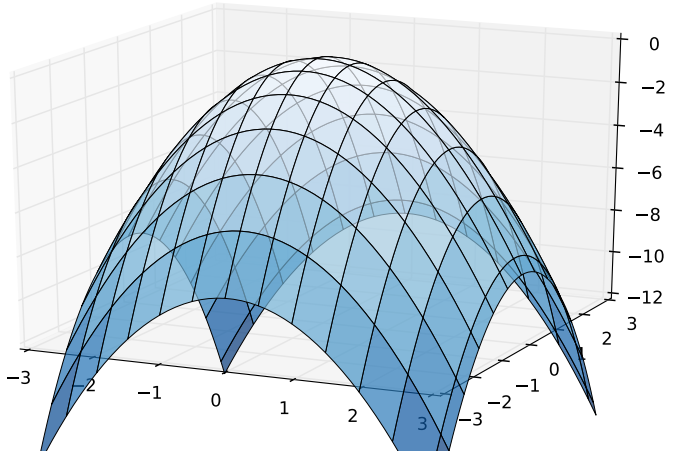


Figure 1.9: The quadratic function $Q(\mathbf{x}) = -x_1^2 - x_2^2$

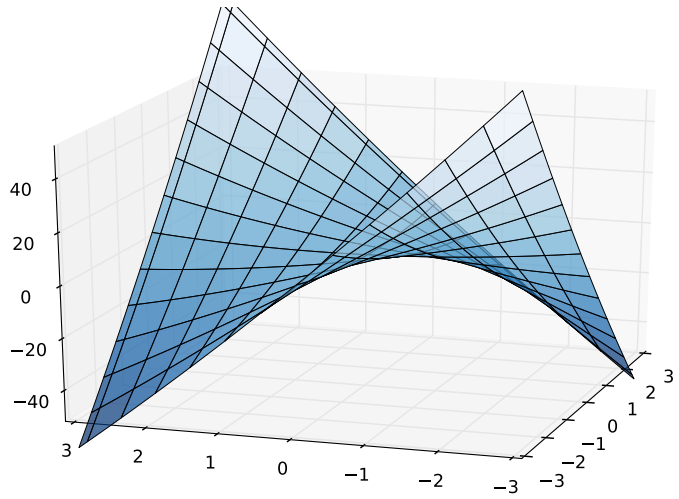


Figure 1.10: The quadratic function $Q(\mathbf{x}) = x_1^2/2 + 8x_1x_2 + x_2^2/2$

If **A** fits none of these categories then **A** is called **indefinite**. Figure 1.9 shows the graph of a negative definite quadratic function. Now the function is hill-shaped, and **0** is the unique global maximum. Figure 1.10 shows an indefinite form.

The easiest case for detecting definiteness is when the matrix **A** is diagonal, since

$$\mathbf{A} = \operatorname{diag}(d_1, \ldots, d_N) \quad \text{implies} \quad Q(\mathbf{x}) = d_1 x_1^2 + \cdots + d_N x_N^2$$

From the right hand expression we see that a diagonal matrix is positive definite if and only if all diagonal elements are positive. Analogous statements are true for nonnegative, nonpositive and negative definite matrices. The next fact generalizes this idea and is proved in §3.1.2.

Fact 1.4.13. Let **A** be any symmetric matrix. **A** is

1. positive definite if and only if its eigenvalues are all positive,
2. negative definite if and only if its eigenvalues are all negative,

and similarly for nonpositive and nonnegative definite.

It immediately follows from fact 1.4.13 (why?) that

Fact 1.4.14. If $\mathbf{A}$ is positive definite, then $\mathbf{A}$ is nonsingular, with $\det(\mathbf{A}) > 0$.

Finally, here's a necessary (but not sufficient) condition for each kind of definiteness.

Fact 1.4.15. If $\mathbf{A}$ is positive definite, then each element a_{nn} on the principal diagonal is positive, and the same for nonnegative, nonpositive and negative.

1.5 Further Reading

To be written.

1.6 Exercises

Ex. 1.6.1. Given two vectors $\mathbf{x}$ and $\mathbf{y}$, show that $|\|\mathbf{x}\| - \|\mathbf{y}\|| \leq \|\mathbf{x} - \mathbf{y}\|$.[3]

Ex. 1.6.2. Use fact 1.1.2 on page 5 to show that if $\mathbf{y} \in \mathbb{R}^N$ is such that $\mathbf{y}'\mathbf{x} = 0$ for every $\mathbf{x} \in \mathbb{R}^N$, then $\mathbf{y} = \mathbf{0}$.

Ex. 1.6.3. Fix nonzero $\mathbf{x} \in \mathbb{R}^N$. Consider the optimization problem

$$\max_{\mathbf{y}} \mathbf{x}'\mathbf{y} \quad \text{subject to} \quad \mathbf{y} \in \mathbb{R}^N \text{ and } \|\mathbf{y}\| = 1$$

Show that the maximizer is $\hat{\mathbf{x}} := \mathbf{x}/\|\mathbf{x}\|$.[4]

Ex. 1.6.4. Show that if S and S' are two linear subspaces of $\mathbb{R}^N$, then $S \cap S'$ is also a linear subspace of $\mathbb{R}^N$.

Ex. 1.6.5. Show that every linear subspace of $\mathbb{R}^N$ contains the origin $\mathbf{0}$.

Ex. 1.6.6. Show that the vectors $(1, 1)$ and $(-1, 2)$ are linearly independent.[5]

[3] Hint: Use the triangle inequality.

[4] Hint: There's no need to go taking derivatives and setting them equal to zero. An easier proof exists. If you're stuck, consider the Cauchy-Schwarz inequality.

[5] Hint: Look at the different definitions of linear independence. Choose the one that's easiest to work with in terms of algebra.

Ex. 1.6.7. Let $\mathbf{a} \in \mathbb{R}^N$ and let $A := \{\mathbf{x} \in \mathbb{R}^N : \mathbf{a}'\mathbf{x} = 0\}$. Show that A is a linear subspace of $\mathbb{R}^N$.

Ex. 1.6.8. Let Q be the subset of $\mathbb{R}^3$ defined by

$$Q := \{(x_1, x_2, x_3) \in \mathbb{R}^3 : x_2 = x_1 + x_3\}$$

Is Q a linear subspace of $\mathbb{R}^3$? Why or why not?

Ex. 1.6.9. Let Q be the subset of $\mathbb{R}^3$ defined by

$$Q := \{(x_1, x_2, x_3) \in \mathbb{R}^3 : x_2 = 1\}$$

Is Q a linear subspace of $\mathbb{R}^3$? Why or why not?

Ex. 1.6.10. Show that if $T\colon \mathbb{R}^N \to \mathbb{R}^N$ is a linear function and λ is any scalar, then $E := \{\mathbf{x} \in \mathbb{R}^N : T\mathbf{x} = \lambda\mathbf{x}\}$ is a linear subspace of $\mathbb{R}^N$.

Ex. 1.6.11. Prove the equivalences in fact 1.1.6 on page 10.

Ex. 1.6.12. Prove fact 1.1.7 on page 11.

Ex. 1.6.13. Show that if S is a linear subspace of $\mathbb{R}^N$ then every basis of S has the same number of elements.

Ex. 1.6.14. Prove fact 1.1.8 on page 14.

Ex. 1.6.15. Prove theorem 1.2.1 on page 16.

Ex. 1.6.16. Show that if $T\colon \mathbb{R}^K \to \mathbb{R}^N$ is linear and $K > N$, then T is not one-to-one.

Ex. 1.6.17. Prove the claim in fact 1.4.14 on page 38 that if $\mathbf{A}$ is positive definite, then $\mathbf{A}$ is nonsingular. If you can, prove it without invoking positivity of its eigenvalues.

Ex. 1.6.18. Prove fact 1.4.15 on page 38.

Ex. 1.6.19. Show that for any two conformable matrices $\mathbf{A}$ and $\mathbf{B}$, we have $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$.[6]

Ex. 1.6.20. Let $\mathbf{A}$ be a constant $N \times N$ matrix. Assuming existence of the inverse $\mathbf{A}^{-1}$, show that $(\mathbf{A}')^{-1} = (\mathbf{A}^{-1})'$.

[6] Hint: Look at the definition of the inverse! Always look at the definition, and then show that the object in question has the stated property.

Ex. 1.6.21. Show that if $\mathbf{e}_i$ and $\mathbf{e}_j$ are the i-th and j-th canonical basis vectors of $\mathbb{R}^N$ respectively, and $\mathbf{A}$ is an $N \times N$ matrix, then $\mathbf{e}_i' \mathbf{A} \mathbf{e}_j = a_{ij}$, the i, j-th element of $\mathbf{A}$.

Ex. 1.6.22. Prove fact 1.3.6 on page 26.

Ex. 1.6.23. Let

$$\mathbf{A} := \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}, \qquad \mathbf{B} := \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$$

Show that

1. $\mathbf{A}$ is nonnegative definite.
2. $\mathbf{B}$ is *not* positive definite.

Ex. 1.6.24. Let $\mathbf{A}_1, \ldots, \mathbf{A}_J$ be invertible matrices. Use proof by induction and fact 1.3.4 on page 24 to show that the product of these matrices is invertible, and, in particular, that

$$(\mathbf{A}_1 \mathbf{A}_2 \cdots \mathbf{A}_J)^{-1} = \mathbf{A}_J^{-1} \cdots \mathbf{A}_2^{-1} \mathbf{A}_1^{-1}$$

Ex. 1.6.25. Show that for any matrix $\mathbf{A}$, the matrix $\mathbf{A}'\mathbf{A}$ is well-defined (i.e., multiplication is possible), square, and nonnegative definite.

Ex. 1.6.26. Show that if $\mathbf{A}$ and $\mathbf{B}$ are positive definite and $\mathbf{A} + \mathbf{B}$ is well defined, then it is also positive definite.

Ex. 1.6.27. Let $\mathbf{A}$ be $N \times K$. Show that if $\mathbf{A}\mathbf{x} = \mathbf{0}$ for all $K \times 1$ vectors $\mathbf{x}$, then $\mathbf{A} = \mathbf{0}$ (i.e., every element of $\mathbf{A}$ is zero). Show as a corollary that if $\mathbf{A}$ and $\mathbf{B}$ are $N \times K$ and $\mathbf{A}\mathbf{x} = \mathbf{B}\mathbf{x}$ for all $K \times 1$ vectors $\mathbf{x}$, then $\mathbf{A} = \mathbf{B}$.

Ex. 1.6.28. Let $\mathbf{I}_N$ be the $N \times N$ identity matrix.

1. Explain briefly why $\mathbf{I}_N$ is full column rank.
2. Show that $\mathbf{I}_N$ is the inverse of itself.
3. Let $\mathbf{A} := \alpha \mathbf{I}_N$. Give a condition on α such that $\mathbf{A}$ is positive definite.

Ex. 1.6.29. Let $\mathbf{X} := \mathbf{I}_N - 2\mathbf{u}\mathbf{u}'$, where $\mathbf{u}$ is an $N \times 1$ vector with $\|\mathbf{u}\| = 1$. Show that $\mathbf{X}$ is symmetric and $\mathbf{X}\mathbf{X} = \mathbf{I}_N$.

Ex. 1.6.30. Recall the definition of similarity of matrices, as given in §1.4.4. Let's write $\mathbf{A} \sim \mathbf{B}$ if $\mathbf{A}$ is similar to $\mathbf{B}$. Show that $\sim$ is an **equivalence relation** on the set of $N \times N$ matrices. In particular, show that, for any $N \times N$ matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$, we have (i) $\mathbf{A} \sim \mathbf{A}$, (ii) $\mathbf{A} \sim \mathbf{B}$ implies $\mathbf{B} \sim \mathbf{A}$ and (iii) $\mathbf{A} \sim \mathbf{B}$ and $\mathbf{B} \sim \mathbf{C}$ implies $\mathbf{A} \sim \mathbf{C}$.

Ex. 1.6.31. Confirm the claim in fact 1.4.8 on page 30.

Ex. 1.6.32. Show that $\mathbf{A}$ is nonsingular if and only if 0 is *not* an eigenvalue for $\mathbf{A}$.

Ex. 1.6.33. Show that the only nonsingular idempotent matrix is the identity matrix.

Ex. 1.6.34. Let $\mathbf{1}$ be an $N \times 1$ vector of ones. Consider the matrix

$$\mathbf{Z} := \frac{1}{N}\mathbf{1}\mathbf{1}'$$

1. Show that if $\mathbf{x}$ is any $N \times 1$ vector, then $\mathbf{Z}\mathbf{x}$ is a vector with all elements equal to the mean of the elements of $\mathbf{x}$.
2. Show that $\mathbf{Z}$ is idempotent.

1.6.1 Solutions to Selected Exercises

Solution to Exercise 1.6.3. Fix nonzero $\mathbf{x} \in \mathbb{R}^N$. Let $\hat{\mathbf{x}} := \mathbf{x}/\|\mathbf{x}\|$. Comparing this point with any other $\mathbf{y} \in \mathbb{R}^N$ satisfying $\|\mathbf{y}\| = 1$, the Cauchy-Schwarz inequality yields

$$\mathbf{y}'\mathbf{x} \leq |\mathbf{y}'\mathbf{x}| \leq \|\mathbf{y}\|\|\mathbf{x}\| = \|\mathbf{x}\| = \frac{\mathbf{x}'\mathbf{x}}{\|\mathbf{x}\|} = \hat{\mathbf{x}}'\mathbf{x}$$

Hence $\hat{\mathbf{x}}$ is the maximizer, as claimed. □

Solution to Exercise 1.6.7. Let $\mathbf{x}, \mathbf{y} \in A$ and let $\alpha, \beta \in \mathbb{R}$. We must show that $\mathbf{z} := \alpha\mathbf{x} + \beta\mathbf{y} \in A$, or, equivalently, $\mathbf{a}'\mathbf{z} = \mathbf{a}'(\alpha\mathbf{x} + \beta\mathbf{y}) = 0$. This is immediate, because $\mathbf{a}'(\alpha\mathbf{x} + \beta\mathbf{y}) = \alpha\mathbf{a}'\mathbf{x} + \beta\mathbf{a}'\mathbf{y} = 0 + 0 = 0$. □

Solution to Exercise 1.6.8. If $\mathbf{a} := (1, -1, 1)$, then Q is all $\mathbf{x}$ with $\mathbf{a}'\mathbf{x} = 0$. This set is a linear subspace of $\mathbb{R}^3$, as shown in exercise 1.6.7. □

Solution to Exercise 1.6.11. We are asked to verify the equivalences in fact 1.1.6 on page 10 for the set $X := \{\mathbf{x}_1, \ldots, \mathbf{x}_K\}$. The right way to do this is to establish a cycle, such as part 1 implies part 2 implies part 3 implies part 1. It is then clear that part i implies part j for any i and j.

First let's show that part 1 implies part 2, which is that if X_0 is a proper subset of X, then $\operatorname{span}(X_0)$ is a proper subset of $\operatorname{span}(X)$. To save fiddly notation let's take $X_0 := \{\mathbf{x}_2, \ldots, \mathbf{x}_K\}$. Suppose to the contrary that $\operatorname{span}(X_0) = \operatorname{span}(X)$. Since $\mathbf{x}_1 \in \operatorname{span}(X)$

we must then have $\mathbf{x}_1 \in \operatorname{span}(X_0)$, from which we deduce the existence of scalars $\alpha_2, \ldots, \alpha_K$ such that $\mathbf{0} = -\mathbf{x}_1 + \alpha_2 \mathbf{x}_2 + \cdots + \alpha_K \mathbf{x}_K$. Since $-1 \neq 0$, this contradicts part 1.

The next claim is that part 2 implies part 3; that is, that so vector in X can be written as a linear combination of the others. Suppose to the contrary that $\mathbf{x}_1 = \alpha_2 \mathbf{x}_2 + \cdots + \alpha_K \mathbf{x}_K$, say. Let $\mathbf{y} \in \operatorname{span}(X)$, so that $\mathbf{y} = \beta_1 \mathbf{x}_1 + \cdots + \beta_K \mathbf{x}_K$. If we use the preceding equality to substitute out $\mathbf{x}_1$, we get $\mathbf{y}$ as a linear combination of $\{\mathbf{x}_2, \ldots, \mathbf{x}_K\}$ alone. In other words, any element of $\operatorname{span}(X)$ is in the span of the proper subset $\{\mathbf{x}_2, \ldots, \mathbf{x}_K\}$. Contradiction.

The final claim is that part 3 implies part 1; that is, that $\alpha_1 = \cdots = \alpha_K = 0$ whenever $\alpha_1 \mathbf{x}_1 + \cdots + \alpha_K \mathbf{x}_K = \mathbf{0}$. Suppose to the contrary that there exist scalars with $\alpha_1 \mathbf{x}_1 + \cdots + \alpha_K \mathbf{x}_K = \mathbf{0}$ and yet $\alpha_k \neq 0$ for at least one k. It follows immediately that $\mathbf{x}_k = (1/\alpha_k) \sum_{j \neq k} \alpha_j \mathbf{x}_j$. Contradiction. □

Solution to Exercise 1.6.12. The aim is to prove fact 1.1.7 on page 11. Regarding the first part, let's take X as linearly independent and show that the subset $X_0 := \{\mathbf{x}_1, \ldots, \mathbf{x}_{K-1}\}$ is linearly independent. (The argument for more general subsets is similar.) Suppose to the contrary that X_0 is linearly dependent. Then by the definition we can take $\alpha_1, \ldots, \alpha_{K-1}$ not all zero with $\sum_{k=1}^{K-1} \alpha_k \mathbf{x}_k = \mathbf{0}$. Letting $\alpha_K = 0$ we can write this as $\sum_{k=1}^{K} \alpha_k \mathbf{x}_k = \mathbf{0}$. Since not all coefficients are zero we have contradicted linear independence of X.

Regarding the second claim, let $X := \{\mathbf{x}_1, \ldots, \mathbf{x}_K\}$ be linearly independent and suppose that $\mathbf{x}_j = \mathbf{0}$. Then by setting $\alpha_k = 0$ for $k \neq j$ and $\alpha_j = 1$ we can form scalars not all equal to zero with $\sum_{k=1}^{K} \alpha_k \mathbf{x}_k = \mathbf{0}$.

Regarding the third claim, let $X := \{\mathbf{x}_1, \ldots, \mathbf{x}_K\} \subset \mathbb{R}^N$ be linearly independent and let $\mathbf{x}_{K+1}$ be any point in $\mathbb{R}^N$ such that $\mathbf{x}_{K+1} \notin \operatorname{span}(X)$. The claim is that $X \cup \{\mathbf{x}_{K+1}\}$ is linearly independent. Suppose to the contrary that there exist $\alpha_1, \ldots, \alpha_K, \alpha_{K+1}$ not all zero such that $\sum_{k=1}^{K+1} \alpha_k \mathbf{x}_k = \mathbf{0}$. There are two possibilities for α_{K+1}, both of which lead to a contradiction: First, if $\alpha_{K+1} = 0$, then, since $\alpha_1, \ldots, \alpha_K, \alpha_{K+1}$ are not all zero, at least one of $\alpha_1, \ldots, \alpha_K$ are nonzero, and, moreover, $\sum_{k=1}^{K} \alpha_k \mathbf{x}_k = \sum_{k=1}^{K+1} \alpha_k \mathbf{x}_k = \mathbf{0}$. This contradicts our assumption of independence on X. On the other hand, if $\alpha_{K+1} \neq 0$, then from $\sum_{k=1}^{K+1} \alpha_k \mathbf{x}_k = \mathbf{0}$ we can express $\mathbf{x}_{K+1}$ as a linear combination of elements of X. This contradicts the hypothesis that $\mathbf{x}_{K+1} \notin \operatorname{span}(X)$. □

Solution to Exercise 1.6.13. Let B_1 and B_2 be two bases of S, with K_1 and K_2 elements respectively. By definition, B_2 is a linearly independent subset of S. Moreover, S is spanned by the set B_1, which has K_1 elements. Applying theorem 1.1.2, we

see that B_2 has at most K_1 elements. That is, $K_2 \leq K_1$. Reversing the roles of B_1 and B_2 gives $K_1 \leq K_2$. □

Solution to Exercise 1.6.14. The aim is to prove fact 1.1.8 on page 14. Suppose that S and T are K-dimensional linear subspaces of $\mathbb{R}^N$ with $S \subset T$. We claim that $S = T$. To see this, observe that by the definition of dimension, S is equal to $\text{span}(B)$ where B is a set of K linearly independent basis vectors $\{\mathbf{b}_1, \ldots, \mathbf{b}_K\}$. If $S \neq T$, then there exists a vector $\mathbf{x} \in T$ such that $\mathbf{x} \notin \text{span}(B)$. In view of fact 1.1.6 on page 10, the set $\{\mathbf{x}, \mathbf{b}_1, \ldots, \mathbf{b}_K\}$ is linearly independent. Moreover, since $\mathbf{x} \in T$ and since $B \subset S \subset T$, we now have $K + 1$ linearly independent vectors inside T. At the same time, being K-dimensional, we know that T is spanned by K vectors. This contradicts theorem 1.1.2 on page 12.

Regarding part 2, suppose that S is an M-dimensional linear subspace of $\mathbb{R}^N$ where $M < N$ and yet $S = \mathbb{R}^N$. Then we have a space S spanned by $M < N$ vectors that at the same time contains the N linearly independent canonical basis vectors. We are led to another contradiction of theorem 1.1.2. Hence $S = \mathbb{R}^N$ cannot hold. □

Solution to Exercise 1.6.15. A collection of equivalent statements are usually proved via a cycle of implications, such as $1 \implies 2 \implies \cdots \implies 5 \implies 1$. However in this case the logic is clearer if we directly show that all statements are equivalent to linear independence of V.

First observe equivalence of the onto property and linear independence of V via

$$T \text{ onto} \iff \text{rng}(T) = \mathbb{R}^N \iff \text{span}(V) = \mathbb{R}^N$$

by lemma 1.2.1, and the last statement is equivalent to linear independence of V by theorem 1.1.3 on page 12.

Next let's show that $\ker(T) = \{\mathbf{0}\}$ implies linear independence of V. To this end, suppose that $\ker(T) = \{\mathbf{0}\}$ and let $\alpha_1, \ldots, \alpha_N$ be such that $\sum_{n=1}^N \alpha_n T\mathbf{e}_n = \mathbf{0}$. By linearity of T we then have $T(\sum_{n=1}^N \alpha_n \mathbf{e}_n) = \mathbf{0}$. Since $\ker(T) = \{\mathbf{0}\}$ this means that $\sum_{n=1}^N \alpha_n \mathbf{e}_n = \mathbf{0}$, which in view of independence of $\{\mathbf{e}_1, \ldots, \mathbf{e}_N\}$, implies $\alpha_1 = \cdots = \alpha_N = 0$. This establishes that V is linearly independent.

Now let's check that linear independence of V implies $\ker(T) = \{\mathbf{0}\}$. To this end, let $\mathbf{x}$ be a vector in $\mathbb{R}^N$ such that $T\mathbf{x} = \mathbf{0}$. We can represent $\mathbf{x}$ in the form $\sum_{n=1}^N \alpha_n \mathbf{e}_n$ for suitable scalars $\{\alpha_n\}$. From linearity and $T\mathbf{x} = \mathbf{0}$ we get $\sum_{n=1}^N \alpha_n T\mathbf{e}_n = \mathbf{0}$. By linear independence of V this implies that each $\alpha_n = 0$, whence $\mathbf{x} = \mathbf{0}$. Thus $\ker(T) = \{\mathbf{0}\}$ as claimed.

From fact 1.2.1 we have $\ker(T) = \{\mathbf{0}\}$ iff T is one-to-one, so we can now state the following equivalences

$$T \text{ onto} \iff V \text{ linearly independent} \iff T \text{ one-to-one} \tag{1.14}$$

Finally, if T is a bijection then T is onto and hence V is linearly independent by (1.14). Conversely, if V is linearly independent then T is both onto and one-to-one by (1.14). Hence T is a bijection. □

Solution to Exercise 1.6.16. Let T be as described in the exercise and let $K > N$. Seeking a contradiction, suppose in addition that T is one-to-one. Let $\{\alpha_k\}_{k=1}^K$ be such that $\sum_{k=1}^K \alpha_k T\mathbf{e}_k = \mathbf{0}$. By linearity, $T(\sum_{k=1}^K \alpha_k \mathbf{e}_k) = \mathbf{0}$, and since T is one-to-one and $T\mathbf{0} = \mathbf{0}$, this in turn implies that $\sum_{k=1}^K \alpha_k \mathbf{e}_k = \mathbf{0}$. Since the canonical basis vectors are linearly independent, it must be that $\alpha_1 = \cdots = \alpha_K = 0$. From this we conclude that $\{T\mathbf{e}_1, \ldots, T\mathbf{e}_K\}$ is linearly independent. Thus $\mathbb{R}^N$ contains K linearly independent vectors, despite the fact that $N < K$. This is impossible by theorem 1.1.2 on page 12. □

Solution to Exercise 1.6.17. Let $\mathbf{A}$ be positive definite and consider the following: If $\mathbf{A}$ is singular, then there exists nonzero $\mathbf{x}$ with $\mathbf{Ax} = \mathbf{0}$ (see theorem 1.3.2 on page 22). But then $\mathbf{x}'\mathbf{Ax} = 0$ for nonzero $\mathbf{x}$. Contradiction. □

Solution to Exercise 1.6.18. If $\mathbf{x} = \mathbf{e}_n$ then $\mathbf{x}'\mathbf{Ax} = a_{nn}$. The claim follows. □

Solution to Exercise 1.6.22. Since the columns of $\mathbf{A}$ consist of K vectors in $\mathbb{R}^N$, the fact that $K > N$ implies that not all of the columns of $\mathbf{A}$ are linearly independent. (Recall theorem 1.1.2 on page 12.) It follows that $\mathbf{Az} = \mathbf{0}$ for some nonzero $\mathbf{z}$ in $\mathbb{R}^K$, and hence that $\mathbf{A}\lambda\mathbf{z} = \mathbf{0}$ for any scalar λ. Now suppose that some $\mathbf{x}$ solves $\mathbf{Ax} = \mathbf{b}$. Then, for any $\lambda \in \mathbb{R}$, we have $\mathbf{Ax} + \mathbf{A}\lambda\mathbf{z} = \mathbf{A}(\mathbf{x} + \lambda\mathbf{z}) = \mathbf{b}$. This proves the claim. □

Solution to Exercise 1.6.28. The solutions are as follows: (1) $\mathbf{I}_N$ is full column rank because its columns are the canonical basis vectors, which are independent. (2) By definition, $\mathbf{B}$ is the inverse of $\mathbf{A}$ if $\mathbf{BA} = \mathbf{AB} = \mathbf{I}_N$. It follows immediately that $\mathbf{I}_N$ is the inverse of itself. (3) A sufficient condition is $\alpha > 0$. If this holds, then given $\mathbf{x} \neq \mathbf{0}$, we have $\mathbf{x}'\alpha\mathbf{I}_N\mathbf{x} = \alpha\|\mathbf{x}\|^2 > 0$. □

Solution to Exercise 1.6.29. First, $\mathbf{X}$ is symmetric because

$$\mathbf{X}' = (\mathbf{I}_N - 2\mathbf{u}\mathbf{u}')' = \mathbf{I}_N' - 2(\mathbf{u}\mathbf{u}')' = \mathbf{I}_N - 2(\mathbf{u}')'\mathbf{u}' = \mathbf{I}_N - 2\mathbf{u}\mathbf{u}' = \mathbf{X}$$

Second, $\mathbf{X}\mathbf{X} = \mathbf{I}_N$, because

$$\begin{aligned}\mathbf{X}\mathbf{X} = (\mathbf{I}_N - 2\mathbf{u}\mathbf{u}')(\mathbf{I}_N' - 2\mathbf{u}\mathbf{u}') &= \mathbf{I}_N\mathbf{I}_N - 2\mathbf{I}_N 2\mathbf{u}\mathbf{u}' + (2\mathbf{u}\mathbf{u}')(2\mathbf{u}\mathbf{u}') \\ &= \mathbf{I}_N - 4\mathbf{u}\mathbf{u}' + 4\mathbf{u}\mathbf{u}'\mathbf{u}\mathbf{u}' = \mathbf{I}_N - 4\mathbf{u}\mathbf{u}' + 4\mathbf{u}\mathbf{u}' = \mathbf{I}_N\end{aligned}$$

The second last equality is due to the assumption that $\mathbf{u}'\mathbf{u} = \|\mathbf{u}\|^2 = 1$. □

Solution to Exercise 1.6.33. Suppose that $\mathbf{A}$ is both idempotent and nonsingular. From idempotence we have $\mathbf{A}\mathbf{A} = \mathbf{A}$. Premultiplying by $\mathbf{A}^{-1}$ gives $\mathbf{A} = \mathbf{I}$. □

Chapter 2

Probability

[Roadmap]

2.1 Probabilistic Models

We begin with some foundations of probability theory. These involve a few set theoretic operations—see Appendix 4.1 for a review.

2.1.1 Sample Spaces

In setting up a probabilistic model, we always start with the notion of a **sample space**, which we think of as being the enumeration or "list" of all possible outcomes in a given random experiment. In general, the sample space can be any nonempty set, and is usually denoted Ω. A typical element of Ω is denoted ω. The general idea is that a realization of uncertainty will lead to the selection of a particular $\omega \in \Omega$.

Example 2.1.1. Let $\Omega := \{1, \ldots, 6\}$ represent the six different faces of a dice. A realization of uncertainty corresponds to a roll of the dice, with the outcome being an integer ω in the set $\{1, \ldots, 6\}$.

The specification of all possible outcomes Ω is one part of our model. The other thing we need to do is to assign probabilities to outcomes. One idea is to start by assigning appropriate probabilities to every ω in Ω. It turns out that this is not the

right way forward. Instead, the standard approach is to *directly assign probabilities to subsets of* Ω *instead*. In the language of probability theory, subsets of Ω are called **events**. The set of events is usually denoted by $\mathcal{F}$, and we follow this convention.[1]

Below we attach probabilities to events using notation $\mathbb{P}(B)$, where B is an event (i.e., $B \in \mathcal{F}$). The symbol $\mathbb{P}(B)$ should be interpreted as representing "the probability that event B occurs." The way you should think about it is this:

> $\mathbb{P}(B)$ represents the probability that when uncertainty is resolved and some $\omega \in \Omega$ is selected by "nature," the statement $\omega \in B$ is true.

To illustrate, consider again example 2.1.1. Let B be the event $\{1, 2\}$. The number $\mathbb{P}(B)$ represents the probability that the face ω selected by the roll is either 1 or 2.

Let Ω be any sample space. Two events we always find in $\mathcal{F}$ are Ω itself and the empty set $\emptyset$, the latter because the empty set is regarded as being a subset of every set.[2] In this context, Ω is called the **certain event** because it always occurs (regardless of which outcome ω is selected, $\omega \in \Omega$ is true by definition). The empty set $\emptyset$ is called the **impossible event**.

Remark 2.1.1. Why not start the other way, assigning probability to individual points $\omega \in \Omega$, and then working out the probability of events by looking at the probability of the points contained in those events? In fact this approach fails in many cases. For example, suppose we take Ω to be the interval $(0, 1)$ and set up a model where all numbers in $(0, 1)$ are "equally likely." In this scenario, it can be shown that the probability of an event of the form (a, b) is just the length of the interval, or $b - a$. Moreover, the probability of an individual point $x \in (0, 1)$ occuring should be less than the probability of some point in the interval $(x - \epsilon, x + \epsilon)$ occuring. Hence the probability of hitting x is less than 2ϵ for any ϵ we choose. No positive number is less than 2ϵ for any ϵ. Hence the probability of x must be zero. As a result, we can't build probabilities of events from probabilities of individual elements, since the probability of hitting any given point is zero. There is no way to build up a proper probabilistic model with this information alone.

[1] I'm skirting some technical details here. In many situations, we exclude some of the more "complex" subsets of Ω from $\mathcal{F}$ because they are so messy that assigning probabilities to these sets causes problems for the theory. See §2.1.4 for more discussion.

[2] Why? Because in mathematics every sensible mathematical statement must be either true or false, so $\emptyset \subset \Omega$ must be true or false. To show the statement true, we need to show that every element of $\emptyset$ is also in Ω. Since $\emptyset$ contains no elements we cannot test this statement directly. However it is certain not false, since no counterexample can be given. As a result we regard it to be true.

2.1.2 Probabilities

The second stage of our model construction is to assign probabilities to elements of $\mathcal{F}$. This is done with a function $\mathbb{P}$ that maps $\mathcal{F}$ to $[0,1]$. In order to make sure our model of probability is well behaved, we need to put certain restrictions on $\mathbb{P}$. For example, we wouldn't want to have a B with $\mathbb{P}(B) = -93$, as negative probabilities don't make much sense. These restrictions are imposed in the next definition:

A **probability** $\mathbb{P}$ on $(\Omega, \mathcal{F})$ is a function from $\mathcal{F}$ to $[0,1]$ that satisfies

1. $\mathbb{P}(\Omega) = 1$, and
2. Additivity: $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$ whenever $A, B \in \mathcal{F}$ with $A \cap B = \varnothing$.

Together, the triple $(\Omega, \mathcal{F}, \mathbb{P})$ is called a **probability space**. It describes a set of events and their probabilities for a given experiment. *Confession: I've simplified the standard definition of a probability space slightly to avoid technical discussions that we don't need to go into right now. An outline of the technical issues can be found in* §2.1.4.

The axioms in the definition of a probability are sensible. For example, it's clear why we require $\mathbb{P}(\Omega) = 1$, since the realization ω will always be chosen from Ω by its definition. Also, the additivity property is natural: To find the probability of a given event, we can determine all the different (i.e., *disjoint*) ways that the event could occur, and then sum their probabilities. The examples below reinforce this idea.

Example 2.1.2. Let $\Omega := \{1, \ldots, 6\}$ represent the six different faces of a dice, as in example 2.1.1. We define a function $\mathbb{P}$ over all $A \in \mathcal{F}$ by

$$\mathbb{P}(A) := \frac{\#A}{6} \quad \text{where} \quad \#A := \text{ number of elements in } A \tag{2.1}$$

For example, $\mathbb{P}\{2,4,6\} = 3/6 = 1/2$. Let's check the axioms that define a probability. It's easy to see that $0 \leq \mathbb{P}(A) \leq 1$ for any $A \in \mathcal{F}$, and that $\mathbb{P}(\Omega) = 1$. Regarding additivity, suppose that A and B are two disjoint subsets of $\{1, \ldots, 6\}$. Then $\#(A \cup B) = \#A + \#B$, since, by disjointness, the number of elements in the union is just the number contributed by A plus the number contributed by B. Hence

$$\mathbb{P}(A \cup B) = \frac{\#(A \cup B)}{6} = \frac{\#A + \#B}{6} = \frac{\#A}{6} + \frac{\#B}{6} = \mathbb{P}(A) + \mathbb{P}(B)$$

In the definition of $\mathbb{P}$, additivity was defined for pairs of sets, but this is enough to imply additivity over any disjoint finite union. In particular, if $A_1, \ldots, A_J$ are disjoint in the sense that $A_i \cap A_j = \emptyset$ whenever $i \neq j$, then

$$\mathbb{P}\left(\cup_{j=1}^{J} A_j\right) = \sum_{j=1}^{J} \mathbb{P}(A_j)$$

also holds. See exercise 2.8.1.

Example 2.1.3. Continuing example 2.1.2, if we roll the dice, the probability of getting an even number is the probability of getting a 2 plus that of getting a 4 plus that of getting a 6. Formally,

$$\begin{aligned}\mathbb{P}\{2,4,6\} &= \mathbb{P}[\{2\} \cup \{4\} \cup \{6\}] \\ &= \mathbb{P}\{2\} + \mathbb{P}\{4\} + \mathbb{P}\{6\} = 1/6 + 1/6 + 1/6 = 1/2\end{aligned}$$

Example 2.1.4. Consider a memory chip in a computer, made up of billions of tiny switches. Imagine that a random number generator accesses a subset of N switches, setting each one to "on" or "off" at random. One sample space for this experiment is

$$\Omega_0 := \{(b_1, \ldots, b_N) : b_n \in \{\text{on, off}\} \text{ for each } n\}$$

Letting zero represent off and one represent on, we can also use the more practical space

$$\Omega := \{(b_1, \ldots, b_N) : b_n \in \{0, 1\} \text{ for each } n\}$$

Thus, Ω is the set of all binary sequences of length N. As our probability, we define

$$\mathbb{P}(A) := 2^{-N}(\#A)$$

To see that this is indeed a probability on $(\Omega, \mathcal{F})$ we need to check that $0 \leq \mathbb{P}(A) \leq 1$ for all $A \subset \Omega$, that $\mathbb{P}(\Omega) = 1$, and that $\mathbb{P}$ is additive. Exercise 2.8.7 asks you to confirm that $\mathbb{P}$ is additive. That $\mathbb{P}(\Omega) = 1$ follows from the fact that the number of distinct binary sequences of length N is 2^N.

Now let's go back to the general case, where $(\Omega, \mathcal{F}, \mathbb{P})$ is an arbitrary probability space. From the axioms above, we can derive a suprising number of properties. Let's list the key ones, starting with the next fact.

Fact 2.1.1. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $A, B \in \mathcal{F}$. If $A \subset B$, then

1. $\mathbb{P}(B \setminus A) = \mathbb{P}(B) - \mathbb{P}(A)$;
2. $\mathbb{P}(A) \leq \mathbb{P}(B)$;
3. $\mathbb{P}(A^c) := \mathbb{P}(\Omega \setminus A) = 1 - \mathbb{P}(A)$; and
4. $\mathbb{P}(\emptyset) = 0$.

These claims are not hard to prove. Regarding the part 1, if $A \subset B$, then we have $B = (B \setminus A) \cup A$. (Sketch the Venn diagram.) Since $B \setminus A$ and A are disjoint, additivity of $\mathbb{P}$ now gives

$$\mathbb{P}(B) = \mathbb{P}(B \setminus A) + \mathbb{P}(A) \qquad \text{(whenever } A \subset B\text{)}$$

This equality implies parts 1–4 of fact 2.1.1. Rearranging gives part 1, while nonnegativity of $\mathbb{P}$ gives part 2. Specializing to $B = \Omega$ gives part 3, and setting $B = A$ gives part 4.

The property that $A \subset B$ implies $\mathbb{P}(A) \leq \mathbb{P}(B)$ is called **monotonicity**, and is fundamental. If $A \subset B$, then we know that B occurs whenever A occurs (because if ω lands in A, then it also lands in B). Hence, the probability of B should be larger. Many crucial ideas in probability boil down to this one point.

Fact 2.1.2. If A and B are any (not necessarily disjoint) events, then

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$$

It follows that for any $A, B \in \mathcal{F}$, we have $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$. This is called **subadditivity**. Thus, probabilities are subadditive over arbitrary pairs of events and additive over disjoint pairs.

2.1.3 Dependence and Independence

If A and B are events, then the **conditional probability of A given B** is

$$\mathbb{P}(A \mid B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \tag{2.2}$$

It represents the probability that A will occur, given the information that B has occurred. For the definition to make sense, it requires that $\mathbb{P}(B) > 0$. Events A and B are called **independent** if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B) \tag{2.3}$$

If A and B are independent, then the conditional probability of A given B is just the probability of A.

Example 2.1.5. Consider an experiment where we roll a dice twice. A suitable sample space is the set of pairs (i, j), where i and j are between 1 and 6. The first element i represents the outcome of the first roll, while the second element j represents the outcome of the second roll. Formally,

$$\Omega := \{(i, j) : i, j \in \{1, \ldots, 6\}\}$$

For our probability, let's define $\mathbb{P}(E) := \#E/36$. (Here elements of E are pairs, so $\#E$ is the number of pairs in E.) Now consider the events

$$A := \{(i, j) \in \Omega : i \text{ is even}\} \quad \text{and} \quad B := \{(i, j) \in \Omega : j \text{ is even}\}$$

In this case we have

$$A \cap B = \{(i, j) \in \Omega : i \text{ and } j \text{ are even}\}$$

We can establish (2.3), indicating independence of A and B under $\mathbb{P}$. To check this we need to be able to count the number of elements in A, B and $A \cap B$. The basic principle for counting ordered tuples is that the total number of possible tuples is the product of the number of possibilities for each element. For example, the number of distinct tuples

$$(i, j, k) \text{ where } i \in I, j \in J \text{ and } k \in K$$

is $(\#I) \times (\#J) \times (\#K)$. Hence, the number of elements in A is $3 \times 6 = 18$, the number of elements in B is $6 \times 3 = 18$, and the number of elements in $A \cap B$ is $3 \times 3 = 9$. As a result,

$$\mathbb{P}(A \cap B) = 9/36 = 1/4 = (18/36) \times (18/36) = \mathbb{P}(A)\mathbb{P}(B)$$

Thus, A and B are independent, as claimed.

A very useful result is the **law of total probability**, which says the following: Let $A \in \mathcal{F}$ and let $B_1, \ldots, B_M$ be a partition of Ω, so that $B_m \in \mathcal{F}$ for each m, $B_j \cap B_k = \varnothing$ when $j \neq k$, and $\cup_{m=1}^{M} B_m = \Omega$. If $\mathbb{P}(B_m) > 0$ for all m, then

$$\mathbb{P}(A) = \sum_{m=1}^{M} \mathbb{P}(A \mid B_m) \cdot \mathbb{P}(B_m)$$

The proof is straightforward, although you should check that the manipulations of intersections and unions if you have not seen them before:

$$\begin{aligned} \mathbb{P}(A) = \mathbb{P}[A \cap (\cup_{m=1}^{M} B_m)] &= \mathbb{P}[\cup_{m=1}^{M} (A \cap B_m)] \\ &= \sum_{m=1}^{M} \mathbb{P}(A \cap B_m) = \sum_{m=1}^{M} \mathbb{P}(A \mid B_m) \cdot \mathbb{P}(B_m) \end{aligned}$$

Example 2.1.6. Suppose we flip a coin to decide whether to take part in a poker game. If we play the chance of losing money is $2/3$. The overall chance of losing money (LM) that evening is

$$\mathbb{P}\{\text{LM }\} = \mathbb{P}\{\text{LM } | \text{play}\}\mathbb{P}\{\text{play}\} + \mathbb{P}\{\text{LM } | \text{don't play}\}\mathbb{P}\{\text{don't play}\}$$

which is $(2/3) \times (1/2) + 0 \times (1/2) = 1/3$.

2.1.4 Technical Details

As alluded to above, in the presentation of probability spaces I've swept some technical details under the carpet to make the presentation smooth. These details won't affect anything that follows, and this whole course can be completed successfully without knowing anything about them. Hence you can skip this section on first reading. Nevertheless, if you intend to keep going deeper into probability and statistics, eventually you will have to work your way through them. So let's note them as points for future study.

First, it turns out that assigning probabilities to all subsets of an arbitrary set Ω in a consistent way can be quite a difficult task. The reason is that if Ω is relatively large—a continuum, say—then it contains an awful lot of subsets, and some of them can be manipulated to exhibit very strange phenomena. (Look up the Banach-Tarski paradox for a hint of what I mean.) Because of this we usually take our set of events $\mathcal{F}$ to be a "well behaved" *sub*set of the subsets of Ω, and only assign probabilities to elements of $\mathcal{F}$.

How to choose $\mathcal{F}$? As stated above, we don't just choose freely because doing so will make it hard to form a consistent theory. Instead, the usual method is to start with a collection of sets that are reasonable and well behaved, and then permit extention to other sets than can be obtained from the original sets by some standard set operations.

For example, let's suppose that $A \in \mathcal{F}$, so that $\mathbb{P}(A)$ is well defined, and represents the "probability of event A." Now, given that we can assign a probability to the event A, it would be a bit unfortunate if we couldn't assign a probability to the event "not A", which corresponds to A^c. So normally we require that if $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$. When this is true, we say that $\mathcal{F}$ is "closed under the taking of complements".

Also, let's suppose that A and B are both in $\mathcal{F}$, so we assign probabilities to these events. In this case, it would be natural to think about the probability of the event

"A and B", which corresponds to $A \cap B$. So we also require that if A and B are in $\mathcal{F}$, then $A \cap B$ is also in $\mathcal{F}$. We say that $\mathcal{F}$ is "closed under the taking of intersections."

Perhaps we should also require that if A and B are in $\mathcal{F}$, then $A \cup B$ is also in $\mathcal{F}$? Actually, we don't have to, because (see fact 4.1.1 on page 145),

$$A \cup B = (A^c \cap B^c)^c$$

Thus, if $\mathcal{F}$ is closed under the taking of complements and intersections, then $\mathcal{F}$ is automatically closed under the taking of unions.

There is one more restriction that's typically placed on $\mathcal{F}$, which is the property of being closed under "countable" unions. This just means that if $A_1, A_2, \ldots$ is a sequence of sets in $\mathcal{F}$, then its union is likewise in $\mathcal{F}$. Since the details don't matter here we won't discuss it further. When $\mathcal{F}$ satisfies all these properties and contains the whole space Ω it is called a **σ-algebra.** Almost all of advanced probability rests on event classes that form σ-algebras.

Also, in standard probability theory there is a restriction placed on the probability $\mathbb{P}$ that has not yet been mentioned, called **countable additivity**. The definition of countable additivity is that if $A_1, A_2, \ldots$ is a disjoint sequence of sets in $\mathcal{F}$ (disjoint means that $A_i \cap A_j = \emptyset$ for any $i \neq j$), then

$$\mathbb{P}(\cup_i A_i) := \mathbb{P}\{\omega \in \Omega : \omega \in A_i \text{ for some } i\} = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$$

Why strengthen additivity to countable additivity? Countable additivity works behind the scenes to make probability theory run smoothly (expectations operators are suitably continuous, and so on). None of these details will concern us in this course.

If you wish, you can learn all about σ-algebras and countable additivity in any text on measure theory. There are many beautiful books on this subject. One of my favorites at the introductory level is Williams (1991).

2.2 Random Variables

[Roadmap]

2.2.1 Definition and Notation

In elementary probability courses, a random variable is defined as a "value that changes randomly," or something to that effect. This isn't a very satisfying definition. The proper definition of a **random variable** is a function from a sample space Ω into $\mathbb{R}$. Think of it this way:

1. "Nature" picks out an element ω in Ω according to some probability.
2. The random variable now sends this ω into numerical value $x(\omega)$.

Thus random variables convert outcomes in sample space—which can be any kind of object—*into numerical outcomes.* This is valuable because numerical outcomes are easy to order, add, subtract, etc. In other words, random variables "report" the outcome of an experiment in a format amenable to further analysis.[3]

Example 2.2.1. Consider the sample space

$$\Omega := \{(b_1, b_2, \ldots) : b_n \in \{0, 1\} \text{ for each } n\}$$

Ω is called the set of all infinite binary sequences. (This is an infinite version of the sample space in example 2.1.4. Imagine a computer with an infinite amount of memory.) Consider an experiment where we flip a coin until we get a "heads". We let 0 represent tails and 1 represent heads. The experiment of flipping until we get a heads can be modeled with the random variable

$$x(\omega) = x(b_1, b_2, \ldots) = \min\{n \in \mathbb{N} : b_n = 1\}$$

The number of heads in the first 10 flips is given by the random variable

$$y(\omega) = y(b_1, b_2, \ldots) = \sum_{n=1}^{10} b_n$$

As per the definition, x and y are well-defined functions from Ω into $\mathbb{R}$.[4]

[3] I'm skipping some technical details again. If Ω is a continuum, then, when identifying random variables with the class of all functions from Ω to $\mathbb{R}$, we typically exclude some particularly complicated functions. The remaining "nice" functions are called the **measurable functions**. These are our random variables. In this course we will never meet the nasty functions, and there's no need to go into further details. Those who want to know more should consult any text on measure theory (e.g., Williams, 1991).

[4] Is this true? What if $\omega = \omega_0$ is an infinite sequence containing only zeros? Then $\{n \in \mathbb{N} : b_n = 1\} = \emptyset$. So what is $x(\omega_0)$. The convention here is to set $\min \emptyset = \infty$. This is reasonable, but now x is not a map into $\mathbb{R}$, because it can take the value ∞. However, in most applications this event has probability zero, and hence we can ignore it. For example, we can set $x(\omega_0) = 0$ without changing anything significant. Now we're back to a well-defined function from Ω to $\mathbb{R}$.

Before going further, let's discuss a common notational convention with random variables that we've adopted above and that will be used below. With a random variable x, you will often see notation such as

$$\{x \text{ has some property}\}$$

This is a shorthand for the *event*

$$\{\omega \in \Omega : x(\omega) \text{ has some property}\}$$

We'll follow this convention, but you should translated it backwards and forwards in your mind to start off with. To give an example, consider the claim that, for any random variable x,

$$\mathbb{P}\{x \leq a\} \leq \mathbb{P}\{x \leq b\} \quad \text{whenever} \quad a \leq b \tag{2.4}$$

This is intuitively obvious. The mathematical argument goes as follows: Pick any $a, b \in \mathbb{R}$ with $a \leq b$. We have

$$\{x \leq a\} := \{\omega \in \Omega : x(\omega) \leq a\} \subset \{\omega \in \Omega : x(\omega) \leq b\} =: \{x \leq b\}$$

where the inclusion $\subset$ holds because if ω is such that $x(\omega) \leq a$, then, since $a \leq b$, we also have $x(\omega) \leq b$. (Any ω in the left-hand side is also in the right-hand side.) The result in (2.4) now follows from monotonicity of $\mathbb{P}$ (fact 2.1.1 on page 49).

Example 2.2.2. Recall example 2.1.4, with sample space

$$\Omega := \{(b_1, \ldots, b_N) : b_n \in \{0, 1\} \text{ for each } n\}$$

The set of events and probability were defined as $\mathcal{F} :=$ all subsets of Ω and $\mathbb{P}(A) := 2^{-N}(\#A)$. Consider a random variable x on Ω that returns the first element of any given sequence. That is,

$$x(\omega) = x(b_1, \ldots, b_N) = b_1$$

The probability that $x = 1$ is $1/2$. Indeed,

$$\mathbb{P}\{x = 1\} := \mathbb{P}\{\omega \in \Omega : x(\omega) = 1\} = \mathbb{P}\{(b_1, \ldots, b_N) : b_1 = 1\}$$

The number of length N binary sequences with $b_1 = 1$ is 2^{N-1}, so $\mathbb{P}\{x = 1\} = 1/2$.

2.2.2 Finite Random Variables

Aside from trivial random variables that take one value no matter what happens (and hence provide no information), the simplest kind of random variables are **binary** or **Bernoulli** random variables. A binary random variable is a random variable that takes values in $\{0, 1\}$. While binary random variables are by themselves rather limited, in fact they play a role analogous to basis vectors in $\mathbb{R}^N$. In particular, a great variety of random variables can be constructed as linear combinations of binary random variables. Things like expectation of a random variable can then be defined in terms of these primitive components. This section explores these ideas.

There is a generic way to create binary random variables, using **indicator functions**. If Q is a statement, such as "on the planet Uranus, there exists a tribe of three-headed monkeys," then $\mathbb{1}\{Q\}$ is considered as equal to 1 when the statement Q is true, and zero when the statement Q is false. Hence, $\mathbb{1}\{Q\}$ is a binary indicator of the truth of the statement Q. Another common variation on the notation is, for arbitrary $C \subset \Omega$,

$$\mathbb{1}_C(\omega) :=: \mathbb{1}\{\omega \in C\} := \begin{cases} 1 & \text{if } \omega \in C \\ 0 & \text{otherwise} \end{cases}$$

Note that $\mathbb{1}_C$ is a binary random variable. In fact, when you think about it, *any* binary random variable has the form

$$x(\omega) = \mathbb{1}_C(\omega) :=: \mathbb{1}\{\omega \in C\} \tag{2.5}$$

where C is some subset of Ω. Indeed, if x is any binary random variable, we can write x in the form of (2.5) by setting $C := \{\omega \in \Omega : x(\omega) = 1\}$.

From binary random variables we can create finite random variables. A **finite random variable** is a random variable with finite range (see §4.2 for the definition of range). We can create finite random variables by taking linear combinations of binary random variables. For example, let A and B be disjoint subsets of Ω. The random variable x defined by

$$x(\omega) = s\mathbb{1}_A(\omega) + t\mathbb{1}_B(\omega) \tag{2.6}$$

is a finite random variable taking the value s when ω falls in A, t when ω falls in B, and zero otherwise. Figure 2.1 shows a graph of x when $\Omega = \mathbb{R}$.

In fact *any* finite random variable can be expressed as the linear combinations of binary random variables. Thus we take as our generic expression for a finite random

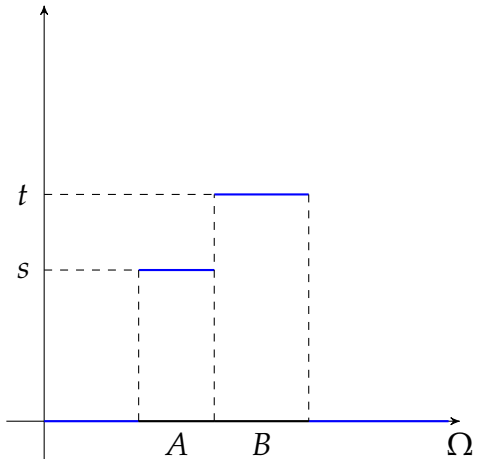


Figure 2.1: Finite random variable $x(\omega) = s\mathbb{1}_A(\omega) + t\mathbb{1}_B(\omega)$

variable the expression

$$x(\omega) = \sum_{j=1}^{J} s_j \mathbb{1}_{A_j}(\omega) \tag{2.7}$$

Whenever we work with this expression we will always assume that

- the scalars $s_1, \ldots, s_J$ are distinct, and
- the sets $A_1, \ldots, A_J$ form a partition of Ω.

By a **partition** we mean that $A_i \cap A_j = \emptyset$ when $i \neq j$, and $\cup_j A_j = \Omega$.

Fact 2.2.1. If x is the random variable in (2.7), then, for all j,

1. $x(\omega) = s_j$ if and only if $\omega \in A_j$.
2. $\{x = s_j\} = A_j$.
3. $\mathbb{P}\{x = s_j\} = \mathbb{P}(A_j)$.

Convince yourself of these results before continuing. (The second two statements follow from the first.)

2.2.3 Expectations

Our next task is to define expectations of random variables and state their basic properties. The idea behind expectations is that they give a kind of weighted average value, defined as the sum of all possible values of of the variable, weighted by their probabilities. Let's start with the finite case, where things are clearest. Letting x be as defined in (2.7), the **expectation** of x is

$$\mathbb{E}[x] := \sum_{j=1}^{J} s_j \mathbb{P}(A_j) \tag{2.8}$$

To understand this expression, recall from fact 2.2.1 that $\mathbb{P}(A_j)$ is also $\mathbb{P}\{x = s_j\}$, so we can also write

$$\mathbb{E}[x] = \sum_{j=1}^{J} s_j \mathbb{P}\{x = s_j\}$$

As required, the expectation is the sum of the different values that x may take, weighted by their probabilities.

We can already make some important observations.

Fact 2.2.2. $\mathbb{E}[\mathbb{1}_A] = \mathbb{P}(A)$ for any $A \in \mathcal{F}$.

Fact 2.2.2 seems trivial but in fact it represents *the* fundamental link between probabilities and expectations. To see why it holds, observe that $\mathbb{1}_A(\omega) = 1 \times \mathbb{1}_A(\omega) + 0 \times \mathbb{1}_{A^c}(\omega)$. Applying (2.8), we get $\mathbb{E}[\mathbb{1}_A] = 1 \times \mathbb{P}(A) + 0 \times \mathbb{P}(A^c) = \mathbb{P}(A)$.

Fact 2.2.3. If $\alpha \in \mathbb{R}$, then $\mathbb{E}[\alpha] = \alpha$.

The right way to understand this is to take α to be the constant random variable $\alpha \mathbb{1}_\Omega$. From (2.8), the expectation of this constant is $\mathbb{E}[\alpha] := \mathbb{E}[\alpha \mathbb{1}_\Omega] = \alpha \mathbb{P}(\Omega) = \alpha$.

How about expectation for arbitrary random variables, such as those with infinite range? The same mode of definition doesn't work but this presents no major problem because any random variable can be well approximated by finite random variables. A finite approximation x_n to an arbitrary random variable x is shown in figure 2.2. This approximation can be improved without limit if we allow the finite approximation to take a larger and larger number of distinct values.

Hence we can take a sequence $\{x_n\}$ of finite random variables converging to any selected arbitrary random variable x. The expectation of x is then defined as

$$\mathbb{E}[x] := \lim_{n \to \infty} \mathbb{E}[x_n] \tag{2.9}$$

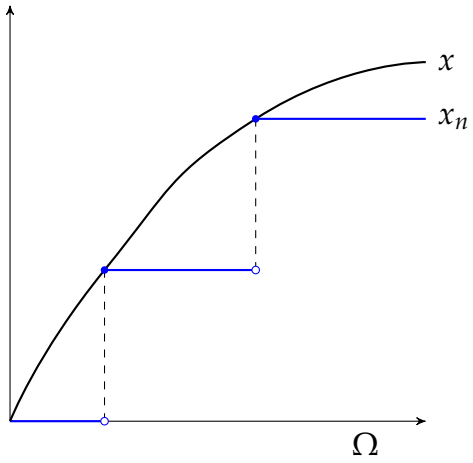


Figure 2.2: Finite approximation to a general random variable

To complete the definition we need to make sure that this limit makes sense. There are a couple of issues that arise here. Suppose first that x is nonnegative (i.e., $x(\omega) \geq 0$ for all $\omega \in \Omega$). It turns out that the sequence $\{x_n\}$ approximating x can then be chosen so that $\mathbb{E}[x_n]$ is monotone increasing, and therefore must converge to something. While that something can be $+\infty$, this causes no real problem as we just say that $\mathbb{E}[x] = \infty$.

If x is not nonnegative then some thought will convince you that we can write it as the sum $x = x^+ - x^-$, where $x^+(\omega) := \max\{x(\omega), 0\}$ and $x^-(\omega) := -\min\{x(\omega), 0\}$. Both x^+ and x^- are nonnegative random variables. The expectation can then be defined as $\mathbb{E}[x] = \mathbb{E}[x^+] - \mathbb{E}[x^-]$. The only issue is that this expression might have the form $\infty - \infty$, which is not allowed. Hence for random variables that are not necessarily nonnegative, we usually restrict attention to integrable random variables. An **integrable** random variable is a random variable x such that $\mathbb{E}[|x|] < \infty$. Since $x^+ \leq |x|$ and $x^- \leq |x|$ this is enough to ensure that both $\mathbb{E}[x^+] < \infty$ and $\mathbb{E}[x^-] < \infty$. Hence $\mathbb{E}[x] = \mathbb{E}[x^+] - \mathbb{E}[x^-]$ is well defined.

To be careful we should also check that the value in (2.9) does not depend on the particular approximating sequence $\{x_n\}$ that we choose. This and other related technical details can be found in a text such as Williams (1991). In that reference the following facts are also established:

Fact 2.2.4. If x and y are integrable random variables and α and β are any constants,

then

$$\mathbb{E}[\alpha x + \beta y] = \alpha \mathbb{E}[x] + \beta \mathbb{E}[y]$$

This is called linearity of expectations. Here $\alpha x + \beta y$ should be understood as the random variable $(\alpha x + \beta y)(\omega) := \alpha x(\omega) + \beta y(\omega)$. Exercises 2.8.10 and 2.8.11 ask you to check fact 2.2.4 in some special cases.

Fact 2.2.5. If x, y are integrable random variables and $x \leq y$, then $\mathbb{E}[x] \leq \mathbb{E}[y]$.

This is called monotonicity of expectations. The statement $x \leq y$ means that x is less than y for any realization of uncertainty. Formally, $x(\omega) \leq y(\omega)$ for all $\omega \in \Omega$. Exercise 2.8.20 asks you to prove fact 2.2.5 in one special case.

Let x be a random variable and let $k \in \mathbb{N}$. If $\mathbb{E}[|x|^k] < \infty$ then the **k-th moment** of x is said to exist, and is defined as the value $\mathbb{E}[x^k]$. For some random variables even the first moment does not exist. For others every moment exists.

If x is a random variable with finite second moment, then the **variance** of x is

$$\text{var}[x] := \mathbb{E}[(x - \mathbb{E}[x])^2]$$

This gives a measure of the dispersion of x. The **standard deviation** of x is often written as σ_x and defined as

$$\sigma_x := \sqrt{\text{var}[x]}$$

Fact 2.2.6. If the k-th moment of x exists, then so does the j-th moment for all $j \leq k$.

Fact 2.2.7. If x and y are random variables with finite second moment, then

$$|\mathbb{E}[xy]| \leq \sqrt{\mathbb{E}[x^2]\mathbb{E}[y^2]} \tag{2.10}$$

Fact 2.2.8. For any nonnegative random variable x any $\delta > 0$, we have

$$\mathbb{P}\{x \geq \delta\} \leq \frac{\mathbb{E}[x]}{\delta} \tag{2.11}$$

Fact 2.2.7 is called the **Cauchy-Schwarz inequality for random variables**, while (2.11) is called **Chebyshev's inequality**. A common variation of Chebyshev's inequality is the bound

$$\mathbb{P}\{|x| \geq \delta\} \leq \frac{\mathbb{E}[x^2]}{\delta^2} \tag{2.12}$$

Exercise 2.8.36 asks you to check (2.11) and (2.12).

In concluding this section, let us agree that to avoid repetition, we will assume that every random variable introduced below is integrable unless stated otherwise. Also, only random variables with finite second moment have a well defined variance, but in what follows we will often talk about the variance of a given random variable without adding the caveat "assuming it exists."

2.3 Distributions

Distributions summarize the probabilities of different outcomes for random variables and help us compute expectations. In this section, we describe the link between random variables and their distributions.

2.3.1 Distribution Functions

Let x be a random variable on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and consider the function F defined by

$$F(s) := \mathbb{P}\{x \leq s\} := \mathbb{P}\{\omega \in \Omega : x(\omega) \leq s\} \qquad (s \in \mathbb{R}) \tag{2.13}$$

It can be shown that, for any choice of random variable x, this function always has the following properties:

1. right-continuity: $F(s_n) \downarrow F(s)$ whenever $s_n \downarrow s$
2. monotonicity: $s \leq s'$ implies $F(s) \leq F(s')$
3. $\lim_{s \to -\infty} F(s) = 0$ and $\lim_{s \to \infty} F(s) = 1$

For example, monotonicity is immediate from (2.4) on page 55. (The other properties are a bit trickier to prove. See, e.g, Williams, 1991.)

Any function $F\colon \mathbb{R} \to [0,1]$ satisfying conditions 1–3 is called a **cumulative distribution function** or **cdf** on $\mathbb{R}$. Thus, to each random variable, we can associate a unique cdf on $\mathbb{R}$. We say that F **is the cdf of** x, or, alternatively, that F **is the distribution of** x, and write $x \sim F$.

Example 2.3.1. The function $F(s) = \arctan(s)/\pi + 1/2$ is a cdf—one variant of the Cauchy distribution. A plot is given in figure 2.3.

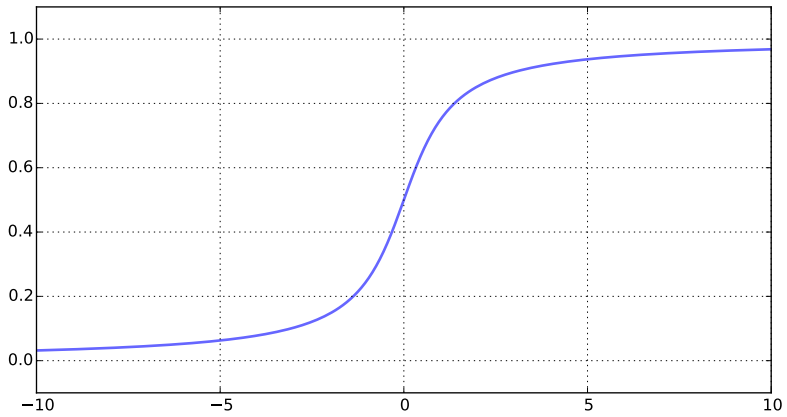


Figure 2.3: Cauchy cdf

It's worth nothing that the following is also true: For every cdf F, there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a random variable $x\colon \Omega \to \mathbb{R}$ such that the distribution of x is F. Exercise 2.8.16 gives some hints on how one construction works.

Fact 2.3.1. If $x \sim F$, then $\mathbb{P}\{a < x \leq b\} = F(b) - F(a)$ for any $a \leq b$.

Indeed, if $a \leq b$, then $\{a < x \leq b\} = \{x \leq b\} \setminus \{x \leq a\}$ and $\{x \leq a\} \subset \{x \leq b\}$. Applying fact 2.1.1 on page 49 gives fact 2.3.1.

A cdf F is called **symmetric** if $F(-s) = 1 - F(s)$ for all $s \in \mathbb{R}$.[5] The proof of the next fact is an exercise (exercise 2.8.14).

Fact 2.3.2. Let F be a cdf and let $x \sim F$. If F is symmetric and $\mathbb{P}\{x = s\} = 0$ for all $s \in \mathbb{R}$, then the distribution $F_{|x|}$ of the absolute value $|x|$ is given by

$$F_{|x|}(s) := \mathbb{P}\{|x| \leq s\} = 2F(s) - 1 \qquad (s \geq 0)$$

One often needs to obtain the cdf of the transform of a random variable. This is easy in the monotone case. For example, if $x \sim F$ and $y = \exp(x)$, then the cdf of y is $G(s) := F(\ln(s))$, because

$$\mathbb{P}\{y \leq s\} = \mathbb{P}\{\exp(x) \leq s\} = \mathbb{P}\{x \leq \ln(s)\} = F(\ln(s)) =: G(s)$$

Note how monotonicity is used in the second equality.

[5] Thus, the probability that $x \leq -s$ is equal to the probability that $x > s$. Centered normal distributions and t-distributions have this property.

2.3.2 Densities and PMFs

Every random variable has a well-defined cdf via (2.13). However, as representations of distributions, cdfs have some disadvantages. For example, and plotting cdfs is a poor way to convey information about probabilities. The amount of probability mass in different regions is determined by the slope of the cdf. Research shows that humans are poor at extracting quantitative information from slopes. They do much better with *heights*, which leads us into our discussion of densities and probability mass functions (pmfs). Densities and pmfs correspond to two different, mutually exclusive cases. The density case arises when the increase of the cdf in question is smooth, and contains no jumps. The pmf case arises when the increase consists of jumps alone.

Let's have a look at these two situations, starting with the second. The pure jump case occurs when the cdf represents a discrete random variable. To understand this, suppose that x takes values $s_1, \ldots, s_J$. Let $p_j := \mathbb{P}\{x = s_j\}$. We then have $0 \leq p_j \leq 1$ for each j, and $\sum_{j=1}^J p_j = 1$ (exercise 2.8.15). A finite collection of numbers $p_1, \ldots, p_J$ such that $0 \leq p_j \leq 1$ and $p_1 + \cdots + p_J = 1$ is called a **probability mass function** (pmf). The cdf corresponding to this random variable is

$$F(s) := \sum_{j=1}^{J} \mathbb{1}\{s_j \leq s\} p_j \tag{2.14}$$

How do we arrive at this expression? Because, for this random variable,

$$\mathbb{P}\{x \leq s\} = \mathbb{P}\left\{ \bigcup_{j \text{ s.t. } s_j \leq s} \{x = s_j\} \right\} = \sum_{j \text{ s.t. } s_j \leq s} \mathbb{P}\{x = s_j\} = \sum_{j=1}^{J} \mathbb{1}\{s_j \leq s\} p_j$$

Visually, F is a step function, with a jump up of size p_j at point s_j. Figure 2.4 gives an example with $J = 2$. The cdf is right continuous but not continuous.

The other case of interest is the density case. A **density** is a nonnegative function p that integrates to 1. For example, suppose that F is a smooth cdf, so that the derivative F' exists. Let $p := F'$. By the fundamental theorem of calculus, we then have

$$\int_r^s p(t)dt = \int_r^s F'(t)dt = F(s) - F(r)$$

From this equality and the properties of cdfs, we can see that p is nonnegative and $\int_{-\infty}^{+\infty} p(s)ds = 1$. In other words, p is a density. Also, taking the limit as $r \to -\infty$ we

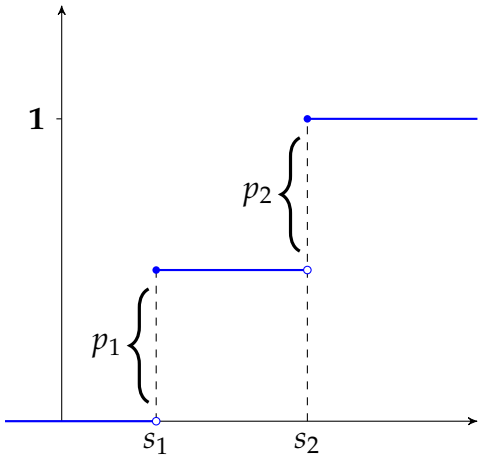


Figure 2.4: Discrete cdf

obtain

$$F(s) = \int_{-\infty}^{s} p(t)dt$$

which tells us that F can be recovered from p.

Not every random variable has a density. The exact necessary and sufficient condition for a density to exist is that F is "absolutely continuous." This is a smoothness condition, an important special case of which is differentiability.[6] If F is absolutely continuous and $p\colon \mathbb{R} \to [0, \infty)$ satisfies

$$F(s) = \int_{-\infty}^{s} p(t)dt \qquad \text{for all } s \in \mathbb{R}$$

then p is called the **density corresponding to** F.

Fact 2.3.3. If x has a density, then $\mathbb{P}\{x = s\} = 0$ for all $s \in \mathbb{R}$.

As discussed above, cdfs are useful because every random variable has one, but pmfs and densities are nicer to work with, and visually more informative. For example, consider figure 2.5, which shows the density corresponding to the Cauchy cdf in figure 2.3. Information about probability mass is now conveyed by height rather than slope, which is easier for us humans to digest.

[6] In elementary texts, random variables with densities are often called "continuous random variables." This terminology is confusing because "continuous" here has nothing to do with the usual definition of continuity of functions.

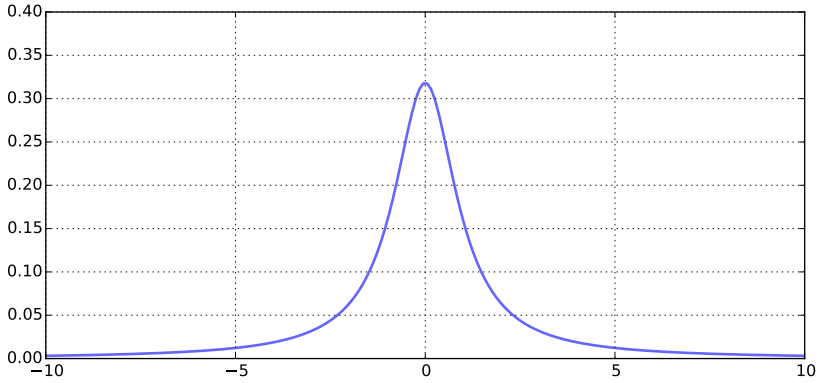


Figure 2.5: Cauchy density

2.3.3 The Quantile Function

Let F be any cdf on $\mathbb{R}$. Suppose that F is strictly increasing, so that the inverse function F^{-1} exists:

$$F^{-1}(q) := \text{ the unique } s \text{ such that } F(s) = q \qquad (0 < q < 1) \tag{2.15}$$

The inverse of the cdf is called the **quantile function**, and has many applications in probability and statistics. For example, the quantile function associated with the Cauchy cdf in example 2.3.1 is $F^{-1}(q) = \tan[\pi(q - 1/2)]$. See figure 2.6.

Things are a bit more complicated when F is not strictly increasing, as the inverse F^{-1} is not well defined. (If F is not strictly increasing, then there exists at least two distinct points s and s' such that $F(s) = F(s')$.) This problem is negotiated by setting

$$F^{-1}(q) := \inf\{s \in \mathbb{R} : F(s) \geq q\} \qquad (0 < q < 1)$$

This expression is a bit more complicated, but in the case where F is strictly increasing, it reduces to (2.15).

The value $F^{-1}(1/2)$ is called the **median** of F. It gives an alternative measure of central tendency (alternative to the mean).

The quantile function features in hypothesis testing, where it can be used to define critical values. An abstract version of the problem is as follows: Let $x \sim F$, where F is strictly increasing, differentiable (so that a density exists and x puts no probability mass on any one point) and symmetric. Given $\alpha \in (0, 1)$, we want to find the c

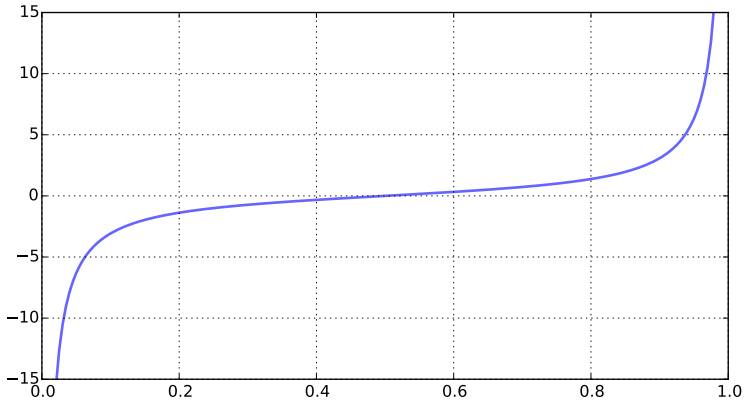


Figure 2.6: Cauchy quantile function

such that $\mathbb{P}\{-c \leq x \leq c\} = 1 - \alpha$ (see figure 2.7). The solution is given by $c := F^{-1}(1 - \alpha/2)$. That is,

$$c = F^{-1}(1 - \alpha/2) \implies \mathbb{P}\{|x| \leq c\} = 1 - \alpha \tag{2.16}$$

To see this, fix $\alpha \in (0, 1)$. From fact 2.3.2, we have

$$\mathbb{P}\{|x| \leq c\} = 2F(c) - 1 = 2F[F^{-1}(1 - \alpha/2)] - 1 = 1 - \alpha$$

In the case where F is the standard normal cdf Φ, this value c is usually denoted by $z_{\alpha/2}$. We will adopt the same notation:

$$z_{\alpha/2} := \Phi^{-1}(1 - \alpha/2) \tag{2.17}$$

2.3.4 Expectations from Distributions

Until now, we've been calculating expectations using the expectation operator $\mathbb{E}$, which was defined from a given probability $\mathbb{P}$ in §2.2.3. One of the most useful facts about distributions is that they encode all the information necessary to calculate $\mathbb{E}[x]$. For a full treatment of this topic you can consult a text such as Williams (1991). Here we'll stick to noting down the main facts. In all of what follows, h is an arbitrary function from $\mathbb{R}$ to $\mathbb{R}$.

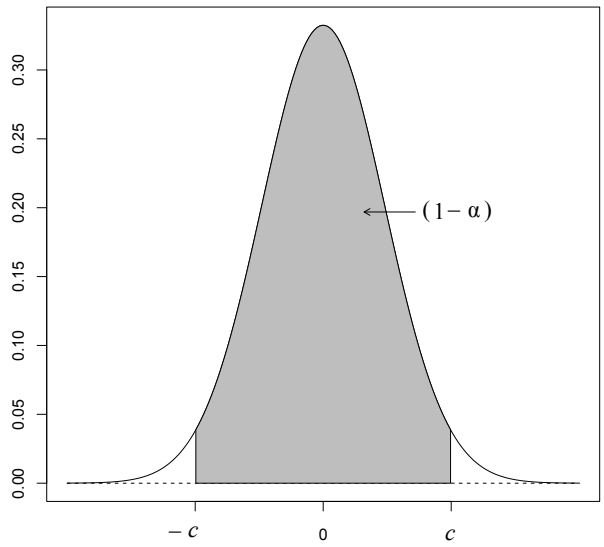


Figure 2.7: Finding critical values

Fact 2.3.4. If x is a discrete random variable taking values $s_1, \ldots, s_J$ with probabilities $p_1, \ldots, p_J$, then

$$\mathbb{E}\left[h(x)\right] = \sum_{j=1}^{J} h(s_j)p_j \tag{2.18}$$

On the other hand, if x has density p, then

$$\mathbb{E}\left[h(x)\right] = \int_{-\infty}^{\infty} h(s)p(s)ds \tag{2.19}$$

It's convenient to have a piece of notation that captures both of these cases. As a result, if $x \sim F$, then we will write

$$\mathbb{E}\left[h(x)\right] = \int h(s)F(ds)$$

The way you should understand this expression is that when F is differentiable with derivative $p = F'$, then $\int h(s)F(ds)$ is defined as $\int_{-\infty}^{\infty} h(s)p(s)ds$. If, on the other hand, F is the step function $F(s) = \sum_{j=1}^{J} \mathbb{1}\{s_j \leq s\}p_j$ corresponding to the discrete random variable in fact 2.3.4, then $\int h(s)F(ds)$ is defined as $\sum_{j=1}^{J} h(s_j)p_j$.

Just for the record, for a given cdf F, the expression $\int h(s)F(ds)$ has its own precise definition, as the "Lebesgue-Stieltjes" integral of h with respect to F. In the special cases where F is discrete or differentiable, one can prove that $\int h(s)F(ds)$ reduces to (2.18) or (2.19) respectively. For details see, e.g., Williams (1991).

As an exercise, let's prove the first part of fact 2.3.4. This is the discrete case, where $x = \sum_{j=1}^{J} s_j \mathbb{1}_{A_j}$ and $p_j := \mathbb{P}\{x = s_j\} = \mathbb{P}(A_j)$. As usual, the values $\{s_j\}$ are distinct and the sets $\{A_j\}$ are a partition of Ω. Let $h\colon \mathbb{R} \to \mathbb{R}$. You should be able to convince yourself that

$$h(x(\omega)) = \sum_{j=1}^{J} h(s_j)\mathbb{1}_{A_j}(\omega)$$

(Pick an arbitrary A_j and check that the left- and right-hand sides are equal when $\omega \in A_j$.) This is a discrete random variable, which we can take the expectation of using (2.8) on page 58. We get

$$\mathbb{E}\left[h(x)\right] = \sum_{j=1}^{J} h(s_j)\mathbb{P}(A_j) = \sum_{j=1}^{J} h(s_j)p_j$$

Equation (2.18) is confirmed.

2.3.5 Common Distributions

Let's list a few well-known distributions. First, given $a < b$, the **uniform distribution** on interval $[a, b]$ is the distribution associated with the density

$$p(s; a, b) := \frac{1}{b-a} \qquad (a \leq s \leq b)$$

(If $s < a$ or $s > b$, then $p(s; a, b) := 0$.) The mean is

$$\int_a^b s\, p(s; a, b)ds = \frac{a+b}{2}$$

The **univariate normal density** or **Gaussian density** is a function p of the form

$$p(s) := p(s; \mu, \sigma) := (2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{1}{2}(s-\mu)^2\sigma^{-2}\right\}$$

for some $\mu \in \mathbb{R}$ and $\sigma > 0$. We represent this distribution symbolically by $\mathcal{N}(\mu, \sigma^2)$. The distribution $\mathcal{N}(0, 1)$ is called the **standard normal distribution**

It is well-known that if $x \sim \mathcal{N}(\mu, \sigma^2)$, then $\mathbb{E}[x] = \mu$, and $\text{var}[x] = \sigma^2$. Hence the two parameters separately define the mean and the variance (or standard deviation), and this is one of many attractive features of the distribution. Here's another:

Fact 2.3.5. If $x_1, \ldots, x_N$ are normally distributed and $\alpha_0, \ldots, \alpha_N$ are any constants, then $\alpha_0 + \sum_{n=1}^{N} \alpha_n x_n$ is also normally distributed.

In other words, linear combinations of normals are normal. This is a fact, telling us that linear models and normal distributions play very well together.[7]

The **chi-squared distribution with k degrees of freedom** is the distribution with density

$$p(s;k) := \frac{1}{2^{k/2}\Gamma(k/2)} s^{k/2-1} e^{-s/2} \qquad (s \geq 0)$$

where Γ is the Gamma function (details omitted). If x has a distribution described by this density, then we write $x \sim \chi^2(k)$.

Student's t-distribution with k degrees of freedom, or, more simply, the t-distribution with k degrees of freedom, is the distribution on $\mathbb{R}$ with density

$$p(s;k) := \frac{\Gamma(\frac{k+1}{2})}{(k\pi)^{1/2}\Gamma(\frac{k}{2})} \left(1 + \frac{s^2}{k}\right)^{-(k+1)/2}$$

The **F-distribution** with parameters k_1, k_2 is the distribution with the unlikely looking density

$$p(s;k_1,k_2) := \frac{\sqrt{(k_1 s)^{k_1} k_2^{k_2} / [k_1 s + k_2^{k_1+k_2}]}}{sB(k_1/2, k_2/2)} \qquad (s \geq 0)$$

where B is the Beta function (details omitted). The F-distribution arises in certain hypothesis tests, some of which we will examine later.

2.4 Joint Distributions and Independence

[roadmap]

[7] We should be a bit careful here—what if $\alpha_n = 0$ for all n? To save ourselves from embarrassment we can declare a random variable concentrated on a point to be a normal random variable with "zero variance".

2.4.1 Distributions Across Random Variables

Consider a collection of N random variables $x_1, \ldots, x_N$. For each individual random variable $x_n : \Omega \to \mathbb{R}$, the distribution F_n of x_n is

$$F_n(s) := \mathbb{P}\{x_n \leq s\} \qquad (-\infty < s < \infty) \tag{2.20}$$

This distribution tells us about the random properties of x_n viewed as a single entity. But we often want to know about the relationships between the variables $x_1, \ldots, x_N$, and outcomes for the group of variables as a whole. To quantify these things, we define the **joint distribution** of $x_1, \ldots, x_N$ to be

$$F(s_1, \ldots, s_N) := \mathbb{P}\{x_1 \leq s_1, \ldots, x_N \leq s_N\} \qquad (-\infty < s_n < \infty \,; n = 1, \ldots, N)$$

In this setting, the distribution F_n of x_n is sometimes called the **marginal distribution**, in order to distinguish it from the joint distribution.

The **joint density** of $x_1, \ldots, x_N$, if it exists, is a function $p : \mathbb{R}^N \to [0, \infty)$ satisfying

$$\int_{-\infty}^{t_N} \cdots \int_{-\infty}^{t_1} p(s_1, \ldots, s_N) ds_1 \cdots ds_N = F(t_1, \ldots, t_N) \tag{2.21}$$

for all $t_n \in \mathbb{R}$, $n = 1, \ldots, N$.

Typically, the joint distribution cannot be determined from the N marginal distributions alone, since the marginals do not tell us about the interactions between the different variables. Once special case where we can tell the joint from the marginals is when there is no interaction. This is called independence, and we treat it in the next section.

From joint densities we can construct conditional densities. The **conditional density** of $x_{k+1}, \ldots, x_N$ given $x_1 = s_1, \ldots, x_k = s_k$ is defined by

$$p(s_{k+1}, \ldots, s_N \,|\, s_1, \ldots, s_k) := \frac{p(s_1, \ldots, s_N)}{p(s_1, \ldots, s_k)} \tag{2.22}$$

Rearranging this expression we obtain a decomposition of the joint density:

$$p(s_1, \ldots, s_N) = p(s_{k+1}, \ldots, s_N \,|\, s_1, \ldots, s_k) p(s_1, \ldots, s_k) \tag{2.23}$$

This decomposition is useful in many situations.

2.4.2 Independence

Let $x_1, \ldots, x_N$ be a collection of random variables with $x_n \sim F_n$, where F_n is a cdf. The variables $x_1, \ldots, x_N$ are called **identically distributed** $F_n = F_m$ for all n, m. They are called **independent** if, given any $s_1, \ldots, s_N$, we have

$$\mathbb{P}\{x_1 \leq s_1, \ldots, x_N \leq s_N\} = \mathbb{P}\{x_1 \leq s_1\} \times \cdots \times \mathbb{P}\{x_N \leq s_N\} \tag{2.24}$$

Equivalently, if F is the joint distribution of $x_1, \ldots, x_N$ and F_n is the marginal distribution of x_n, then independence states that

$$F(s_1, \ldots, s_N) = F_1(s_1) \times \cdots \times F_N(s_N) = \prod_{n=1}^{N} F_n(s_n)$$

We use the abbreviation IID for collections of random variables that are both independent and identically distributed.

Example 2.4.1. Consider a monkey throwing darts at a dartboard. Let x denote the horizontal location of the dart relative to the center of the board, and let y denote the vertical location. (For example, if $x = -1$ and $y = 3$, then the dart is 1cm to the left of the center, and 3cm above.) At first pass, we might suppose that x and y are independent and identically distributed.

Fact 2.4.1. If $x_1, \ldots, x_M$ are independent and $\mathbb{E}|x_m|$ is finite for each m, then

$$\mathbb{E}\left[\prod_{m=1}^{M} x_m\right] = \prod_{m=1}^{M} \mathbb{E}[x_m]$$

We won't prove the last fact in the general case, as this involves measure theory. However, we can illustrate the idea by showing that $\mathbb{E}[xy] = \mathbb{E}[x]\mathbb{E}[y]$ when x and y are independent and defined by (2.40). In this case, it can be shown (details omitted) that the random variables x and y are independent precisely when the events A and B are independent. Now observe that

$$(xy)(\omega) := x(\omega)y(\omega) = s\mathbb{1}\{\omega \in A\}t\mathbb{1}\{\omega \in B\} = st\mathbb{1}\{\omega \in A \cap B\}$$

Hence, by the definition of expectations, we have

$$\mathbb{E}[xy] = st\mathbb{P}(A \cap B) = st\mathbb{P}(A)\mathbb{P}(B) = s\mathbb{P}(A)t\mathbb{P}(B) = \mathbb{E}[x]\mathbb{E}[y]$$

Fact 2.4.2. If x and y are independent and g and f are any functions, then $f(x)$ and $g(y)$ are independent.

An important special case of the "independence means multiply" rule is as follows.

Fact 2.4.3. If random variables $x_1, \ldots, x_N$ are independent, and each has density p_n, then the joint density p exists, and is the product of the marginal densities:

$$p(s_1, \ldots, s_N) = \prod_{n=1}^{N} p_n(s_n)$$

Here are some useful facts relating independence and certain common distributions.

Fact 2.4.4. If $x_1, \ldots, x_k \overset{\text{IID}}{\sim} \mathcal{N}(0,1)$, then

$$Q := \sum_{i=1}^{k} x_i^2 \sim \chi^2(k)$$

Fact 2.4.5. If $Q_1, \ldots, Q_J$ are independent with $Q_j \sim \chi^2(k_j)$, then $\sum_{j=1}^{J} Q_j \sim \chi^2(\sum_j k_j)$.

Fact 2.4.6. If Z and Q are two random variables such that

1. $Z \sim \mathcal{N}(0,1)$,
2. $Q \sim \chi^2(k)$, and
3. Z and Q are independent,

then $Z(k/Q)^{1/2}$ has the t-distribution with k degrees of freedom.

Fact 2.4.7. If $Q_1 \sim \chi^2(k_1)$ and $Q_2 \sim \chi^2(k_2)$ are independent, then

$$\frac{Q_1/k_1}{Q_2/k_2}$$

is distributed as $F(k_1, k_2)$.

2.4.3 Covariance

The **covariance** of random variables x and y is defined as

$$\text{cov}[x, y] := \mathbb{E}\left[(x - \mathbb{E}[x])(y - \mathbb{E}[y])\right]$$

Fact 2.4.8. If $x_1, \dots, x_N$ are random variables and $\alpha_1, \dots, \alpha_N$ are constant scalars, then

$$\operatorname{var}\left[\sum_{n=1}^{N} \alpha_n x_n\right] = \sum_{n=1}^{N} \alpha_n^2 \operatorname{var}[x_n] + 2 \sum_{n<m} \alpha_n \alpha_m \operatorname{cov}[x_n, x_m]$$

In particular, if α and β are real numbers and x and y are random variables, then $\operatorname{var}[\alpha] = 0$,[8] $\operatorname{var}[\alpha + \beta x] = \beta^2 \operatorname{var}[x]$, and

$$\operatorname{var}[\alpha x + \beta y] = \alpha^2 \operatorname{var}[x] + \beta^2 \operatorname{var}[y] + 2\alpha\beta \operatorname{cov}[x, y]$$

Given two random variables x and y with finite variances σ_x^2 and σ_y^2 respectively, the **correlation** of x and y is defined as

$$\operatorname{corr}[x, y] := \frac{\operatorname{cov}[x, y]}{\sigma_x \, \sigma_y}$$

If $\operatorname{corr}[x, y] = 0$, we say that x and y are **uncorrelated**. For this to occur, it is necessary and sufficient that $\operatorname{cov}[x, y] = 0$. Positive correlation means that $\operatorname{corr}[x, y]$ is positive, while negative correlation means that $\operatorname{corr}[x, y]$ is negative. The first part of the next fact follows immediately from fact 2.2.7, while the second is just algebra.

Fact 2.4.9. Given any two random variables x, y and positive constants α, β, we have

$$-1 \leq \operatorname{corr}[x, y] \leq 1 \quad \text{and} \quad \operatorname{corr}[\alpha x, \beta y] = \operatorname{corr}[x, y]$$

Fact 2.4.10. If x and y are independent, then $\operatorname{cov}[x, y] = \operatorname{corr}[x, y] = 0$.

Note that the converse is not true: One can construct examples of dependent random variables with zero covariance.

2.4.4 Best Linear Predictors

Let's consider the problem of predicting the value of a random variable y given knowledge of the value of a second random variable x (and also knowledge of the underlying probability distributions, which makes this a problem in probability

[8] Here $\operatorname{var}[\alpha]$ should be understood as $\operatorname{var}[\alpha \mathbb{1}\{\omega \in \Omega\}]$, as was the case when we discussed fact 2.2.3 on page 58.

rather than statistics). Thus, we seek a function f such that $f(x)$ is close to y on average. To measure the "average distance" between $f(x)$ and y, we will use the mean squared deviation between $f(x)$ and y, which is

$$\mathbb{E}\left[(y - f(x))^2\right]$$

As we will learn in chapter 3, the minimizer of the mean squared deviation over all functions of x is obtained by choosing $f(x) = \mathbb{E}[y \mid x]$, where the right-hand size is the conditional expectation of y given x. However, the conditional expectation may be nonlinear and complicated, so let's now consider the simpler problem of finding a good predictor of y within a small and well-behaved class of functions. The class of functions we will consider is the set of "linear" functions

$$\mathcal{L} := \{ \text{ all functions of the form } \ell(x) = \alpha + \beta x\}$$

(While elementary courses refer to these functions as linear, in fact they are not linear unless $\alpha = 0$ (see §1.2.1). The class of functions $\mathcal{L}$ is more correctly known as the set of **affine** functions.) Thus, we consider the problem

$$\min_{\ell \in \mathcal{L}} \mathbb{E}\left[(y - \ell(x))^2\right] = \min_{\alpha,\beta \in \mathbb{R}} \mathbb{E}\left[(y - \alpha - \beta x)^2\right] \tag{2.25}$$

Expanding the square on the right-hand side and using linearity of $\mathbb{E}$, the objective function becomes

$$\psi(\alpha, \beta) := \mathbb{E}\left[y^2\right] - 2\alpha\mathbb{E}\left[y\right] - 2\beta\mathbb{E}\left[xy\right] + 2\alpha\beta\mathbb{E}\left[x\right] + \alpha^2 + \beta^2\mathbb{E}\left[x^2\right]$$

Computing the derivatives and solving the equations

$$\frac{\partial\psi(\alpha,\beta)}{\partial\alpha} = 0 \quad \text{and} \quad \frac{\partial\psi(\alpha,\beta)}{\partial\beta} = 0$$

We obtain (exercise 2.8.29) the minimizers

$$\beta^* := \frac{\text{cov}[x,y]}{\text{var}[x]} \quad \text{and} \quad \alpha^* := \mathbb{E}\left[y\right] - \beta^*\mathbb{E}\left[x\right] \tag{2.26}$$

The best linear predictor is therefore

$$\ell^*(x) := \alpha^* + \beta^* x$$

If you've studied elementary linear least squares regression before, you will realize that α^* and β^* are the "population" counterparts for the coefficient estimates in the regression setting. We'll talk more about the connections below.

2.5 Asymptotics

[Roadmap]

2.5.1 Modes of Convergence

Let $\{x_n\}_{n=1}^{\infty}$ be a sequence of random variables. We say that $\{x_n\}_{n=1}^{\infty}$ converges to random variable x **in probability** if

$$\text{for any } \delta > 0, \quad \mathbb{P}\{|x_n - x| > \delta\} \to 0 \quad \text{as} \quad n \to \infty$$

In symbols, this convergence is represented by $x_n \xrightarrow{p} x$. In almost all the applications we consider, the limit x will be a constant. The next example illustrates the definition for this case.

Example 2.5.1. If $x_n \sim \mathcal{N}(\alpha, 1/n)$, then $x_n \xrightarrow{p} \alpha$. That is, for any $\delta > 0$, we have $\mathbb{P}\{|x_n - \alpha| > \delta\} \to 0$. Fixing $\delta > 0$, the probability $\mathbb{P}\{|x_n - \alpha| > \delta\}$ is shown in figure 2.8 for two different values of n, where it corresponds to the size of the shaded areas. This probability collapses to zero as $n \to \infty$, decreasing the variance and causing the density to become more peaked.

A full proof of the convergence result in example 2.5.1 can be found by looking at the normal density and bounding tail probabilities. However, a much simpler proof can also be obtained by exploiting the connection between convergence in probability and convergence in mean squared error. The details are below.

Fact 2.5.1. Regarding convergence in probability, the following statements are true:

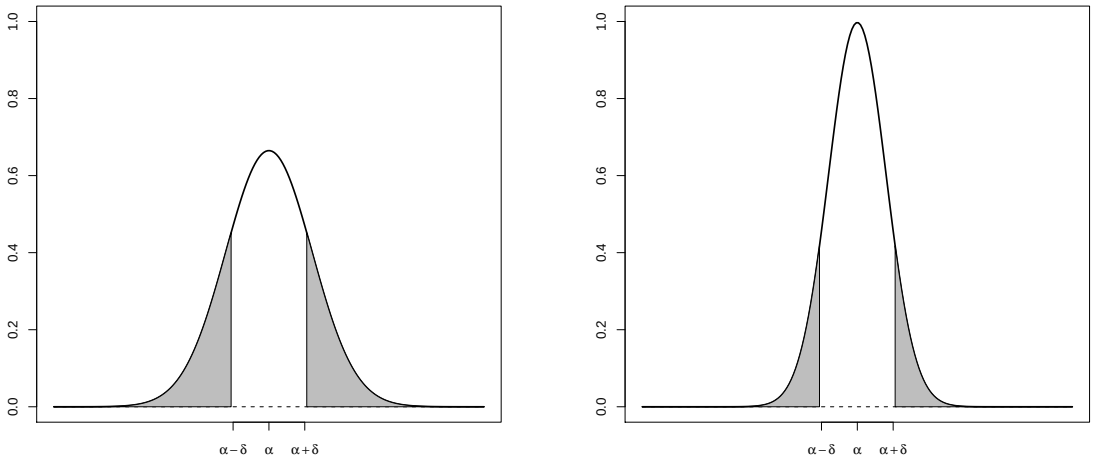
1. If $g\colon \mathbb{R} \to \mathbb{R}$ is continuous at x and $x_n \xrightarrow{p} x$, then $g(x_n) \xrightarrow{p} g(x)$.
2. If $x_n \xrightarrow{p} x$ and $y_n \xrightarrow{p} y$, then $x_n + y_n \xrightarrow{p} x + y$ and $x_n y_n \xrightarrow{p} xy$.
3. If $x_n \xrightarrow{p} x$ and $\alpha_n \to \alpha$, then $x_n + \alpha_n \xrightarrow{p} x + \alpha$ and $x_n \alpha_n \xrightarrow{p} x\alpha$.[9]

We say that $\{x_n\}$ converges to x **in mean square** if

$$\mathbb{E}\left[(x_n - x)^2\right] \to 0 \quad \text{as } n \to \infty \tag{2.27}$$

In symbols, this convergence is represented by $x_n \xrightarrow{ms} x$.

[9] Here $\{\alpha_n\}$ is a nonrandom scalar sequence.



(a) $n = 10$ (b) $n = 20$

Figure 2.8: $\mathbb{P}\{|x_n - \alpha| > \delta\}$ for $x_n \sim \mathcal{N}(\alpha, 1/n)$

Fact 2.5.2. Regarding convergence in mean square, the following statements are true:

1. If $x_n \overset{ms}{\to} x$, then $x_n \overset{p}{\to} x$.
2. If α is constant, then $x_n \overset{ms}{\to} \alpha$ if and only if $\mathbb{E}[x_n] \to \alpha$ and $\text{var}[x_n] \to 0$.

Part 1 of fact 2.5.2 follows from Chebyshev's inequality (page 60). Using monotonicity of $\mathbb{P}$ and then applying (2.12) to $x_n - x$, we obtain

$$\mathbb{P}\{|x_n - x| > \delta\} \leq \mathbb{P}\{|x_n - x| \geq \delta\} \leq \frac{\mathbb{E}[(x_n - x)^2]}{\delta^2}$$

Part 1 of fact 2.5.2 follows. Part 2 is implied by the equality

$$\mathbb{E}[(x_n - \alpha)^2] = \text{var}[x_n] + (\mathbb{E}[x_n] - \alpha)^2$$

Verification of this equality is an exercise.

Example 2.5.2. In example 2.5.1, we stated that if $x_n \sim \mathcal{N}(\alpha, 1/n)$, then $x_n \overset{p}{\to} \alpha$. This follows from parts 1 and 2 of fact 2.5.2, since $\mathbb{E}[x_n] = \alpha$ and $\text{var}[x_n] = 1/n \to 0$.

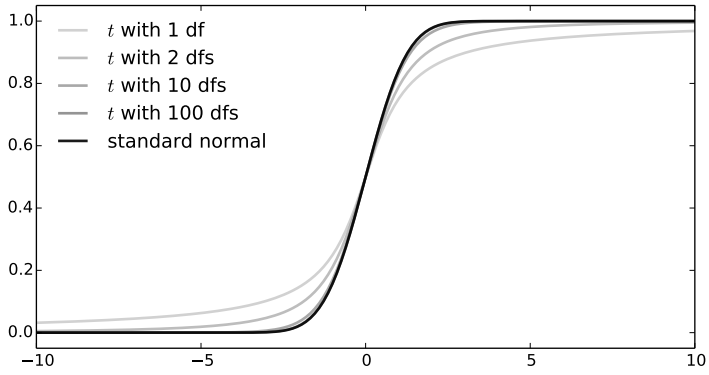


Figure 2.9: t-distribution with k df converges to $\mathcal{N}(0,1)$ as $k \to \infty$

Let $\{F_n\}_{n=1}^{\infty}$ be a sequence of cdfs, and let F be a cdf. We say that F_n **converges weakly** to F if, for any s such that F is continuous at s, we have

$$F_n(s) \to F(s) \quad \text{as} \quad n \to \infty$$

Example 2.5.3. It is well-known that the cdf of the t-distribution with k degrees of freedom converges to the standard normal cdf as $k \to \infty$. This convergence is illustrated in figure 2.9.

Sometimes densities are easier to work with than cdfs. In this connection, note that pointwise convergence of densities implies weak convergence of the corresponding distribution functions:

Fact 2.5.3. Let $\{F_n\}_{n=1}^{\infty}$ be a sequence of cdfs, and let F be a cdf. Suppose that all these cdfs are differentiable, and let p_n and p be the densities of F_n and F respectively. If $p_n(s) \to p(s)$ for all $s \in \mathbb{R}$, then F_n converges weakly to F.

Let $\{x_n\}_{n=1}^{\infty}$ and x be random variables, where $x_n \sim F_n$ and $x \sim F$. We say that x_n converges **in distribution** to x if F_n converges weakly to F. In symbols, this convergence is represented by $x_n \xrightarrow{d} x$.

Fact 2.5.4. Regarding convergence in distribution, the following statements are true:

1. If $g \colon \mathbb{R} \to \mathbb{R}$ is continuous at x and $x_n \xrightarrow{d} x$, then $g(x_n) \xrightarrow{d} g(x)$.

2. If $x_n \xrightarrow{p} x$, then $x_n \xrightarrow{d} x$.

3. If α is constant and $x_n \xrightarrow{d} \alpha$, then $x_n \xrightarrow{p} \alpha$.

The next result is sometimes known as **Slutsky's theorem**.

Fact 2.5.5. If α is constant, $x_n \xrightarrow{p} \alpha$ and $y_n \xrightarrow{d} y$, then $x_n + y_n \xrightarrow{d} \alpha + y$ and $x_n y_n \xrightarrow{d} \alpha y$.

An immediate but useful conseqence is that

Fact 2.5.6. If $x_n \xrightarrow{p} 0$ and $y_n \xrightarrow{d} y$, then $x_n y_n \xrightarrow{p} 0$.

Indeed, by Slutsky's theorem (fact 2.5.5) we have $x_n y_n \xrightarrow{d} 0$. Since the limit is constant, 2.5.4 then tells us that convergence is in probability as well.

2.5.2 The Law of Large Numbers

Two of the most important theorems in both probability and statistics are the law of large numbers and the central limit theorem. In their simplest forms, these theorems deal with averages of independent and identically distributed (IID) sequences. The law of large numbers tells us that these averages converge in probability to the mean of the distribution in question. The central limit theorem tells us that a simple transform of the average converges to a normal distribution.

Let's start with the **law of large numbers**, which relates to the sample mean

$$\bar{x}_N := \frac{1}{N} \sum_{n=1}^{N} x_n$$

of a given sample $x_1, \ldots, x_N$

Theorem 2.5.1. *Let $\{x_n\}$ be an* IID *sequence of random variables with common distribution F. If the first moment $\int |s| F(ds)$ is finite, then*

$$\bar{x}_N \xrightarrow{p} \mathbb{E}[x_n] = \int s F(ds) \quad \text{as} \quad N \to \infty \tag{2.28}$$

To prove theorem 2.5.1, we can use fact 2.5.2 on page 76. In view of this fact, it suffices to show that $\mathbb{E}[\bar{x}_N] \to \int sF(ds)$ and $\text{var}[\bar{x}_N] \to 0$ as $N \to \infty$. These steps are left as an exercise (exercise 2.8.39). When you do the exercise, note to yourself exactly where independence bites.[10]

Example 2.5.4. To illustrate the law of large numbers, consider flipping a coin until 10 heads have occurred. The coin is not fair: The probability of heads is 0.4. Let x be the number of tails observed in the process. It is known that such an x has the negative binomial distribution, and, with a little bit of googling, we find that the mean $\mathbb{E}[x]$ is 15. This means that if we simulate many observations of x and take the sample mean, we should get a value close to 15. Code to do this is provided in the next listing. Can you see how this program works?[11]

Listing 1 Illustrates the LLN

```
import numpy as np
from random import uniform
num_repetitions = 10000
outcomes = np.empty(num_repetitions)

for i in range(num_repetitions):
    num_tails = 0
    num_heads = 0
    while num_heads < 10:
        b = uniform(0, 1)
        num_heads = num_heads + (b < 0.4)
        num_tails = num_tails + (b >= 0.4)
    outcomes[i] = num_tails

print(outcomes.mean())
```

At first glance, the law of large numbers (2.28) appears to only be a statement about the sample mean, but actually it can be applied to functions of the random variable

[10] The proof involves a bit of cheating, because it assumes that the variance of each x_n is finite. This second moment assumption is not necessary for the result, but it helps to simplify the proof.

[11] Hint: If u is uniform on $[0,1]$ and $q \in [0,1]$, then $\mathbb{P}\{u \leq q\} = q$. This fact is used to simulate the coin flips. Also recall that the logical values TRUE and FALSE are treated as 1 and 0 respectively in algebraic expressions.

as well. For example, if $h\colon \mathbb{R} \to \mathbb{R}$ is such that $\int |h(s)|F(ds)$ is finite, then

$$\frac{1}{N}\sum_{n=1}^{N} h(x_n) \xrightarrow{p} \mathbb{E}\left[h(x_n)\right] = \int h(s)F(ds) \tag{2.29}$$

This can be confirmed by letting $y_n := h(x_n)$ and then applying theorem 2.5.1.

Also, the law of large numbers applies to probabilities as well as expectations. To see this, let $x \sim F$, fix $B \subset \mathbb{R}$, and consider the probability $\mathbb{P}\{x \in B\}$. Let h be the function defined by $h(s) = \mathbb{1}\{s \in B\}$ for all $s \in \mathbb{R}$. Using the principle that expectations of indicator functions equal probabilities of events (page 58), we have

$$\mathbb{E}\left[h(x)\right] = \mathbb{E}\left[\mathbb{1}\{x \in B\}\right] = \mathbb{P}\{x \in B\}$$

It now follows from (2.29) that if $\{x_n\}$ is an IID sample from F, then

$$\frac{1}{N}\sum_{n=1}^{N} \mathbb{1}\{x_n \in B\} \xrightarrow{p} \mathbb{P}\{x_n \in B\} \tag{2.30}$$

The left hand side is the fraction of the sample that falls in the set B, and (2.30) tells us that this fraction converges to the probability that $x_n \in B$.

2.5.3 The Central Limit Theorem

The **central limit theorem** is another classical result from probability theory. It is arguably one of the most beautiful and important results in all of mathematics. Relative to the LLN, it requires an additional second moment condition.

Theorem 2.5.2. *Assume the conditions of theorem 2.5.1. If, in addition, the second moment $\int s^2 F(ds)$ is finite, then*

$$\sqrt{N}(\bar{x}_N - \mu) \xrightarrow{d} y \sim \mathcal{N}(0, \sigma^2) \quad \text{as} \quad N \to \infty \tag{2.31}$$

where $\mu := \int sF(ds) = \mathbb{E}\left[x_n\right]$ and $\sigma^2 := \int (s-\mu)^2 F(ds) = \operatorname{var}[x_n]$.

Another common statement of the central limit theorem is as follows: If all the conditions of theorem 2.5.2 are satisfied, then

$$z_N := \sqrt{N}\left\{\frac{\bar{x}_N - \mu}{\sigma}\right\} \xrightarrow{d} z \sim \mathcal{N}(0,1) \quad \text{as} \quad N \to \infty \tag{2.32}$$

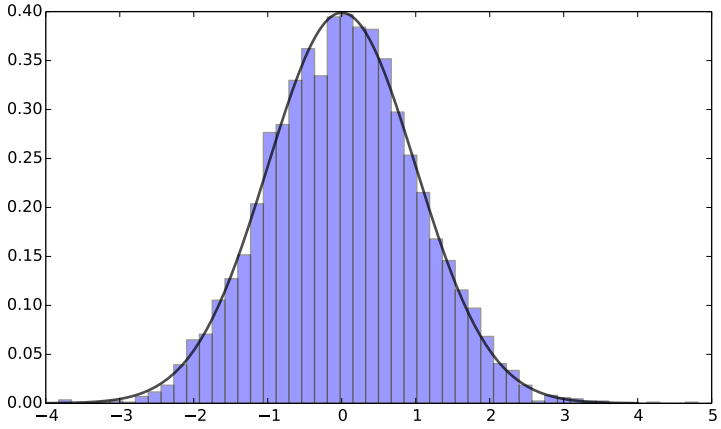


Figure 2.10: Illustration of the CLT

Exercise 2.8.40 asks you to confirm this via theorem 2.5.2 and fact 2.5.4.

The central limit theorem tells us about the distribution of the sample mean when N is large. Arguing informally, for N large we have

$$\sqrt{N}(\bar{x}_N - \mu) \approx y \sim \mathcal{N}(0, \sigma^2)$$

$$\therefore \quad \bar{x}_N \approx \frac{y}{\sqrt{N}} + \mu \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{N}\right)$$

Here $\approx$ means that the distributions are approximately equal. We see that $\bar{x}_N$ is approximately normal, with mean equal to $\mu := \mathbb{E}[x_1]$ and variance converging to zero at a rate proportional to $1/N$.

The convergence in (2.32) is illustrated by listing 2, the output of which is given in figure 2.10. The listing generates 5,000 observations of the random variable z_N defined in (2.32), where each x_n is $\chi^2(5)$. (The mean of this distribution is 5, and the variance is $2 \times 5 = 10$.) The observations of z_N are stored in the vector `outcomes`, and then histogrammed. At the end of the listing we superimpose the density of the standard normal distribution over the histogram. As predicted, the fit is relatively good.

Before finishing this section, we briefly note the following asymptotic result, which is frequently used in conjunction with the central limit theorem:

Listing 2 Illustrates the CLT

```
import numpy as np
import scipy.stats as st
import matplotlib.pyplot as plt

num_replications = 5000
outcomes = np.empty(num_replications)
N = 1000
k = 5      # Degrees of freedom
chi = st.chi2(k)

for i in range(num_replications):
    xvec = chi.rvs(N)
    outcomes[i] = np.sqrt(N / (2 * k)) * (xvec.mean() - k)

xmin, xmax = -4, 4
grid = np.linspace(xmin, xmax, 200)
fig, ax, = plt.subplots()
ax.hist(outcomes, bins=50, normed=True, alpha=0.4)
ax.plot(grid, st.norm.pdf(grid), 'k-', lw=2, alpha=0.7)
plt.show()
```

Theorem 2.5.3. *Let $\{t_n\}$ be a sequence of random numbers and let θ be a constant. Suppose that $\sqrt{n}(t_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$ for some $\sigma > 0$. Suppose further that $g\colon \mathbb{R} \to \mathbb{R}$ is differentiable at θ and $g'(\theta) \neq 0$. Under these conditions we have*

$$\sqrt{n}\{g(t_n) - g(\theta)\} \xrightarrow{d} \mathcal{N}(0, g'(\theta)^2\sigma^2) \quad \text{as} \quad n \to \infty \tag{2.33}$$

The technique illustrated in theorem 2.5.3 is referred to as the **delta method**. The delta method is extremely useful, particularly when one seeks the asymptotic distribution of certain kinds estimators. We will see its importance in some applications later on. The proof of theorem 2.5.3 is based on a Taylor expansion of g around the point θ, and can be found in almost any text on mathematical statistics. Exercise 2.8.43 walks you throught the most important ideas.

Instead of giving a full proof here, we will cover some parts of the proof of the following corollary: If the conditions of theorem 2.5.2 are satisfied and $g\colon \mathbb{R} \to \mathbb{R}$ is

differentiable at μ with $g'(\mu) \neq 0$, then

$$\sqrt{N}\{g(\bar{x}_N) - g(\mu)\} \xrightarrow{d} \mathcal{N}(0, g'(\mu)^2\sigma^2) \quad \text{as} \quad N \to \infty \tag{2.34}$$

2.6 Random Vectors and Matrices

A **random vector** $\mathbf{x}$ is just a sequence of K random variables $(x_1, \ldots, x_K)$. Each realization of $\mathbf{x}$ is an element of $\mathbb{R}^K$. The distribution (or cdf) of $\mathbf{x}$ is the joint distribution F of $(x_1, \ldots, x_K)$. That is,

$$F(\mathbf{s}) :=: F(s_1, \ldots, s_K) := \mathbb{P}\{x_1 \leq s_1, \ldots, x_K \leq s_K\} :=: \mathbb{P}\{\mathbf{x} \leq \mathbf{s}\} \tag{2.35}$$

for each $\mathbf{s}$ in $\mathbb{R}^K$. (Here and in what follows, the statement $\mathbf{x} \leq \mathbf{s}$ means that $x_n \leq s_n$ for $n = 1, \ldots, K$.)

Just as some but not all distributions on $\mathbb{R}$ have a density representation (see §2.3.2), some but not all distributions on $\mathbb{R}^K$ can be represented by a density. We say that $f\colon \mathbb{R}^K \to \mathbb{R}$ is the density of random vector $\mathbf{x} := (x_1, \ldots, x_K)$ if

$$\int_B f(\mathbf{s})\, d\mathbf{s} = \mathbb{P}\{\mathbf{x} \in B\} \tag{2.36}$$

for every subset B of $\mathbb{R}^K$.[12] Most of the distributions we work with in this course have density representations.

For random vectors, the definition of independence mirrors the scalar case. In particular, a collection of random vectors $\mathbf{x}_1, \ldots, \mathbf{x}_N$ is called **independent** if, given any $\mathbf{s}_1, \ldots, \mathbf{s}_N$, we have

$$\mathbb{P}\{\mathbf{x}_1 \leq \mathbf{s}_1, \ldots, \mathbf{x}_N \leq \mathbf{s}_N\} = \mathbb{P}\{\mathbf{x}_1 \leq \mathbf{s}_1\} \times \cdots \times \mathbb{P}\{\mathbf{x}_N \leq \mathbf{s}_N\}$$

We note the following multivariate version of fact 2.4.2:

Fact 2.6.1. If $\mathbf{x}$ and $\mathbf{y}$ are independent and g and f are any functions, then $f(\mathbf{x})$ and $g(\mathbf{y})$ are also independent.

A **random $N \times K$ matrix** $\mathbf{X}$ is a rectangular $N \times K$ array of random variables. In this section, we briefly review some properties of random vectors and matrices.

[12] Actually, some subsets of $\mathbb{R}^K$ are so messy that it's not possible to integrate over them, so we only require (2.36) to hold for a large but suitably well-behaved class of sets called the *Borel* sets. See any text on measure theory for details.

2.6.1 Expectations for Vectors and Matrices

Let $\mathbf{x} = (x_1, \ldots, x_K)$ be a random vector taking values in $\mathbb{R}^K$ with $\mu_k := \mathbb{E}[x_k]$ for all $k = 1, \ldots, K$. The **expectation** $\mathbb{E}[\mathbf{x}]$ of vector $\mathbf{x}$ is defined as the vector of expectations:

$$\mathbb{E}[\mathbf{x}] := \begin{pmatrix} \mathbb{E}[x_1] \\ \mathbb{E}[x_2] \\ \vdots \\ \mathbb{E}[x_K] \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_K \end{pmatrix} =: \boldsymbol{\mu}$$

More generally, if $\mathbf{X}$ is a random $N \times K$ matrix, then its expectation $\mathbb{E}[\mathbf{X}]$ is the matrix of the expectations:

$$\mathbb{E}[\mathbf{X}] := \begin{pmatrix} \mathbb{E}[x_{11}] & \mathbb{E}[x_{12}] & \cdots & \mathbb{E}[x_{1K}] \\ \mathbb{E}[x_{21}] & \mathbb{E}[x_{22}] & \cdots & \mathbb{E}[x_{2K}] \\ \vdots & \vdots & & \vdots \\ \mathbb{E}[x_{N1}] & \mathbb{E}[x_{N2}] & \cdots & \mathbb{E}[x_{NK}] \end{pmatrix}$$

Expectation of vectors and matrices maintains the linearity of scalar expectations:

Fact 2.6.2. If $\mathbf{X}$ and $\mathbf{Y}$ are random and $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ are conformable constant matrices, then

$$\mathbb{E}[\mathbf{A} + \mathbf{B}\mathbf{X} + \mathbf{C}\mathbf{Y}] = \mathbf{A} + \mathbf{B}\mathbb{E}[\mathbf{X}] + \mathbf{C}\mathbb{E}[\mathbf{Y}]$$

The **covariance** between random $N \times 1$ vectors $\mathbf{x}$ and $\mathbf{y}$ is

$$\operatorname{cov}[\mathbf{x}, \mathbf{y}] := \mathbb{E}[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{y} - \mathbb{E}[\mathbf{y}])']$$

The **variance-covariance matrix** of random vector $\mathbf{x}$ with $\boldsymbol{\mu} := \mathbb{E}[\mathbf{x}]$ is defined as

$$\operatorname{var}[\mathbf{x}] := \operatorname{cov}(\mathbf{x}, \mathbf{x}) = \mathbb{E}[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])'] = \mathbb{E}[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})']$$

Expanding this out, we get

$$\operatorname{var}[\mathbf{x}] = \begin{pmatrix} \mathbb{E}[(x_1 - \mu_1)(x_1 - \mu_1)] & \cdots & \mathbb{E}[(x_1 - \mu_1)(x_N - \mu_N)] \\ \mathbb{E}[(x_2 - \mu_2)(x_1 - \mu_1)] & \cdots & \mathbb{E}[(x_2 - \mu_2)(x_N - \mu_N)] \\ \vdots & \vdots & \vdots \\ \mathbb{E}[(x_N - \mu_N)(x_1 - \mu_1)] & \cdots & \mathbb{E}[(x_N - \mu_N)(x_N - \mu_N)] \end{pmatrix}$$

The j,k-th term is the scalar covariance between x_j and x_k. As a result, the principle diagonal contains the variance of each x_n.

Some simple algebra yields the alternative expressions

$$\operatorname{cov}[\mathbf{x}, \mathbf{y}] = \mathbb{E}[\mathbf{x}\mathbf{y}'] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{y}]' \quad \text{and} \quad \operatorname{var}[\mathbf{x}] = \mathbb{E}[\mathbf{x}\mathbf{x}'] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{x}]'$$

Fact 2.6.3. For any random vector $\mathbf{x}$, the variance-covariance matrix $\operatorname{var}[\mathbf{x}]$ is square, symmetric and nonnegative definite.

Fact 2.6.4. For any random vector $\mathbf{x}$, any constant conformable matrix $\mathbf{A}$ and any constant conformable vector $\mathbf{a}$, we have

$$\operatorname{var}[\mathbf{a} + \mathbf{A}\mathbf{x}] = \mathbf{A} \operatorname{var}[\mathbf{x}] \mathbf{A}'$$

2.6.2 Multivariate Gaussians

The **multivariate normal density** or **Gaussian density** in $\mathbb{R}^N$ is a function p of the form

$$p(\mathbf{s}) = (2\pi)^{-N/2} \det(\boldsymbol{\Sigma})^{-1/2} \exp\left\{ -\frac{1}{2} (\mathbf{s} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{s} - \boldsymbol{\mu}) \right\}$$

where $\boldsymbol{\mu}$ is any $N \times 1$ vector and $\boldsymbol{\Sigma}$ is a symmetric, positive definite $N \times N$ matrix. In symbols, we represent this distribution by $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Although we omit the derivations, it can be shown that if $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then

$$\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu} \quad \text{and} \quad \operatorname{var}[\mathbf{x}] = \boldsymbol{\Sigma}$$

We say that $\mathbf{x}$ is **normally distributed** if $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for some $N \times 1$ vector $\boldsymbol{\mu}$ and symmetric, positive definite $N \times N$ matrix $\boldsymbol{\Sigma}$. We say that $\mathbf{x}$ is **standard normal** if $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Sigma} = \mathbf{I}$.

Fact 2.6.5. $N \times 1$ random vector $\mathbf{x}$ is normally distributed if and only if $\mathbf{a}'\mathbf{x}$ is normally distributed in $\mathbb{R}$ for every constant $N \times 1$ vector $\mathbf{a}$.[13]

Fact 2.6.6. If $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\mathbf{a} + \mathbf{A}\mathbf{x} \sim \mathcal{N}(\mathbf{a} + \mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$.

Here, the fact that $\mathbf{a} + \mathbf{A}\mathbf{x}$ has mean $\mathbf{a} + \mathbf{A}\boldsymbol{\mu}$ and variance-covariance matrix $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'$ is not surprising. What is important is that normality is preserved .

Fact 2.6.7. Normally distributed random variables are independent if and only if they are uncorrelated. In particular, if both x and y are normally distributed and $\operatorname{cov}[x, y] = 0$, then x and y are independent.

Fact 2.6.8. If $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $(\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \sim \chi^2(k)$, where $k :=$ length of $\mathbf{x}$.

Fact 2.6.9. If $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\mathbf{A}$ is a conformable idempotent and symmetric matrix with $\operatorname{rank}(\mathbf{A}) = j$, then $\mathbf{x}'\mathbf{A}\mathbf{x} \sim \chi^2(j)$. (In view of fact 1.4.7, when using this result it is sufficient to show that $\operatorname{trace}(\mathbf{A}) = j$.)

[13] If $\mathbf{a} = \mathbf{0}$ then we can interpret $\mathbf{a}'\mathbf{x}$ as a "normal" random variable with zero variance.

2.6.3 Convergence of Random Matrices

Next we extend the probabilistic notions of convergence discussed in §2.5.1 to random vectors and matrices. Beginning with the notion of convergence in probability (see §2.5.1 for the scalar case), let $\{\mathbf{X}_n\}_{n=1}^{\infty}$ be a sequence of random $N \times K$ matrices. We say that $\mathbf{X}_n$ converges to a random $N \times K$ matrix $\mathbf{X}$ **in probability** and write $\mathbf{X}_n \xrightarrow{p} \mathbf{X}$ if every element of $\mathbf{X}_n$ converges to the corresponding element of $\mathbf{X}$ in probability in the scalar sense. That is,

$$\mathbf{X}_n \xrightarrow{p} \mathbf{X} \quad \iff \quad x_{ij}^n \xrightarrow{p} x_{ij} \text{ for all } i \text{ and } j$$

Similarly, a sequence of random vectors is said to converge in probability when the individual components converge in probability. That is,

$$\begin{pmatrix} x_1^n \\ \vdots \\ x_K^n \end{pmatrix} \xrightarrow{p} \begin{pmatrix} x_1 \\ \vdots \\ x_K \end{pmatrix} \quad \iff \quad x_k^n \xrightarrow{p} x_k \text{ for all } k$$

Fact 2.6.10. Let $\{\mathbf{X}_n\}$ be a sequence of random $N \times K$ matrices, let $\{\mathbf{x}_n\}$ is a sequence of random vectors in $\mathbb{R}^K$ and let $\mathbf{X}$ and $\mathbf{x}$ be, respectively, a random matrix and vector of the same dimensions. Then,

1. $\mathbf{x}_n \xrightarrow{p} \mathbf{x}$ if and only if $\|\mathbf{x}_n - \mathbf{x}\| \xrightarrow{p} 0$.
2. $\mathbf{X}_n \xrightarrow{p} \mathbf{X}$ if and only if $\|\mathbf{X}_n - \mathbf{X}\| \xrightarrow{p} 0$.

Here the first norm is the ordinary Euclidean norm and the second is the matrix norm of §1.4.5.

Now let's extend the notion of convergence in *distribution* to random vectors. The definition is almost identical to the scalar case, with only the obvious modifications. Let $\{F_n\}_{n=1}^{\infty}$ be a sequence of cdfs on $\mathbb{R}^K$, and let F be a cdf on $\mathbb{R}^K$. We say that F_n converges to F **weakly** if, for any $\mathbf{s}$ such that F is continuous at $\mathbf{s}$, we have

$$F_n(\mathbf{s}) \to F(\mathbf{s}) \quad \text{as } n \to \infty$$

Let $\{\mathbf{x}_n\}_{n=1}^{\infty}$ and $\mathbf{x}$ be random vectors in $\mathbb{R}^K$, where $\mathbf{x}_n \sim F_n$ and $\mathbf{x} \sim F$. We say that $\mathbf{x}_n$ converges **in distribution** to $\mathbf{x}$ if F_n converges weakly to F. In symbols, this convergence is represented by $\mathbf{x}_n \xrightarrow{d} \mathbf{x}$.

As discussed above, convergence of $\mathbf{x}_n$ to $\mathbf{x}$ in probability simply requires that the elements of $\mathbf{x}_n$ converge in probability (in the scalar sense) to the corresponding elements of $\mathbf{x}$. For convergence in distribution this is not generally true:

$$x_k^n \xrightarrow{d} x_k \text{ for all } k \text{ does } \textit{not} \text{ imply } \mathbf{x}_n := \begin{pmatrix} x_1^n \\ \vdots \\ x_K^n \end{pmatrix} \xrightarrow{d} \begin{pmatrix} x_1 \\ \vdots \\ x_K \end{pmatrix} =: \mathbf{x}$$

Put differently, convergence of the marginals does not necessarily imply convergence of the joint distribution. (As you might have guessed, one setting where convergence of the marginals implies convergence of the joint is when the elements of the vectors are independent, and the joint is just the product of the marginals.)

The fact that elementwise convergence in distribution does not necessarily imply convergence of the vectors is problematic, because vector convergence is harder to work with than scalar convergence. Fortunately, we have the following results, which provide a link from scalar to vector convergence:

Fact 2.6.11. Let $\mathbf{x}_n$ and $\mathbf{x}$ be random vectors in $\mathbb{R}^K$.

1. If $\mathbf{a}'\mathbf{x}_n \xrightarrow{d} \mathbf{a}'\mathbf{x}$ for any $\mathbf{a} \in \mathbb{R}^K$, then $\mathbf{x}_n \xrightarrow{d} \mathbf{x}$.
2. If $\mathbf{a}'\mathbf{x}_n \xrightarrow{p} \mathbf{a}'\mathbf{x}$ for any $\mathbf{a} \in \mathbb{R}^K$, then $\mathbf{x}_n \xrightarrow{p} \mathbf{x}$.

The second of these results is quite straightforward to prove (exercise 2.8.45). The first is more difficult (the standard argument uses characteristic functions). It is often referred as the Cramer-Wold device.

We noted a variety of useful results pertaining to convergence in probability and distribution for sums and products in the scalar case. Most of these carry over to the vector case essentially unchanged. For example,

Fact 2.6.12. Assuming conformability, the following statements are true:

1. If $\mathbf{X}_n \xrightarrow{p} \mathbf{X}$ and $\mathbf{X}_n$ and $\mathbf{X}$ are nonsingular, then $\mathbf{X}_n^{-1} \xrightarrow{p} \mathbf{X}^{-1}$.
2. If $\mathbf{X}_n \xrightarrow{p} \mathbf{X}$ and $\mathbf{Y}_n \xrightarrow{p} \mathbf{Y}$, then

$$\mathbf{X}_n + \mathbf{Y}_n \xrightarrow{p} \mathbf{X} + \mathbf{Y}, \quad \mathbf{X}_n\mathbf{Y}_n \xrightarrow{p} \mathbf{X}\mathbf{Y} \quad \text{and} \quad \mathbf{Y}_n\mathbf{X}_n \xrightarrow{p} \mathbf{Y}\mathbf{X}$$

3. If $\mathbf{X}_n \xrightarrow{p} \mathbf{X}$ and $\mathbf{A}_n \to \mathbf{A}$, then

$$\mathbf{X}_n + \mathbf{A}_n \xrightarrow{p} \mathbf{X} + \mathbf{A}, \quad \mathbf{X}_n\mathbf{A}_n \xrightarrow{p} \mathbf{X}\mathbf{A} \quad \text{and} \quad \mathbf{A}_n\mathbf{X}_n \xrightarrow{p} \mathbf{A}\mathbf{X}$$

In part 3 of fact 2.6.12, the matrices $\mathbf{A}_n$ and $\mathbf{A}$ are nonrandom. The convergence $\mathbf{A}_n \to \mathbf{A}$ means that each element of $\mathbf{A}_n$ converges in the usual scalar sense to the corresponding element of $\mathbf{A}$:

$$\mathbf{A}_n \to \mathbf{A} \text{ means } a_{ij}^n \to a_{ij} \text{ for all } i \text{ and } j$$

Alternatively, we can stack the matrices into vectors and take the norms, as discussed above. Then we say that $\mathbf{A}_n \to \mathbf{A}$ if $\|\mathbf{A}_n - \mathbf{A}\| \to 0$. The two definitions can be shown to be equivalent.

Example 2.6.1. To see how fact 2.6.12 can be used, let's establish convergence of the quadratic form

$$\mathbf{a}'\mathbf{X}_n\mathbf{a} \xrightarrow{p} \mathbf{a}'\mathbf{X}\mathbf{a} \quad \text{whenever } \mathbf{a} \text{ is a conformable constant vector and } \mathbf{X}_n \xrightarrow{p} \mathbf{X} \tag{2.37}$$

This follows from two applications of fact 2.6.12. Applying fact 2.6.12 once we get $\mathbf{a}'\mathbf{X}_n \xrightarrow{p} \mathbf{a}'\mathbf{X}$. Applying it a second time yields the convergence in (2.37).

As in the scalar case, convergence in probability and convergence in distribution are both preserved under continuous transformations:

Fact 2.6.13 (Continuous mapping theorem, vector case)**.** Let $\mathbf{x}_n$ and $\mathbf{x}$ be random vectors in $\mathbb{R}^K$, and let $g\colon \mathbb{R}^K \to \mathbb{R}^J$ be continuous at $\mathbf{x}$. In this setting,

1. If $\mathbf{x}_n \xrightarrow{d} \mathbf{x}$, then $g(\mathbf{x}_n) \xrightarrow{d} g(\mathbf{x})$.
2. If $\mathbf{x}_n \xrightarrow{p} \mathbf{x}$, then $g(\mathbf{x}_n) \xrightarrow{p} g(\mathbf{x})$.

Another result used routinely in econometric theory is the vector version of Slutsky's theorem:

Fact 2.6.14 (Slutsky's theorem, vector case)**.** Let $\mathbf{x}_n$ and $\mathbf{x}$ be random vectors in $\mathbb{R}^K$, let $\mathbf{Y}_n$ be random matrices, and let $\mathbf{C}$ be a constant matrix. If $\mathbf{Y}_n \xrightarrow{p} \mathbf{C}$ and $\mathbf{x}_n \xrightarrow{d} \mathbf{x}$, then

$$\mathbf{Y}_n\mathbf{x}_n \xrightarrow{d} \mathbf{C}\mathbf{x} \quad \text{and} \quad \mathbf{Y}_n + \mathbf{x}_n \xrightarrow{d} \mathbf{C} + \mathbf{x}$$

whenever the matrices are conformable.

The delta method from theorem 2.5.3 on page 81 extends to random vectors. For example, let $g\colon \mathbb{R}^K \to \mathbb{R}$ be differentiable at a vector $\boldsymbol{\theta} \in \mathbb{R}^K$, in the sense that the **gradiant vector**

$$\nabla g(\boldsymbol{\theta}) := \begin{pmatrix} \frac{\partial g(\boldsymbol{\theta})}{\partial \theta_1} \\ \vdots \\ \frac{\partial g(\boldsymbol{\theta})}{\partial \theta_K} \end{pmatrix}$$

is well defined (i.e., the limit defining each of the partial derivatives exists). In this context,

Fact 2.6.15. If $\{\mathbf{t}_n\}$ is a sequence of random vectors in $\mathbb{R}^K$ with $\sqrt{n}(\mathbf{t}_n - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}(0, \Sigma)$ for some $\boldsymbol{\theta} \in \mathbb{R}^K$ and positive definite $K \times K$ matrix Σ, then

$$\sqrt{n}\{g(\mathbf{t}_n) - g(\boldsymbol{\theta})\} \xrightarrow{d} \mathcal{N}(0, \nabla g(\boldsymbol{\theta})' \Sigma \nabla g(\boldsymbol{\theta})) \quad \text{as} \quad n \to \infty \tag{2.38}$$

whenever $\nabla g(\boldsymbol{\theta})' \Sigma \nabla g(\boldsymbol{\theta})$ is positive. This last assumption will be satisfied if, for example, at least one of the partial derivatives in $\nabla g(\boldsymbol{\theta})$ is nonzero (why?).

2.6.4 Vector LLN and CLT

With the above definitions of convergence in hand, we can move on to the next topic: Vector LLN and CLT. The scalar LLN and CLT that we discussed in §2.5 extend to the vector case in a natural way. For example, we have the following result:

Theorem 2.6.1. *Let $\{\mathbf{x}_n\}$ be an* IID *sequence of random vectors in $\mathbb{R}^K$ with $\mathbb{E}\left[\|\mathbf{x}_n\|^2\right] < \infty$. Let $\boldsymbol{\mu} := \mathbb{E}\left[\mathbf{x}_n\right]$ and let $\Sigma := \operatorname{var}[\mathbf{x}_n]$. For this sequence we have*

$$\bar{\mathbf{x}}_N := \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n \xrightarrow{p} \boldsymbol{\mu} \quad \textit{and} \quad \sqrt{N}\left(\bar{\mathbf{x}}_N - \boldsymbol{\mu}\right) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma) \tag{2.39}$$

Figure 2.11 illustrates the LLN in two dimensions. The green dot is the point $\mathbf{0} = (0, 0)$ in $\mathbb{R}^2$. The black dots are IID observations $\mathbf{x}_1, \ldots, \mathbf{x}_N$ of a random vector with mean $\boldsymbol{\mu} = \mathbf{0}$. The red dot is the sample mean $\frac{1}{N}\sum_{n=1}^{N} \mathbf{x}_n$. (Remember that we are working with vectors here, so the summation and scalar multiplication in the sample mean $\bar{\mathbf{x}}_N$ is done elementwise—in this case for two elements. In particular, the sample mean is a linear combination of the observations $\mathbf{x}_1, \ldots, \mathbf{x}_N$.) By the vector LLN, the red dot converges to the green dot.

The vector LLN in theorem 2.6.1 follows from the scalar LLN. To see this, let $\mathbf{x}_n$ be as in theorem 2.6.1, let $\mathbf{a}$ be any constant vector in $\mathbb{R}^K$ and consider the scalar sequence

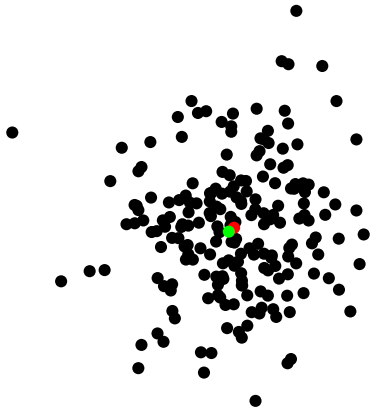


Figure 2.11: LLN, vector case

$\{y_n\}$ defined by $y_n = \mathbf{a}'\mathbf{x}_n$. The sequence $\{y_n\}$ inherets the IID property from $\{\mathbf{x}_n\}$.[14] By the scalar LLN (theorem 2.5.1) we have

$$\frac{1}{N}\sum_{n=1}^{N} y_n \xrightarrow{p} \mathbb{E}[y_n] = \mathbb{E}[\mathbf{a}'\mathbf{x}_n] = \mathbf{a}'\mathbb{E}[\mathbf{x}_n] = \mathbf{a}'\boldsymbol{\mu}$$

But

$$\frac{1}{N}\sum_{n=1}^{N} y_n = \frac{1}{N}\sum_{n=1}^{N} \mathbf{a}'\mathbf{x}_n = \mathbf{a}'\left[\frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n\right] = \mathbf{a}'\bar{\mathbf{x}}_N$$

Since $\mathbf{a}$ was arbitrary, we have shown that

$$\mathbf{a}'\bar{\mathbf{x}}_N \xrightarrow{p} \mathbf{a}'\boldsymbol{\mu} \text{ for any } \mathbf{a} \in \mathbb{R}^K$$

The claim $\bar{\mathbf{x}}_N \xrightarrow{p} \boldsymbol{\mu}$ now follows from fact 2.6.11.

The vector CLT in theorem 2.6.1 also follows from the scalar case. The proof is rather similar to the vector LLN proof we have just completed. See exercise 2.8.48.

2.7 Further Reading

To be written

[14] Functions of independent random variables are themselves independent (fact 2.4.2, page 71).

2.8 Exercises

Ex. 2.8.1. Suppose that $\mathbb{P}$ is a probability on $(\Omega, \mathcal{F})$, so that $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$ whenever A and B are disjoint. Show that if A, B and C are disjoint, then $\mathbb{P}(A \cup B \cup C) = \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C)$.

Ex. 2.8.2. Prove fact 2.1.2: $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$ for any A, B.[15]

Ex. 2.8.3. Given sample space $\Omega := \{1, 2, 3\}$, let $A := \{1\}$, $B := \{2\}$ and $C := \{3\}$. Let $\mathbb{P}(A) = \mathbb{P}(B) = 1/3$. Compute $\mathbb{P}(C)$, $\mathbb{P}(A \cup B)$, $\mathbb{P}(A \cap B)$, $\mathbb{P}(A^c)$, $\mathbb{P}(A^c \cup B^c)$ and $\mathbb{P}(A \mid B)$. Are A and C independent?

Ex. 2.8.4. A dice is designed so that the probability of getting face m is qm, where $m \in \{1, \ldots, 6\}$ and q is a constant. Compute q.

Ex. 2.8.5. Let Ω be a nonempty finite set, and let ω_0 be a fixed element of Ω. For each $A \subset \Omega$, define $\mathbb{P}(A) := \mathbb{1}\{\omega_0 \in A\}$. Is $\mathbb{P}$ a probability on Ω? Why or why not?

Ex. 2.8.6. Let Ω be any sample space, and let $\mathbb{P}$ be a probability on the subsets $\mathcal{F}$. Let $A \in \mathcal{F}$. Show that if $\mathbb{P}(A) = 0$ or $\mathbb{P}(A) = 1$, then A is independent of every other event in $\mathcal{F}$. Show that if A is independent of itself, then either $\mathbb{P}(A) = 0$ or $\mathbb{P}(A) = 1$. Show that if A and B are independent, then A^c and B^c are also independent.

Ex. 2.8.7. Let $\mathbb{P}$ and Ω be defined as in example 2.1.4. Show that $\mathbb{P}$ is additive, in the sense that if A and B are disjoint events, then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.

Ex. 2.8.8. Let $\mathbb{P}$ and Ω be defined as in example 2.1.4. Let A be the event that the first switch is on, and let B be the event that the second switch is on. Show that A and B are independent under $\mathbb{P}$.

Ex. 2.8.9. Show that when Ω is finite, a random variable x on Ω can only take on a finite set of values (i.e., has finite range).[16]

Ex. 2.8.10. Fact 2.2.4 on page 59 states that, among other things, if x is a random variable, then $\mathbb{E}[\alpha x] = \alpha \mathbb{E}[x]$. Show this for the case where x is a finite random variable.

[15] Hint: Sketching the Venn diagram, convince yourself that $A = [(A \cup B) \setminus B] \cup (A \cap B)$. Finish the proof using the definition of a probability and fact 2.1.1 (page 49).

[16] Hint: Have a look at the definition of a function in §4.2.

Ex. 2.8.11. Fact 2.2.4 on page 59 states that, among other things, if x and y are two random variables, then $\mathbb{E}[x+y] = \mathbb{E}[x] + \mathbb{E}[y]$. Instead of giving the full proof, try to show this for the binary random variables

$$x(\omega) = \mathbb{1}_A(\omega) \quad \text{and} \quad y(\omega) = \mathbb{1}_B(\omega) \tag{2.40}$$

Note that even this is a little fiddly—see the solution if you get lost.

Ex. 2.8.12. Recall F defined in (2.13). We claimed that $\lim_{s\to\infty} F(s) = 1$. Verify this when x is the finite-valued random variable in (2.7).

Ex. 2.8.13. Recall F defined in (2.13). Suppose that x is the finite-valued random variable in (2.7). Show that $\lim_{s\to-\infty} F_x(s) = 0$. If you can, show that F is right-continuous.

Ex. 2.8.14. Prove the claim in fact 2.3.2 on page 62.

Ex. 2.8.15. Let x be a discrete random variable taking values $s_1, \ldots, s_J$, and let $p_j := \mathbb{P}\{x = s_j\}$. Show that $0 \leq p_j \leq 1$ for each j, and $\sum_{j=1}^J p_j = 1$.

Ex. 2.8.16. This exercise describes the **inverse transform** method for generating random variables with arbitrary distribution from uniform random variables. The uniform cdf on $[0,1]$ is given by $F(s) = 0$ if $s < 0$, $F(s) = s$ if $0 \leq s \leq 1$, and $F(s) = 1$ if $s > 1$. Let G be another cdf on $\mathbb{R}$. Suppose that G is strictly increasing, and let G^{-1} be the inverse (quantile). Show that if $u \sim F$, then $G^{-1}(u) \sim G$.

Ex. 2.8.17. Let $x \sim F$ where F is the uniform cdf on $[0,1]$. Give an expression for the cdf G of the random variable $y = x^2$.

Ex. 2.8.18. Let F be the cdf on $\mathbb{R}$ defined by $F(s) = e^s/(1+e^s)$ for $s \in \mathbb{R}$.

1. Obtain the quantile function corresponding to F.
2. Let $x \sim F$ and compute $\mathbb{P}\{0 \leq x \leq \ln 2\}$.

Ex. 2.8.19. Let $y \sim F$, where F is a cdf. Show that $F(s) = \mathbb{E}[\mathbb{1}\{y \leq s\}]$ for any s.

Ex. 2.8.20. Confirm monotonicity of expectations (fact 2.2.5 on page 60) for the special case where x and y are the random variables in (2.40).

Ex. 2.8.21. Prove fact 2.2.6 (existence of k-th moment implies existence of j-th moment for all $j \leq k$).

Ex. 2.8.22. Confirm the expression for variance of linear combinations in fact 2.4.8.

Ex. 2.8.23. Let x and y be scalar random variables. With reference to fact 2.4.9 on page 73, is it true that $\operatorname{corr}[\alpha x, \beta y] = \operatorname{corr}[x, y]$ for *any* constant scalars α and β? Why or why not?

Ex. 2.8.24. Confirm the claim in fact 2.4.10: If x and y are independent, then $\operatorname{cov}[x, y] = \operatorname{corr}[x, y] = 0$.

Ex. 2.8.25. Let x_1 and x_2 be random variables with densities p_1 and p_2. Let q be their joint density. Show that x_1 and x_2 are independent whenever $q(s, s') = p_1(s)p_2(s')$ for every $s, s' \in \mathbb{R}$.

Ex. 2.8.26. Fact 2.4.2 tells us that if x and y are independent random variables and g and f are any two functions, then $f(x)$ and $g(y)$ are independent. Prove this for the case where $f(x) = 2x$ and $g(y) = 3y - 1$.

Ex. 2.8.27. Let x and y be independent uniform random variables on $[0, 1]$. Let $z := \max\{x, y\}$. Compute the cdf, density and mean of z.[17] In addition, compute the cdf of $w := \min\{x, y\}$.

Ex. 2.8.28. A discrete random variable x taking values in the set $\mathbb{N}_0 := \{0, 1, 2, \ldots\}$ has pmf $\{p_j\} = \{p_0, p_1, \ldots\}$, if $\mathbb{P}\{x = j\} = p_j$ for all $j \in \mathbb{N}_0$. (Here $\{p_j\}$ is a sequence in $[0, 1]$ with $\sum_{j=0}^{\infty} p_j = 1$.) Let u and v be indepedent random variables taking values in $\mathbb{N}_0$ with pmfs $\{p_j\}$ and $\{q_j\}$ respectively, and let $z := u + v$. Obtain an expression for the pmf of z in terms of $\{p_j\}$ and $\{q_j\}$.

Ex. 2.8.29. Confirm the solutions in (2.26).

Ex. 2.8.30. Consider the setting of §2.4.4. Let α^*, β^* and ℓ^* be as defined there. Let the prediction error u be defined as $u := y - \ell^*(x)$. Show that

1. $\mathbb{E}[\ell^*(x)] = \mathbb{E}[y]$
2. $\operatorname{var}[\ell^*(x)] = \operatorname{corr}[x, y]^2 \operatorname{var}[y]$
3. $\operatorname{var}[u] = (1 - \operatorname{corr}[x, y]^2) \operatorname{var}[y]$

Ex. 2.8.31. Continuing on from exercise 2.8.30, show that $\operatorname{cov}[\ell^*(x), u] = 0$.

[17] Hint: Fix $s \in \mathbb{R}$ and compare the sets $\{z \leq s\}$ and $\{x \leq s\} \cap \{y \leq s\}$. What is the relationship between these two sets?

Ex. 2.8.32. Show that if $\mathbf{x}$ is a random vector with $\mathbb{E}[\mathbf{x}\mathbf{x}'] = \mathbf{I}$ and $\mathbf{A}$ is a conformable constant matrix, then $\mathbb{E}[\mathbf{x}'\mathbf{A}\mathbf{x}] = \text{trace}(\mathbf{A})$.

Ex. 2.8.33. Let $\mathbf{x}$ be random and let $\mathbf{a}$ be constant. Show that if $\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}$ and $\text{var}[\mathbf{x}] = \boldsymbol{\Sigma}$, then $\mathbb{E}[\mathbf{a}'\mathbf{x}] = \mathbf{a}'\boldsymbol{\mu}$ and $\text{var}[\mathbf{a}'\mathbf{x}] = \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}$.

Ex. 2.8.34. Let $\mathbf{x}$ be a random $K \times 1$ vector. Show that $\mathbb{E}[\mathbf{x}\mathbf{x}']$ is nonnegative definite.

Ex. 2.8.35. Let $\mathbf{x} = (x_1, \ldots, x_N) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_N)$.

1. Are x_1 and x_2 independent? Why or why not?
2. What is the distribution of x_1^2? Why?
3. What is the distribution of x_1^2/x_2^2? Why?
4. What is the distribution of $x_1[2/(x_2^2 + x_3^2)]^{1/2}$? Why?
5. What is the distribution of $\|\mathbf{x}\|^2$? Why?
6. If $\mathbf{a}$ is an $N \times 1$ constant vector, what is the distribution of $\mathbf{a}'\mathbf{x}$?

Ex. 2.8.36. Prove the Chebyshev inequalities (2.11) and (2.12).

Ex. 2.8.37. Let $\{x_n\}$ be a sequence of random variables satisfying $x_n = y$ for all n, where y is a single random variable. Show that if $\mathbb{P}\{y = -1\} = \mathbb{P}\{y = 1\} = 0.5$, then $x_n \xrightarrow{p} 0$ fails. Show that if $\mathbb{P}\{y = 0\} = 1$, then $x_n \xrightarrow{p} 0$ holds.

Ex. 2.8.38. We saw in fact 2.5.4 that if $x_n \xrightarrow{p} x$, then $x_n \xrightarrow{d} x$. Show that the converse is not generally true. In other words, give an example of a sequence of random variables $\{x_n\}$ and random variable x such that x_n converges to x in distribution, but not in probability.

Ex. 2.8.39. In this exercise, we complete the proof of the LLN on page 78. Let $\{x_n\}$ be an IID sequence of random variables with common distribution F. Show that $\mathbb{E}[\bar{x}_N] \to \int sF(ds)$ and $\text{var}[\bar{x}_N] \to 0$ as $N \to \infty$.

Ex. 2.8.40. Confirm (2.32) via theorem 2.5.2 and fact 2.5.4.

Ex. 2.8.41. Let $u \sim U[0,1]$ and $x_n := n\mathbb{1}\{0 \leq u \leq 1/n\}$ for $n = 1, 2, \ldots$

1. Calculate the expectation of x_n.

2. Show that $x_n \xrightarrow{p} 0$ as $n \to \infty$.

Ex. 2.8.42. Let x be any random variable with $\mathbb{E}[x] = \mu$ and $\text{var}[x] = \sigma^2 < \infty$. Show that $x_n := x/n$ converges to zero in probability as $n \to \infty$.

Ex. 2.8.43. This exercise covers some of the proof behind theorem 2.5.3 on page 81. Suppose that $\{t_n\}$ is a sequence of random variables, θ is a constant, and

$$\sqrt{n}(t_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2) \quad \text{as} \quad n \to \infty$$

Let $g \colon \mathbb{R} \to \mathbb{R}$ be differentiable at θ with $g'(\theta) \neq 0$. Taking a first order Taylor expansion of g around θ, we can write $g(t_n) = g(\theta) + g'(\theta)(t_n - \theta) + R(t_n - \theta)$, where $R(t_n - \theta)$ is a remainder term. It turns out that under these conditions we have $\sqrt{n}R(t_n - \theta) \xrightarrow{p} 0$. The details are omitted. Using this fact, prove carefully that $\sqrt{n}\{g(t_n) - g(\theta)\} \xrightarrow{d} \mathcal{N}(0, g'(\theta)^2\sigma^2)$.

Ex. 2.8.44. Using fact 2.5.1 (page 75) as appropriate, prove the following part of fact 2.6.12: If $\mathbf{X}_n \xrightarrow{p} \mathbf{X}$ and $\mathbf{Y}_n \xrightarrow{p} \mathbf{Y}$, then $\mathbf{X}_n\mathbf{Y}_n \xrightarrow{p} \mathbf{XY}$ whenever the matrices are conformable.

Ex. 2.8.45. Confirm the following claim in fact 2.6.11: If $\mathbf{a}'\mathbf{x}_n \xrightarrow{p} \mathbf{a}'\mathbf{x}$ for every $\mathbf{a} \in \mathbb{R}^K$, then $\mathbf{x}_n \xrightarrow{p} \mathbf{x}$.

Ex. 2.8.46. Let $\{\mathbf{x}_n\}$ be a sequence of vectors in $\mathbb{R}^2$, where $\mathbf{x}_n := (x_n, y_n)$ for each n. Suppose that $\mathbf{x}_n \xrightarrow{p} \mathbf{0}$ (i.e., $x_n \xrightarrow{p} 0$ and $y_n \xrightarrow{p} 0$). Show that $\|\mathbf{x}_n\| \xrightarrow{p} 0$.

Ex. 2.8.47. Verify the first part of fact 2.6.10 on page 86. (Note that exercise 2.8.46 is a warm up to this exercise.)

Ex. 2.8.48. Confirm the claim $\sqrt{N}(\bar{\mathbf{x}}_N - \boldsymbol{\mu}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma)$ in theorem 2.6.1.

Ex. 2.8.49. Let $\{\mathbf{x}_n\}$ be an IID sequence of random vectors in $\mathbb{R}^K$ with $\mathbb{E}[\mathbf{x}_n] = \mathbf{0}$ and $\text{var}[\mathbf{x}_n] = \mathbf{I}_K$. Let

$$\bar{\mathbf{x}}_N := \frac{1}{N}\sum_{n=1}^{N} \mathbf{x}_n \quad \text{and} \quad y_N := N \cdot \|\bar{\mathbf{x}}_N\|^2$$

What is the asymptotic distribution of $\{y_N\}$?

2.8.1 Solutions to Selected Exercises

Solution to Exercise 2.8.1. If A, B and C are disjoint, then $A \cup B$ and C are also disjoint, and $A \cup B \cup C = (A \cup B) \cup C$. As a result, using additivity over pairs,

$$\mathbb{P}(A \cup B \cup C) = \mathbb{P}((A \cup B) \cup C) = \mathbb{P}(A \cup B) + \mathbb{P}(C) = \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C)$$

This result can be extended to an arbitrary number of sets by using induction. □

Solution to Exercise 2.8.2. Pick any sets $A, B \in \mathcal{F}$. To show that

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$$

we start by decomposing A into the union of two disjoint sets: $A = [(A \cup B) \setminus B] \cup (A \cap B)$. Using additivity of $\mathbb{P}$, we then have

$$\mathbb{P}(A) = \mathbb{P}[(A \cup B) \setminus B] + \mathbb{P}(A \cap B)$$

Since $B \subset (A \cup B)$, we can apply part 1 of fact 2.1.1 (page 49) to obtain

$$\mathbb{P}(A) = \mathbb{P}(A \cup B) - \mathbb{P}(B) + \mathbb{P}(A \cap B)$$

Rearranging this expression gives the result that we are seeking. □

Solution to Exercise 2.8.3. First, $\mathbb{P}(C) = 1/3$ as $1 = \mathbb{P}(\Omega) = \mathbb{P}(A \cup B \cup C) = \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) = 1/3 + 1/3 + \mathbb{P}(C)$, and hence $\mathbb{P}(C) = 1/3$. In addition, $\mathbb{P}(A \cup B) = 2/3$, $\mathbb{P}(A \cap B) = 0$, $\mathbb{P}(A^c) = 2/3$, $\mathbb{P}(A^c \cup B^c) = \mathbb{P}((A \cap B)^c) = \mathbb{P}(\Omega) = 1$, and $\mathbb{P}(A \cap C) = 0 \neq 1/9 = \mathbb{P}(A)\mathbb{P}(C)$. Therefore A is not independent of C. □

Solution to Exercise 2.8.4. When the dice is rolled one face must come up, so the sum of the probabilities is one. More formally, letting $\Omega = \{1, \ldots, 6\}$ be the sample space, we have

$$\mathbb{P}\{1, \ldots, 6\} = \mathbb{P} \cup_{m=1}^{6} \{m\} = \sum_{m=1}^{6} \mathbb{P}\{m\} = \sum_{m=1}^{6} qm = 1$$

Solving the last equality for q, we get $q = 1/21$. □

Solution to Exercise 2.8.5. To show that $\mathbb{P}$ is a probability on Ω we need to check that

1. $\mathbb{1}\{\omega_0 \in A\} \in [0,1]$ for every $A \subset \Omega$.
2. $\mathbb{1}\{\omega_0 \in \Omega\} = 1$
3. If $A \cap B = \varnothing$, then $\mathbb{1}\{\omega_0 \in A \cup B\} = \mathbb{1}\{\omega_0 \in A\} + \mathbb{1}\{\omega_0 \in B\}$

1 is immediate from the definition of an indicator function. 2 holds because $\omega_0 \in \Omega$. Regarding 3, pick any disjoint A and B. If $\omega_0 \in A$, then $\omega_0 \notin B$, and we have

$$\mathbb{1}\{\omega_0 \in A \cup B\} = 1 = \mathbb{1}\{\omega_0 \in A\} + \mathbb{1}\{\omega_0 \in B\}$$

If $\omega_0 \in B$, then $\omega_0 \notin A$, and once again we have

$$\mathbb{1}\{\omega_0 \in A \cup B\} = 1 = \mathbb{1}\{\omega_0 \in A\} + \mathbb{1}\{\omega_0 \in B\}$$

Finally, if ω_0 is in neither A nor B, then

$$\mathbb{1}\{\omega_0 \in A \cup B\} = 0 = \mathbb{1}\{\omega_0 \in A\} + \mathbb{1}\{\omega_0 \in B\}$$

We have shown that 1–3 hold, and hence $\mathbb{P}$ is a probability on Ω. □

Solution to Exercise 2.8.6. Suppose that $\mathbb{P}(A) = 0$ and that $B \in \mathcal{F}$. We claim that $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$, or, in this case, $\mathbb{P}(A \cap B) = 0$. Using nonnegativity and monotonicity of $\mathbb{P}$ (fact 2.1.1), we obtain

$$0 \leq \mathbb{P}(A \cap B) \leq P(A) = 0$$

Therefore $\mathbb{P}(A \cap B) = 0$ as claimed.

Now suppose that $\mathbb{P}(A) = 1$. We claim that $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$, or, in this case, $\mathbb{P}(A \cap B) = \mathbb{P}(B)$. In view of fact 2.1.2 on page 50, we have

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cup B)$$

Since $\mathbb{P}(A) = 1$, it suffices to show that $\mathbb{P}(A \cup B) = 1$. This last equality is implied by monotonicity of $\mathbb{P}$, because $1 = \mathbb{P}(A) \leq \mathbb{P}(A \cup B) \leq 1$.

Next, suppose that A is independent of itself. Then $\mathbb{P}(A) = \mathbb{P}(A \cap A) = \mathbb{P}(A)\mathbb{P}(A) = \mathbb{P}(A)^2$. If $a = a^2$, then either $a = 0$ or $a = 1$.

Finally, let A and B be independent. We have

$$\mathbb{P}(A^c \cap B^c) = \mathbb{P}((A \cup B)^c) = 1 - \mathbb{P}(A \cup B)$$

Applying fact 2.1.2 and independence, we can transform the right-hand side to obtain

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

In other words, A^c and B^c are independent. □

Solution to Exercise 2.8.7. The proof is almost identical to the proof of additivity in example 2.1.3 (page 49). □

Solution to Exercise 2.8.8. The proof of independence is essentially the same as the proof of independence of A and B in example 2.1.5 (page 51). □

Solution to Exercise 2.8.10. We want to show that $\mathbb{E}[\alpha x] = \alpha \mathbb{E}[x]$ is valid. To see this, observe that

$$\alpha x(\omega) = \alpha \left[\sum_{j=1}^{J} s_j \mathbb{1}_{A_j}(\omega) \right] = \sum_{j=1}^{J} \alpha s_j \mathbb{1}_{A_j}(\omega)$$

Hence, applying (2.8),

$$\mathbb{E}[\alpha x] = \sum_{j=1}^{J} \alpha s_j \mathbb{P}(A_j) = \alpha \left[\sum_{j=1}^{J} s_j \mathbb{P}(A_j) \right] = \alpha \mathbb{E}[x]$$

□

Solution to Exercise 2.8.11. Consider the sum $x + y$. By this, we mean the random variable $(x+y)(\omega) := x(\omega) + y(\omega)$. We claim that $\mathbb{E}[x+y] = \mathbb{E}[x] + \mathbb{E}[y]$. To see that this is the case, note first that

$$(x+y)(\omega) = \mathbb{1}_{A \setminus B}(\omega) + \mathbb{1}_{B \setminus A}(\omega) + 2\mathbb{1}_{A \cap B}(\omega)$$

(To check this, just go through the different cases for ω, and verify that the right hand side of this expression agrees with $x(\omega) + y(\omega)$. Sketching a Venn diagram will help.) Therefore, by the definition of expectation,

$$\mathbb{E}[x+y] = \mathbb{P}(A \setminus B) + \mathbb{P}(B \setminus A) + 2\mathbb{P}(A \cap B) \tag{2.41}$$

Now observe that $A = (A \setminus B) \cup (A \cap B)$ and hence, by disjointness,

$$\mathbb{E}[x] := \mathbb{P}(A) = \mathbb{P}(A \setminus B) + \mathbb{P}(A \cap B)$$

Performing a similar calculation with y produces

$$\mathbb{E}[y] := \mathbb{P}(B) = \mathbb{P}(B \setminus A) + \mathbb{P}(A \cap B)$$

Adding these two produces the value on the right-hand side of (2.41), and we have confirmed that $\mathbb{E}[x+y] = \mathbb{E}[x] + \mathbb{E}[y]$. □

Solution to Exercise 2.8.12. We are assuming that x has finite range, and hence takes only finitely many different values. Let m be the largest such value. For this m, we have

$$\lim_{s\to\infty} F_x(s) \geq F_x(m) = \mathbb{P}\{\omega \in \Omega : x(\omega) \leq m\} = \mathbb{P}(\Omega) = 1$$

(The inequality is due to the fact that F_x is increasing.) On the other hand,

$$\lim_{s\to\infty} F_x(s) = \lim_{s\to\infty} \mathbb{P}\{x \leq s\} \leq \lim_{s\to\infty} \mathbb{P}(\Omega) = 1$$

From these two inequalities we get $1 \leq \lim_{s\to\infty} F_x(s) \leq 1$, which is equivalent to $\lim_{s\to\infty} F_x(s) = 1$. □

Solution to Exercise 2.8.14. Fix $s \geq 0$. Using additivity over disjoint sets, we have

$$F_{|x|}(s) := \mathbb{P}\{|x| \leq s\} = \mathbb{P}\{-s \leq x \leq s\} = \mathbb{P}\{x = -s\} + \mathbb{P}\{-s < x \leq s\}$$

By assumption, $\mathbb{P}\{x = -s\} = 0$. Applying fact 2.3.1 on page 62 then yields

$$F_{|x|}(s) = \mathbb{P}\{-s < x \leq s\} = F(s) - F(-s)$$

The claim $F_{|x|}(s) = 2F(s) - 1$ now follows from the definition of symmetry. □

Solution to Exercise 2.8.15. That $0 \leq p_j \leq 1$ for each j follows immediately from the definition of $\mathbb{P}$. In addition, using additivity of $\mathbb{P}$, we have

$$\sum_{j=1}^{J} p_j = \sum_{j=1}^{J} \mathbb{P}\{x = s_j\} = \mathbb{P} \cup_{j=1}^{J} \{x = s_j\} = \mathbb{P}(\Omega) = 1 \tag{2.42}$$

(We are using the fact that the sets $\{x = s_j\}$ disjoint. Why is this always true? Look carefully at the definition of a function given in §4.2.) □

Solution to Exercise 2.8.16. Let $z := G^{-1}(u)$. We want to show that $z \sim G$. Since G is monotone increasing we have $G(a) \leq G(b)$ whenever $a \leq b$. As a result, for any $s \in \mathbb{R}$,

$$\mathbb{P}\{z \leq s\} = \mathbb{P}\{G^{-1}(u) \leq s\} = \mathbb{P}\{G(G^{-1}(u)) \leq G(s)\} = \mathbb{P}\{u \leq G(s)\} = G(s)$$

We have shown that $z \sim G$ as claimed. □

Solution to Exercise 2.8.17. Evidently $G(s) = 0$ when $s < 0$. For $s \geq 0$ we have

$$\mathbb{P}\{x^2 \leq s\} = \mathbb{P}\{|x| \leq \sqrt{s}\} = \mathbb{P}\{x \leq \sqrt{s}\} = F(\sqrt{s})$$

Thus, $G(s) = F(\sqrt{s})\mathbb{1}\{s \geq 0\}$. □

Solution to Exercise 2.8.20. If $x(\omega) := \mathbb{1}\{\omega \in A\} \leq \mathbb{1}\{\omega \in B\} =: y(\omega)$ for any $\omega \in \Omega$, then $A \subset B$. (If $\omega \in A$, then $x(\omega) = 1$. Since $x(\omega) \leq y(\omega) \leq 1$, we then have $y(\omega) = 1$, and hence $\omega \in B$.) Using fact 2.2.2 and monotonicity of $\mathbb{P}$, we then have

$$\mathbb{E}[x] = \mathbb{E}[\mathbb{1}\{\omega \in A\}] = \mathbb{P}(A) \leq \mathbb{P}(B) = \mathbb{E}[\mathbb{1}\{\omega \in B\}] = \mathbb{E}[y]$$

as was to be shown. □

Solution to Exercise 2.8.21. Let a be any nonnegative number, and let $j \leq k$. If $a \geq 1$, then $a^j \leq a^k$. If $a < 1$, then $a^j \leq 1$. Thus, for any $a \geq 0$, we have $a^j \leq a^k + 1$, and for any random variable x we have $|x|^j \leq |x|^k + 1$. Using monotonicity of expectations (fact 2.2.5 on page 60) and $\mathbb{E}[1] = 1$, we then have $\mathbb{E}[|x|^j] \leq \mathbb{E}[|x^k|] + 1$. Hence the j-th moment exists whenever the k-th moment exists. □

Solution to Exercise 2.8.22. We have

$$\begin{aligned}
\operatorname{var}\left[\sum_{n=1}^N \alpha_n x_n\right] &= \mathbb{E}\left[\left(\sum_{n=1}^N \alpha_n x_n - \mathbb{E}\left[\sum_{n=1}^N \alpha_n x_n\right]\right)^2\right] \\
&= \mathbb{E}\left[\left(\sum_{n=1}^N \alpha_n (x_n - \mathbb{E}[x_n])\right)^2\right] \\
&= \mathbb{E}\left[\sum_{n=1}^N \alpha_n^2 (x_n - \mathbb{E}[x_n])^2 + 2\sum_{n<m} \alpha_n \alpha_m (x_n - \mathbb{E}[x_n])(x_m - \mathbb{E}[x_m])\right] \\
&= \sum_{n=1}^N \alpha_n^2 \operatorname{var}[x_n] + \sum_{n<m} \alpha_n \alpha_m \operatorname{cov}[x_n, x_m]
\end{aligned}$$

as required. □

Solution to Exercise 2.8.26. Let $u := f(x) = 2x$ and $v := g(y) = 3y - 1$, where x and y are independent. Independence of u and v can be confirmed via (2.24) on page 71. Fixing s_1 and s_2 in $\mathbb{R}$, we have

$$\begin{aligned}
\mathbb{P}\{u \leq s_1, v \leq s_2\} &= \mathbb{P}\{x \leq s_1/2, y \leq (s_2 + 1)/3\} \\
&= \mathbb{P}\{x \leq s_1/2\}\mathbb{P}\{y \leq (s_2 + 1)/3\} = \mathbb{P}\{u \leq s_1\}\mathbb{P}\{v \leq s_2\}
\end{aligned}$$

Thus u and v are independent as claimed. □

Solution to Exercise 2.8.27. As in the statement of the exercise, x and y are independent uniform random variables on $[0,1]$, $z := \max\{x,y\}$ and $w := \min\{x,y\}$. As a first step to the proofs, you should convince yourself that if a, b and c are three numbers, then

- $\max\{a,b\} \leq c$ if and only if $a \leq c$ and $b \leq c$
- $\min\{a,b\} \leq c$ if and only if $a \leq c$ or $b \leq c$

Using these facts, next convince yourself that, for any $s \in \mathbb{R}$,

- $\{z \leq s\} = \{x \leq s\} \cap \{y \leq s\}$
- $\{w \leq s\} = \{x \leq s\} \cup \{y \leq s\}$

(For each, equality, show that if ω is in the right-hand side, then ω is in the left-hand side, and vice versa.) Now, for $s \in [0,1]$, we have

$$\mathbb{P}\{z \leq s\} = \mathbb{P}[\{x \leq s\} \cap \{y \leq s\}] = \mathbb{P}\{x \leq s\}\mathbb{P}\{y \leq s\} = s^2$$

By differentiating we get the density $p(s) = 2s$, and by integrating $\int_0^1 sp(s)ds$ we get $\mathbb{E}[z] = 2/3$. Finally, regarding the cdf of w, for $s \in [0,1]$ we have

$$\begin{aligned}\mathbb{P}\{w \leq s\} &= \mathbb{P}[\{x \leq s\} \cup \{y \leq s\}] \\ &= \mathbb{P}\{x \leq s\} + \mathbb{P}\{y \leq s\} - \mathbb{P}[\{x \leq s\} \cap \{y \leq s\}]\end{aligned}$$

Hence $\mathbb{P}\{w \leq s\} = 2s - s^2$. □

Solution to Exercise 2.8.31. Using $y = \ell^*(x) + u$ and the results form exercise 2.8.30, we have

$$\begin{aligned}\operatorname{var}[\ell^*(x) + u] &= \operatorname{var}[y] \\ &= \operatorname{corr}[x,y]^2 \operatorname{var}[y] + (1 - \operatorname{corr}[x,y]^2) \operatorname{var}[y] \\ &= \operatorname{var}[\ell^*(x)] + \operatorname{var}[u]\end{aligned}$$

It follows (why?) that $\operatorname{cov}[\ell^*(x), u] = 0$ as claimed. □

Solution to Exercise 2.8.35. First note that since $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_N)$ we have $\operatorname{cov}[x_i, x_j] = 0$ for all $i \neq j$. Since uncorrelated normal random variables are independent, we then have $x_1, \ldots, x_N \overset{\text{IID}}{\sim} \mathcal{N}(0,1)$. Since sums of squares of independent standard normals are chi-squared, we have in particular that

$$\sum_{n=1}^{k} x_n^2 \sim \chi^2(k) \tag{2.43}$$

for any $k \leq N$. The solutions to the exercise can now be given:

1. Yes, for the reason just described.
2. $x_1^2 \sim \chi^2(1)$ by (2.43)
3. $x_1^2 / x_2^2 \sim F(1,1)$, because if $Q_1 \sim \chi^2(k_1)$ and $Q_2 \sim \chi^2(k_2)$ and Q_1 and Q_2 and independent, then $(Q_1/k_1)/(Q_2/k_2) \sim F(k_1, k_2)$.
4. $x_1[2/(x_2^2 + x_3^2)]^{1/2} \sim t(2)$, because if $Z \sim \mathcal{N}(0,1)$ and $Q \sim \chi^2(k)$ and Z and Q are independent, then $Z(k/Q)^{1/2} \sim t(k)$.
5. $\|\mathbf{x}\|^2 = \sum_{n=1}^N x_N^2 \sim \chi^2(N)$ by (2.43).
6. Linear combinations of normals are normal, so $y := \mathbf{a}'\mathbf{x}$ is normal. Evidently $\mathbb{E}[y] = \mathbb{E}[\mathbf{a}'\mathbf{x}] = \mathbf{a}'\mathbb{E}[\mathbf{x}] = 0$. Using independence, we obtain

$$\operatorname{var}[y] = \sum_{n=1}^{N} a_n^2 \operatorname{var}[x_n] = \sum_{n=1}^{N} a_n^2$$

Hence $y \sim \mathcal{N}(0, \sum_{n=1}^N a_n^2)$.

□

Solution to Exercise 2.8.36. Pick any nonnegative random variable x and $\delta > 0$. By considering what happens at an arbitrary $\omega \in \Omega$, you should be able to convince yourself that

$$x = \mathbb{1}\{x \geq \delta\}x + \mathbb{1}\{x < \delta\}x \geq \mathbb{1}\{x \geq \delta\}\delta$$

Using fact 2.2.5 (page 60), fact 2.2.2 (page 58) and rearranging gives (2.11). Regarding (2.12), observe that

$$x^2 = \mathbb{1}\{|x| \geq \delta\}x^2 + \mathbb{1}\{|x| < \delta\}x^2 \geq \mathbb{1}\{|x| \geq \delta\}\delta^2$$

Proceding as before leads to (2.12). □

Solution to Exercise 2.8.37. From the definition of convergence in probability (see §2.5.1), the statement $x_n \xrightarrow{p} 0$ means that, given any $\delta > 0$, we have $\mathbb{P}\{|x_n| > \delta\} \to 0$. Consider first the case where $\mathbb{P}\{y = -1\} = \mathbb{P}\{y = 1\} = 0.5$. Take $\delta = 0.5$. Then, since $x_n = y$ for all n,

$$\mathbb{P}\{|x_n| > \delta\} = \mathbb{P}\{|y| > 0.5\} = 1$$

Thus, the sequence does not converge to zero. Hence $x_n \xrightarrow{p} 0$ fails. On the other hand, if $\mathbb{P}\{y = 0\} = 1$, then for any $\delta > 0$ we have

$$\mathbb{P}\{|x_n| > \delta\} = \mathbb{P}\{|y| > \delta\} = 0$$

This sequence does converge to zero (in fact it's constant at zero), and $x_n \xrightarrow{p} 0$ holds. □

Solution to Exercise 2.8.38. We want to give an example of a sequence of random variables $\{x_n\}$ and random variable x such that x_n converges to x in distribution, but not in probability. Many examples can be found by using IID sequences. For example, if $\{x_n\}_{n=1}^{\infty}$ and x are IID standard normal random variables, then x_n and x have the same distribution for all n, and hence x_n converges in distribution to x. However, $z_n := x_n - x$ has distribution $\mathcal{N}(0, 2)$ for all n. Letting z be any random variable with distribution $\mathcal{N}(0, 2)$ and δ be any strictly positive constant, we have $\mathbb{P}\{|x_n - x| \geq \delta\} = \mathbb{P}\{|z| \geq \delta\} > 0$. Thus, $\mathbb{P}\{|x_n - x| \geq \delta\}$ does not converge to zero. □

Solution to Exercise 2.8.39. By linearity of expectations,

$$\mathbb{E}[\bar{x}_N] = \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}[x_n] = \frac{N}{N}\int sF(ds) = \int sF(ds)$$

This confirms that $\mathbb{E}[\bar{x}_N] \to \int sF(ds)$ as claimed. To see that $\operatorname{var}[\bar{x}_N] \to 0$ as $N \to \infty$, let σ^2 be the common variance of each x_n. Using fact 2.4.8, we obtain

$$\operatorname{var}\left[\frac{1}{N}\sum_{n=1}^{N} x_n\right] = \frac{1}{N^2}\sum_{n=1}^{N} \sigma^2 + \frac{2}{N^2}\sum_{n<m} \operatorname{cov}[x_n, x_m]$$

By independence, this reduces to $\operatorname{var}[\bar{x}_N] = \sigma^2/N$, which converges to zero. □

Solution to Exercise 2.8.44. Let $\mathbf{X}_n \xrightarrow{p} \mathbf{X}$ and $\mathbf{Y}_n \xrightarrow{p} \mathbf{Y}$. To prove that $\mathbf{X}_n\mathbf{Y}_n \xrightarrow{p} \mathbf{XY}$, we need to show that the i, j-th element of $\mathbf{X}_n\mathbf{Y}_n$ converges in probability to the i, j-th element of $\mathbf{XY}$. By hypothesis, we have

$$x_{ik}^n \xrightarrow{p} x_{ik} \quad \text{and} \quad y_{kj}^n \xrightarrow{p} y_{kj} \quad \text{for all} \quad k$$

Applying fact 2.5.1 on page 75 twice, we obtain

$$x_{ik}^n y_{kj}^n \xrightarrow{p} x_{ik} y_{kj} \quad \text{for all} \quad k$$

and then

$$\sum_k x_{ik}^n y_{kj}^n \xrightarrow{p} \sum_k x_{ik} y_{kj}$$

In other words, the i, j-th element of $\mathbf{X}_n\mathbf{Y}_n$ converges in probability to the i, j-th element of $\mathbf{XY}$. □

Solution to Exercise 2.8.45. If $\mathbf{a}'\mathbf{x}_n \xrightarrow{p} \mathbf{a}'\mathbf{x}$ for every $\mathbf{a} \in \mathbb{R}^K$, then we know in particular that this convergence holds for the canonical basis vectors. Hence

$$\mathbf{e}_k'\mathbf{x}_n \xrightarrow{p} \mathbf{e}_k'\mathbf{x} \quad \text{for every } k$$

$$\therefore \quad x_n^k \xrightarrow{p} x^k \quad \text{for every } k \quad \text{(elementwise convergence)}$$

$$\therefore \quad \mathbf{x}_n \xrightarrow{p} \mathbf{x} \quad \text{(vector convergence, by definition)}$$

□

Solution to Exercise 2.8.46. From fact 2.5.1 on page 75, we know that if $g\colon \mathbb{R} \to \mathbb{R}$ is continuous and $\{u_n\}$ is a scalar sequence of random variables with $u_n \xrightarrow{p} u$, then $g(u_n) \xrightarrow{p} g(u)$. We also know that if $u_n \xrightarrow{p} u$ and $v_n \xrightarrow{p} v$, then $u_n + v_n \xrightarrow{p} u + v$. By assumption, we have

$$x_n \xrightarrow{p} 0 \quad \text{and} \quad y_n \xrightarrow{p} 0$$

$$\therefore \quad x_n^2 \xrightarrow{p} 0^2 = 0 \quad \text{and} \quad y_n^2 \xrightarrow{p} 0^2 = 0$$

$$\therefore \quad \|\mathbf{x}_n\|^2 = x_n^2 + y_n^2 \xrightarrow{p} 0 + 0 = 0$$

$$\therefore \quad \|\mathbf{x}_n\| = \sqrt{\|\mathbf{x}_n\|^2} \xrightarrow{p} \sqrt{0} = 0$$

□

Solution to Exercise 2.8.47. Let $\{\mathbf{x}_n\}$ be a sequence of random vectors in $\mathbb{R}^K$ and $\mathbf{x}$ be a random vector in $\mathbb{R}^K$. We need to show that

$$x_k^n \xrightarrow{p} x_k \text{ for all } k \iff \|\mathbf{x}_n - \mathbf{x}\| \xrightarrow{p} 0$$

A special case of this argument can be found in the solution to exercise 2.8.46. The general case is similar: Suppose first that $x_k^n \xrightarrow{p} x_k$ for all k. Combining the various results about scalar convergence in probability in fact 2.5.1 (page 75), one can then verify (details left to you) that

$$\|\mathbf{x}_n - \mathbf{x}\| := \sqrt{\sum_{k=1}^K (x_k^n - x_k)^2} \xrightarrow{p} 0 \qquad (n \to \infty)$$

Regarding the converse, suppose now that $\|\mathbf{x}_n - \mathbf{x}\| \xrightarrow{p} 0$. Fix $\epsilon > 0$ and arbitrary k. From the definition of the norm we see that $|x_k^n - x_k| \leq \|\mathbf{x}_n - \mathbf{x}\|$ is always true, and hence

$$\begin{aligned} |x_k^n - x_k| > \epsilon &\implies \|\mathbf{x}_n - \mathbf{x}\| > \epsilon \\ \therefore \quad \{|x_k^n - x_k| > \epsilon\} &\subset \{\|\mathbf{x}_n - \mathbf{x}\| > \epsilon\} \\ \therefore \quad 0 \leq \mathbb{P}\{|x_k^n - x_k| > \epsilon\} &\leq \mathbb{P}\{\|\mathbf{x}_n - \mathbf{x}\| > \epsilon\} \to 0 \end{aligned}$$

The proof is done. □

Solution to Exercise 2.8.48. Define

$$\mathbf{z}_n := \sqrt{N}\,(\bar{\mathbf{x}}_N - \boldsymbol{\mu}) \quad \text{and} \quad \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \Sigma)$$

We need to show that $\mathbf{z}_n \xrightarrow{d} \mathbf{z}$. To do this, we apply the Cramer-Wold device (fact 2.6.11, page 87) and the scalar CLT (theorem 2.5.2, page 80). To begin, fix $\mathbf{a} \in \mathbb{R}^K$. Observe that

$$\mathbf{a}'\mathbf{z}_n := \sqrt{N}\,(\bar{y}_n - \mathbb{E}\,[y_n])$$

where $y_n := \mathbf{a}'\mathbf{x}_n$. Since y_n is IID (in particular, functions of independent random variables are independent) and

$$\operatorname{var}[y_n] = \operatorname{var}[\mathbf{a}'\mathbf{x}_n] = \mathbf{a}' \operatorname{var}[\mathbf{x}_n]\mathbf{a} = \mathbf{a}'\Sigma\mathbf{a}$$

the scalar CLT yields

$$\mathbf{a}'\mathbf{z}_n \xrightarrow{d} \mathcal{N}(0, \mathbf{a}'\Sigma\mathbf{a})$$

Since $\mathbf{a}'\mathbf{z} \sim \mathcal{N}(0, \mathbf{a}'\Sigma\mathbf{a})$, we have shown that $\mathbf{a}'\mathbf{z}_n \xrightarrow{d} \mathbf{a}'\mathbf{z}$. Since $\mathbf{a}$ was arbitrary, the Cramer-Wold device tells us that $\mathbf{z}_n$ converges in distribution to $\mathbf{z}$. □

Solution to Exercise 2.8.49. By assumption, $\{\mathbf{x}_n\}$ is an IID sequence in $\mathbb{R}^K$ with $\mathbb{E}[\mathbf{x}_n] = \mathbf{0}$ and $\text{var}[\mathbf{x}_n] = \mathbf{I}_K$. It follows from the vector central limit theorem that

$$\sqrt{N}\bar{\mathbf{x}}_N \xrightarrow{d} \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_K)$$

Letting $g(\mathbf{s}) := \|\mathbf{s}\|^2$ and applying the continuous mapping theorem (fact 2.6.13 on page 88), we obtain

$$y_N = \|\sqrt{N}\bar{\mathbf{x}}_N\|^2 \xrightarrow{d} \|\mathbf{z}\|^2 = \sum_{k=1}^{K} z_k^2$$

From fact 2.4.4 on page 72 we conclude that $y_N \xrightarrow{d} \chi^2(K)$. □

Chapter 3

Orthogonality and Projections

[roadmap]

3.1 Orthogonality

[roadmap]

3.1.1 Definition and Basic Properties

Let $\mathbf{x}$ and $\mathbf{z}$ be vectors in $\mathbb{R}^N$. If $\mathbf{x}'\mathbf{z} = 0$, then we write $\mathbf{x} \perp \mathbf{z}$ and call $\mathbf{x}$ and $\mathbf{z}$ **orthogonal**. In $\mathbb{R}^2$, $\mathbf{x}$ and $\mathbf{z}$ are orthogonal when they are perpendicular to one another, as in figure 3.1. For $\mathbf{x} \in \mathbb{R}^N$ and $S \subset \mathbb{R}^N$, we say that **$\mathbf{x}$ is orthogonal to S** if $\mathbf{x} \perp \mathbf{z}$ for all $\mathbf{z} \in S$ (figure 3.2). In this case we write $\mathbf{x} \perp S$.

Fact 3.1.1. If $B \subset S$ with $\text{span}(B) = S$, then $\mathbf{x} \perp S$ if and only if $\mathbf{x} \perp \mathbf{b}$ for all $\mathbf{b} \in B$.

A set of vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_K\} \subset \mathbb{R}^N$ is called an **orthogonal set** if its elements are mutually orthogonal; that is, if $\mathbf{x}_j \perp \mathbf{x}_k$ whenever j and k are not equal.

Fact 3.1.2 (Pythagorian law)**.** If $\{\mathbf{x}_1, \ldots, \mathbf{x}_K\}$ is an orthogonal set, then

$$\|\mathbf{x}_1 + \cdots + \mathbf{x}_K\|^2 = \|\mathbf{x}_1\|^2 + \cdots + \|\mathbf{x}_K\|^2$$

Orthogonal sets and linear independence are closely related. In particular,

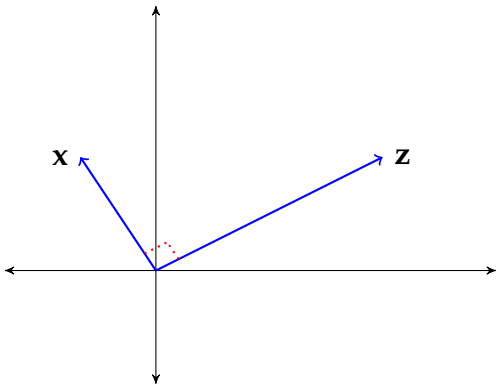


Figure 3.1: $\mathbf{x} \perp \mathbf{z}$

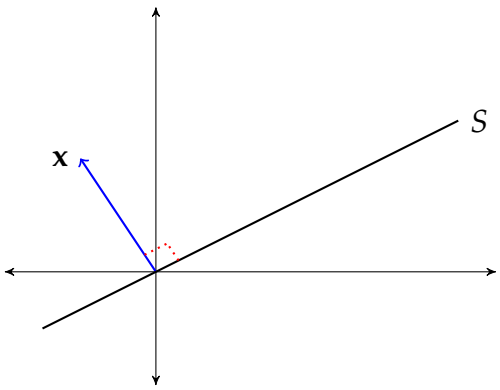


Figure 3.2: $\mathbf{x} \perp S$

Fact 3.1.3. Let $O = \{\mathbf{x}_1, \ldots, \mathbf{x}_K\} \subset \mathbb{R}^N$ be an orthogonal set. If $\mathbf{0} \notin O$, then O is linearly independent.

While not every linearly independent set is orthogonal, a partial converse to fact 3.1.3 is given in §3.3.3.

A set of vectors $O \subset \mathbb{R}^N$ is called an **orthonormal set** if it is an orthogonal set and $\|\mathbf{u}\| = 1$ for all $\mathbf{u} \in O$. An orthonormal set that spans a linear subspace S of $\mathbb{R}^N$ is called an **orthonormal basis** for S. The standard example of an orthonormal basis for all of $\mathbb{R}^N$ is the canonical basis vectors $\mathbf{e}_1, \ldots, \mathbf{e}_N$.

One neat thing about orthonormal bases is the following: If $O = \{\mathbf{u}_1, \ldots, \mathbf{u}_K\}$ is any basis of $S \subset \mathbb{R}^N$, then we can write $\mathbf{x}$ in terms of the basis vectors as in $\mathbf{x} = \sum_{k=1}^K \alpha_k \mathbf{u}_k$ for suitable scalars $\alpha_1, \ldots, \alpha_K$. The value of these scalars is not always transparent, but for an orthonormal basis we have $\alpha_k = \mathbf{x}'\mathbf{u}_k$ for each k. That is,

Fact 3.1.4. If $\{\mathbf{u}_1, \ldots, \mathbf{u}_K\}$ is an orthonormal set and $\mathbf{x} \in \operatorname{span}\{\mathbf{u}_1, \ldots, \mathbf{u}_K\}$, then

$$\mathbf{x} = \sum_{k=1}^{K} (\mathbf{x}'\mathbf{u}_k)\mathbf{u}_k \tag{3.1}$$

The proof is an exercise.

An **orthogonal matrix** is an $N \times N$ matrix $\mathbf{Q}$ such that the columns of $\mathbf{Q}$ form an orthonormal set. If $\mathbf{Q}$ is an orthonormal matrix then the columns of $\mathbf{Q}$ must form an orthonormal basis for $\mathbb{R}^N$. (Why?) Since the columns of an orthogonal matrix $\mathbf{Q}$ are linearly independent, we know that $\mathbf{Q}$ is invertible. Indeed, the inverse is just the transpose. This and other useful facts are collected below:

Fact 3.1.5. Let $\mathbf{Q}$ and $\mathbf{P}$ be orthogonal $N \times N$ matrices. The following statements are true:

1. $\mathbf{Q}^{-1} = \mathbf{Q}'$, and $\mathbf{Q}'$ is also orthogonal.
2. $\mathbf{QP}$ is an orthogonal matrix.
3. $\det(\mathbf{Q}) \in \{-1, 1\}$.

3.1.2 Orthogonal Decompositions

Recall from §1.4.4 that an $N \times N$ matrix $\mathbf{A}$ is diagonalizable if it is similar to a diagonal matrix, and this occurs precisely when the eigenvectors of $\mathbf{A}$ span $\mathbb{R}^N$. In particular, by theorem 1.4.1 on page 33, we have $\mathbf{A} = \mathbf{P}\mathbf{D}\mathbf{P}^{-1}$, where $(\text{col}_n(\mathbf{P}), \lambda_n)$ is an eigenpair of $\mathbf{A}$ for each n. There is one important case where we can say more:

Theorem 3.1.1. *If $\mathbf{A}$ is a symmetric $N \times N$ matrix, then an orthonormal basis of $\mathbb{R}^N$ can be formed from the eigenvectors of $\mathbf{A}$. In particular, $\mathbf{A}$ can be diagonalized as $\mathbf{A} = \mathbf{Q}\mathbf{D}\mathbf{Q}'$, where $\mathbf{Q}$ is the orthogonal matrix constructed from this orthonormal basis, and $\mathbf{D}$ is the diagonal matrix formed from the N eigenvalues of $\mathbf{A}$.*

This theorem is sometimes called the **spectral decomposition theorem**. One nice application is a proof of fact 1.4.13 on page 37, which states among other things that symmetric matrix $\mathbf{A}$ is positive definite if and only if its eigenvalues are all positive. Exercise 3.6.5 and its solution step you through the arguments.

There is another kind of matrix decomposition using orthogonality that has many important applications: the **QR decomposition**.

Theorem 3.1.2. *If $\mathbf{A}$ is an $N \times K$ matrix with linearly independent columns, then there exists an invertible upper triangular $K \times K$ matrix $\mathbf{R}$ and an $N \times K$ matrix $\mathbf{Q}$ with orthonormal columns such that $\mathbf{A} = \mathbf{Q}\mathbf{R}$.*

Remark: $\mathbf{Q}$ isn't referred to as an orthogonal matrix here because it isn't in general square. We'll give a constructive proof of theorem 3.1.2 when we get to §3.3.3.

3.1.3 Orthogonal Complements

Given $S \subset \mathbb{R}^N$, the **orthogonal complement** of S is defined as

$$S^\perp := \{\mathbf{x} \in \mathbb{R}^N : \mathbf{x} \perp S\}$$

In other words, $S^\perp$ is the set of all vectors that are orthogonal to S. Figure 3.3 gives an example in $\mathbb{R}^2$.

Fact 3.1.6. For any nonempty $S \subset \mathbb{R}^N$, the set $S^\perp$ is a linear subspace of $\mathbb{R}^N$.

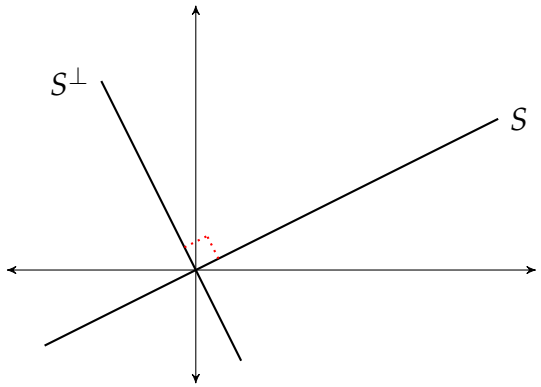


Figure 3.3: Orthogonal complement of S

This is easy enough to confirm: Looking back at the definition of linear subspaces, we see that the following statement must be verified: Given $\mathbf{x}, \mathbf{y} \in S^\perp$ and $\alpha, \beta \in \mathbb{R}$, the vector that $\alpha\mathbf{x} + \beta\mathbf{y}$ is also in $S^\perp$. Clearly this is the case, because if $\mathbf{z} \in S$, then

$$\begin{aligned}(\alpha\mathbf{x} + \beta\mathbf{y})'\mathbf{z} &= \alpha\mathbf{x}'\mathbf{z} + \beta\mathbf{y}'\mathbf{z} && (\because \text{ linearity of inner products}) \\ &= \alpha \times 0 + \beta \times 0 = 0 && (\because \ \mathbf{x}, \mathbf{y} \in S^\perp \text{ and } \mathbf{z} \in S)\end{aligned}$$

We have shown that $\alpha\mathbf{x} + \beta\mathbf{y} \perp \mathbf{z}$ for any $\mathbf{z} \in S$, thus confirming that $\alpha\mathbf{x} + \beta\mathbf{y} \in S^\perp$.

Fact 3.1.7. For $S \subset \mathbb{R}^N$, we have $S \cap S^\perp = \{\mathbf{0}\}$.

3.2 Orthogonal Projections

[Roadmap]

3.2.1 The Orthogonal Projection Theorem

One problem that comes up in many different contexts is approximation of an element $\mathbf{y}$ of $\mathbb{R}^N$ by an element of a given subspace S of $\mathbb{R}^N$. Stated more precisely, the problem is, given $\mathbf{y}$ and S, to find the closest element $\hat{\mathbf{y}}$ of S to $\mathbf{y}$. Closeness is in terms of Euclidean norm, so $\hat{\mathbf{y}}$ is the minimizer of $\|\mathbf{y} - \mathbf{z}\|$ over all $\mathbf{z} \in S$:

$$\hat{\mathbf{y}} = \operatorname*{argmin}_{\mathbf{z} \in S} \|\mathbf{y} - \mathbf{z}\| \tag{3.2}$$

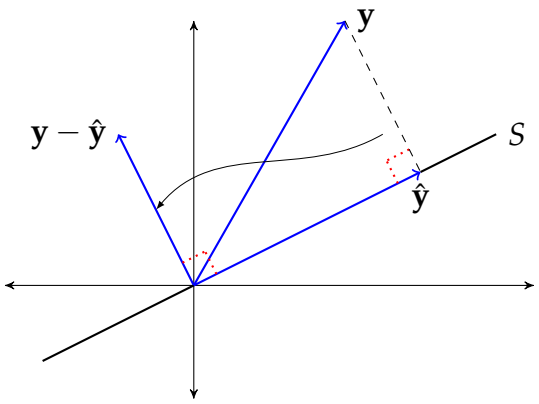


Figure 3.4: Orthogonal projection

Existence of a minimizer is not immediately obvious, suggesting that $\hat{\mathbf{y}}$ may not be well-defined. However, it turns out that we need not be concerned, as $\hat{\mathbf{y}}$ always exists (given any S and $\mathbf{y}$). The next theorem states this fact, as well as providing a way to identify $\hat{\mathbf{y}}$.

Theorem 3.2.1 (Orthogonal Projection Theorem I)**.** *Let $\mathbf{y} \in \mathbb{R}^N$ and let S be any nonempty linear subspace of $\mathbb{R}^N$. The following statements are true:*

1. *The optimization problem (3.2) has exactly one solution.*
2. *Vector $\hat{\mathbf{y}} \in \mathbb{R}^N$ is the unique solution to (3.2) if and only if $\hat{\mathbf{y}} \in S$ and $\mathbf{y} - \hat{\mathbf{y}} \perp S$.*

The vector $\hat{\mathbf{y}}$ in theorem 3.2.1 is called the **orthogonal projection of y onto S**. Although we do not prove the theorem here, the intuition is easy to grasp from a graphical presentation. Figure 3.4 illustrates. Looking at the figure, we can see that the closest point $\hat{\mathbf{y}}$ to $\mathbf{y}$ within S is indeed the one and only point in S such that $\mathbf{y} - \hat{\mathbf{y}}$ is orthogonal to S.

[Add a proof?]

Example 3.2.1. Let $\mathbf{y} \in \mathbb{R}^N$ and let $\mathbf{1} \in \mathbb{R}^N$ be the vector of ones. Let S be the set of constant vectors in $\mathbb{R}^N$, meaning that all elements are equal. Evidently S is the span of $\{\mathbf{1}\}$. The orthogonal projection of $\mathbf{y}$ onto S is $\hat{\mathbf{y}} := \bar{y}\mathbf{1}$, where $\bar{y}$ is the scalar

$$\bar{y} := \frac{1}{N} \sum_{n=1}^{N} y_n$$

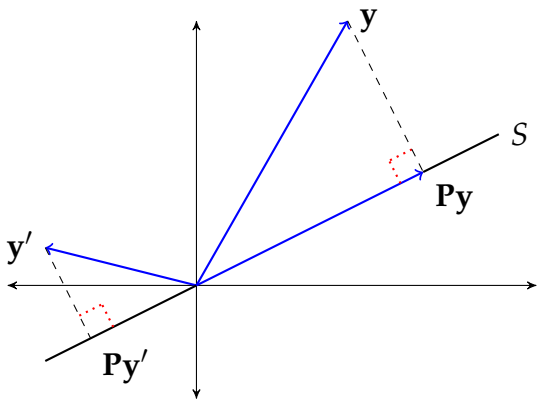


Figure 3.5: Orthogonal projection under **P**

formed by averaging the elements of $\mathbf{y}$. Since $\hat{\mathbf{y}} \in S$ clearly holds, to verify claim this we only need to check that $\mathbf{y} - \hat{\mathbf{y}}$ is orthogonal to S, for which it suffices to show that $(\mathbf{y} - \hat{\mathbf{y}})'\mathbf{1} = 0$. This is true because

$$(\mathbf{y} - \hat{\mathbf{y}})'\mathbf{1} = \mathbf{y}'\mathbf{1} - \hat{\mathbf{y}}'\mathbf{1} = \sum_{n=1}^{N} y_n - \bar{y}\mathbf{1}'\mathbf{1} = \sum_{n=1}^{N} y_n - \frac{1}{N}\sum_{n=1}^{N} y_n N = 0$$

3.2.2 Orthogonal Projection as a Mapping

Holding S fixed, we can think of the operation

$$\mathbf{y} \;\mapsto\; \text{the orthogonal projection of } \mathbf{y} \text{ onto } S$$

as a function from $\mathbb{R}^N$ to $\mathbb{R}^N$. The function is typically denoted by $\mathbf{P}$, so that, for each $\mathbf{y} \in \mathbb{R}^N$, the symbol $\mathbf{Py}$ represents image of $\mathbf{y}$ under $\mathbf{P}$, which is the orthogonal projection $\hat{\mathbf{y}}$. (We should perhaps write $\mathbf{P}_S$ but to simplify notation we'll stick with $\mathbf{P}$. Hopefully the subspace that determines $\mathbf{P}$ will be clear from context.) In general, $\mathbf{P}$ is called the **orthogonal projection onto** S. Figure 3.5 illustrates the action of $\mathbf{P}$ on two different vectors.

Using this notation, we can restate the orthogonal projection theorem, as well as adding some properties of $\mathbf{P}$:

Theorem 3.2.2 (Orthogonal Projection Theorem II)**.** *Let S be any linear subspace, and let $\mathbf{P}\colon \mathbb{R}^N \to \mathbb{R}^N$ be the orthogonal projection onto S. The following statements are true:*

1. *The function* $\mathbf{P}$ *is linear.*

Moreover, for any $\mathbf{y} \in \mathbb{R}^N$*, we have*

2. $\mathbf{P}\mathbf{y} \in S$,
3. $\mathbf{y} - \mathbf{P}\mathbf{y} \perp S$,
4. $\|\mathbf{y}\|^2 = \|\mathbf{P}\mathbf{y}\|^2 + \|\mathbf{y} - \mathbf{P}\mathbf{y}\|^2$,
5. $\|\mathbf{P}\mathbf{y}\| \leq \|\mathbf{y}\|$, *and*
6. $\mathbf{P}\mathbf{y} = \mathbf{y}$ *if and only if* $\mathbf{y} \in S$.

These results are not difficult to prove, given theorem 3.2.1. Linearity of $\mathbf{P}$ is left as an exercise (exercise 3.6.11). Parts 2 and 3 follow directly from theorem 3.2.1. To see part 4, observe that $\mathbf{y}$ can be decomposed as $\mathbf{y} = \mathbf{P}\mathbf{y} + \mathbf{y} - \mathbf{P}\mathbf{y}$. Part 4 now follows from parts 2–3 and the Pythagorean law. (Why?) Part 5 follows from part 4. (Why?) Part 6 follows from the definition of $\mathbf{P}\mathbf{y}$ as the closest point to $\mathbf{y}$ in S.

Fact 3.2.1. If $\{\mathbf{u}_1, \ldots, \mathbf{u}_K\}$ is an orthonormal basis for S, then, for each $\mathbf{y} \in \mathbb{R}^N$,

$$\mathbf{P}\mathbf{y} = \sum_{k=1}^{K} (\mathbf{y}'\mathbf{u}_k)\mathbf{u}_k \tag{3.3}$$

Fact 3.2.1 is a fundamental result. We can see it's true immediately because the right hand side of (3.3) clearly lies in S (being a linear combination of basis functions) and, for any $\mathbf{u}_j$ in the basis set

$$(\mathbf{y} - \mathbf{P}\mathbf{y})'\mathbf{u}_j = \mathbf{y}'\mathbf{u}_j - \sum_{k=1}^{K} (\mathbf{y}'\mathbf{u}_k)(\mathbf{u}_k'\mathbf{u}_j) = \mathbf{y}'\mathbf{u}_j - \mathbf{y}'\mathbf{u}_j = 0$$

This confirms $\mathbf{y} - \mathbf{P}\mathbf{y} \perp S$ by fact 3.1.1 on page 107.

There's one more very important property of $\mathbf{P}$ that we need to make note of: Suppose we have two linear subspaces S_1 and S_2 of $\mathbb{R}^N$, where $S_1 \subset S_2$. What then is the difference between (a) first projecting a point onto the bigger subspace S_2, and then projecting the result onto the smaller subspace S_1, and (b) projecting directly to the smaller subspace S_1? The answer is none—we get the same result.

Fact 3.2.2. Let S_1 and S_2 be two subspaces of $\mathbb{R}^N$, and let $\mathbf{y} \in \mathbb{R}^N$. Let $\mathbf{P}_1$ and $\mathbf{P}_2$ be the projections onto S_1 and S_2 respectively. If $S_1 \subset S_2$, then

$$\mathbf{P}_1\mathbf{P}_2\mathbf{y} = \mathbf{P}_2\mathbf{P}_1\mathbf{y} = \mathbf{P}_1\mathbf{y}$$

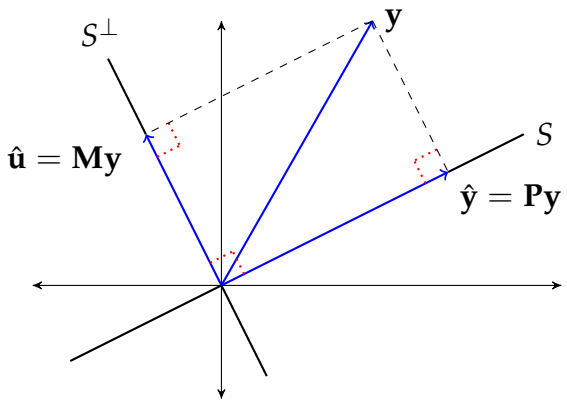


Figure 3.6: Orthogonal projection

3.2.3 Projection as Decomposition

There's yet another way of stating the orthogonal projection theorem, which is also informative. Recall the definition of orthogonal complements from §3.1.3. In the orthogonal projection theorem, our interest was in projecting $\mathbf{y}$ onto S, but we have just learned that $S^\perp$ is itself a linear subspace, so we can also project $\mathbf{y}$ onto $S^\perp$. Just as we used $\mathbf{P}$ to denote the function sending $\mathbf{y}$ into its projection onto S, so we'll use $\mathbf{M}$ to denote the function sending $\mathbf{y}$ into its projection onto $S^\perp$. The result we'll denote by $\hat{\mathbf{u}}$, so that $\hat{\mathbf{u}} := \mathbf{My}$. Figure 3.6 illustrates. The figure suggests that we will have $\mathbf{y} = \hat{\mathbf{y}} + \hat{\mathbf{u}}$, and indeed that is the case. The next theorem states this somewhat more mathematically.

Theorem 3.2.3 (Orthogonal Projection Theorem III)**.** *Let S be a linear subspace of $\mathbb{R}^N$. If $\mathbf{P}$ is the orthogonal projection onto S and $\mathbf{M}$ is the orthogonal projection onto $S^\perp$, then $\mathbf{Py}$ and $\mathbf{My}$ are orthogonal, and*

$$\mathbf{y} = \mathbf{Py} + \mathbf{My}$$

If S_1 and S_2 are two subspaces of $\mathbb{R}^N$ with $S_1 \subset S_2$, then $S_2^\perp \subset S_1^\perp$. This means that the result in fact 3.2.2 is reversed for $\mathbf{M}$.

Fact 3.2.3. Let S_1 and S_2 be two subspaces of $\mathbb{R}^N$ and let $\mathbf{y} \in \mathbb{R}^N$. Let $\mathbf{M}_1$ and $\mathbf{M}_2$ be the projections onto $S_1^\perp$ and $S_2^\perp$ respectively. If $S_1 \subset S_2$, then,

$$\mathbf{M}_1\mathbf{M}_2\mathbf{y} = \mathbf{M}_2\mathbf{M}_1\mathbf{y} = \mathbf{M}_2\mathbf{y}$$

Fact 3.2.4. $\mathbf{Py} = \mathbf{0}$ if and only if $\mathbf{y} \in S^\perp$, and $\mathbf{My} = \mathbf{0}$ if and only if $\mathbf{y} \in S$.[1]

[1]For example, if $\mathbf{Py} = \mathbf{0}$, then $\mathbf{y} = \mathbf{Py} + \mathbf{My} = \mathbf{My}$. Hence $\mathbf{M}$ does not shift $\mathbf{y}$. If an orthogonal

3.3 Applications of Projection

[Roadmap]

3.3.1 Projection Matrices

As stated in theorem 3.2.2, given any subspace S, the corresponding orthogonal projection $\mathbf{P}$ is a linear map from $\mathbb{R}^N$ to $\mathbb{R}^N$. In view of theorem 1.3.1 on page 20, it then follows that there exists an $N \times N$ matrix $\hat{\mathbf{P}}$ such that $\mathbf{P}\mathbf{x} = \hat{\mathbf{P}}\mathbf{x}$ for all $\mathbf{x} \in \mathbb{R}^N$. In fact we've anticipated this in the notation $\mathbf{P}$, and from now on $\mathbf{P}$ will also represent the corresponding matrix. But what does this matrix look like?

Theorem 3.3.1. *Let S be a subspace of $\mathbb{R}^N$. If $\mathbf{P}$ is the orthogonal projection onto S and $\{\mathbf{b}_1, \ldots, \mathbf{b}_K\}$ is any basis for S, then*

$$\mathbf{P} = \mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}' \tag{3.4}$$

where $\mathbf{B}$ is the matrix formed by taking this basis as its columns.

This result is actually a generalization of fact 3.2.1 on page 114. While the expression isn't quite as neat, the advantage is that it applies to any basis, not just orthonormal ones. We further explore the connection between the two results in §3.3.3.

Proof of theorem 3.3.1. Fix $\mathbf{y} \in \mathbb{R}^N$. The claim is that the vector $\hat{\mathbf{y}} := \mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'\mathbf{y}$ is the orthogonal projection of $\mathbf{y}$ onto S, where $\mathbf{B}$ is the matrix defined by $\text{col}_k(\mathbf{B}) = \mathbf{b}_k$. To verify this, we need to show that

1. $\hat{\mathbf{y}} \in S$, and
2. $\mathbf{y} - \hat{\mathbf{y}} \perp S$.

Part 1 is true because $\hat{\mathbf{y}}$ can be written as $\hat{\mathbf{y}} = \mathbf{B}\mathbf{x}$ where $\mathbf{x} := (\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'\mathbf{y}$. The vector $\mathbf{B}\mathbf{x}$ is a linear combination of the columns of $\mathbf{B}$. Since these columns form a basis of S they must lie in S. Hence $\hat{\mathbf{y}} \in S$ as claimed.

projection onto a subspace doesn't shift a point, that's because the point is already in that subspace (see, e.g., theorem 3.2.2). In this case the subspace is $S^\perp$, and we conclude that $\mathbf{y} \in S^\perp$.

Regarding part 2, from the assumption that $\mathbf{B}$ gives a basis for S, all points in S have the form $\mathbf{Bx}$ for some $\mathbf{x} \in \mathbb{R}^K$. Thus part 2 translates to the claim that

$$\mathbf{y} - \mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'\mathbf{y} \perp \mathbf{Bx} \qquad \text{for all } \mathbf{x} \in \mathbb{R}^K$$

This is true, because if $\mathbf{x} \in \mathbb{R}^K$, then

$$(\mathbf{Bx})'[\mathbf{y} - \mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'\mathbf{y}] = \mathbf{x}'[\mathbf{B}'\mathbf{y} - \mathbf{B}'\mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'\mathbf{y}] = \mathbf{x}'[\mathbf{B}'\mathbf{y} - \mathbf{B}'\mathbf{y}] = 0$$

The proof of theorem 3.3.1 is done. □

Theorem 3.3.1 and its proof implicitly assume that $\mathbf{B}'\mathbf{B}$ is nonsingular, but this is justified because $\mathbf{B}$ is assumed to be full column rank (exercise 3.6.17).

The matrix $\mathbf{P} := \mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'$ defined in (3.4) is called the **projection matrix** associated with the basis $\mathbf{B}$ (or the subspace S). It is common to also define the **annihilator** associated with $\mathbf{B}$ as

$$\mathbf{M} := \mathbf{I} - \mathbf{P} \tag{3.5}$$

Here $\mathbf{I}$ is, as usual, the identity matrix (in this case $N \times N$). In view of theorem 3.2.3, the annihilator matrix projects any vector onto the orthogonal complement of S.

Fact 3.3.1. Both $\mathbf{P}$ and $\mathbf{M}$ are symmetric and idempotent.

The proof is an exercise (exercise 3.6.19). Idempotence is actually immediate because both $\mathbf{P}$ and $\mathbf{M}$ represent orthogonal projections onto their respective linear subspaces. Applying the mapping a second time has no effect, because the vector is already in the subspace. However you can obtain a more concrete proof by direct computation.

3.3.2 Overdetermined Systems of Equations

An initial discussion of overdetermined systems was given in §1.3.5. Let's recall the main idea, this time using notation oriented towards linear regression (which is probably the most important application of the theory). We consider a system of equations $\mathbf{X}\boldsymbol{\beta} = \mathbf{y}$ where $\mathbf{X}$ is $N \times K$, $\boldsymbol{\beta}$ is $K \times 1$, and $\mathbf{y}$ is $N \times 1$. We regard $\mathbf{X}$ and $\mathbf{y}$ as given, and seek a $\boldsymbol{\beta} \in \mathbb{R}^K$ that solves this equation.

If $K = N$ and $\mathbf{X}$ has full column rank, then theorem 1.3.2 implies that this system has precisely one solution, which is $\mathbf{X}^{-1}\mathbf{y}$. When $N > K$ the system is overdetermined

and we cannot in general find a $\boldsymbol{\beta}$ that satisfies all N equations. We aim instead for the "best available" vector, which is the $\boldsymbol{\beta} \in \mathbb{R}^K$ making $\mathbf{X}\boldsymbol{\beta}$ is as close to $\mathbf{y}$ as possible (in the Euclidean sense). Using the orthogonal projection theorem, the minimizer is easy to identify:

Theorem 3.3.2. *Let $\mathbf{X}$ be an $N \times K$ matrix with full column rank and let $\mathbf{y} \in \mathbb{R}^N$. The minimization problem $\min_{\boldsymbol{\beta} \in \mathbb{R}^K} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|$ has a unique solution. The solution is*

$$\hat{\boldsymbol{\beta}} := (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{3.6}$$

Proof. Let $\mathbf{X}$ and $\mathbf{y}$ be as in the statement of the theorem. Let $\hat{\boldsymbol{\beta}}$ be as in (3.6) and let $S := \text{span}(\mathbf{X})$. By the full column rank assumption, $\mathbf{X}$ forms a basis for S. In view of theorem 3.3.1, the orthogonal projection of $\mathbf{y}$ onto S is

$$\hat{\mathbf{y}} := \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

Pick any $\boldsymbol{\beta} \in \mathbb{R}^K$ such that $\boldsymbol{\beta} \neq \hat{\boldsymbol{\beta}}$. By the definition of S we have $\tilde{\mathbf{y}} := \mathbf{X}\boldsymbol{\beta} \in S$. Moreover, since $\boldsymbol{\beta} \neq \hat{\boldsymbol{\beta}}$ we must have $\tilde{\mathbf{y}} \neq \hat{\mathbf{y}}$. (Why?!) Hence, by the orthogonal projection theorem (page 112), we have $\|\mathbf{y} - \hat{\mathbf{y}}\| < \|\mathbf{y} - \tilde{\mathbf{y}}\|$. In other words,

$$\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\| < \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|$$

Since $\boldsymbol{\beta}$ was arbitrary the proof is now done. □

The solution $\hat{\boldsymbol{\beta}}$ in theorem 3.3.2 is sometimes called the **least squares solution** to the minimization problem described in that theorem. The intuition behind this notation will be given later on.

Remark 3.3.1. What happens if we drop the assumption that the columns of $\mathbf{X}$ are linearly independent? The set $\text{span}(\mathbf{X})$ is still a linear subspace, and the orthogonal projection theorem still gives us a closest point $\hat{\mathbf{y}}$ to $\mathbf{y}$ in $\text{span}(\mathbf{X})$. Since $\hat{\mathbf{y}} \in \text{span}(\mathbf{X})$, there still exists a vector $\hat{\boldsymbol{\beta}}$ such that $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$. The problems is that now there exists an infinity of such vectors. Exercise 3.6.18 asks you to prove this.

The projection matrix and the annihilator in this case are

$$\mathbf{P} := \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \quad \text{and} \quad \mathbf{M} := \mathbf{I} - \mathbf{P} \tag{3.7}$$

Fact 3.3.2. The annihilator $\mathbf{M}$ associated with $\mathbf{X}$ satisfies $\mathbf{MX} = \mathbf{0}$ (exercise 3.6.21).

3.3.3 Gram Schmidt Orthogonalization

The expression for projection onto a basis given in theorem 3.3.1 is extremely useful. At the same time, it isn't quite as neat as the expression for projection onto an orthonormal basis given in fact 3.2.1 (page 114). For this and other reasons, it's nice to know that any basis of a linear subspace can be converted to an orthonormal basis in a straightforward way. The rest of this section gives details.

Theorem 3.3.3. *Given any linearly independent subset* $\{\mathbf{b}_1, \dots, \mathbf{b}_K\}$ *of* $\mathbb{R}^N$*, there exists an orthonormal set* $\{\mathbf{u}_1, \dots, \mathbf{u}_K\}$ *such that*

$$\operatorname{span}\{\mathbf{b}_1, \dots, \mathbf{b}_k\} = \operatorname{span}\{\mathbf{u}_1, \dots, \mathbf{u}_k\} \quad \textit{for} \quad k = 1, \dots, K$$

The proof of theorem 3.3.3 provides a concrete algorithm for generating the orthonormal set $\{\mathbf{u}_1, \dots, \mathbf{u}_K\}$. The first step is to construct orthogonal sets $\{\mathbf{v}_1, \dots, \mathbf{v}_k\}$ with span identical to $\{\mathbf{b}_1, \dots, \mathbf{b}_k\}$ for each k. At the end one can just normalize each vector to produce orthonormal sets with the same span.

The construction of $\{\mathbf{v}_1, \dots, \mathbf{v}_K\}$ uses the so called **Gram Schmidt orthogonalization** procedure, which runs as follows: First, for each $k = 1, \dots, K$, let $\mathbf{B}_k$ be the $N \times k$ matrix formed by columns $\mathbf{b}_1, \dots, \mathbf{b}_k$. Let $\mathbf{P}_k := \mathbf{B}_k(\mathbf{B}_k'\mathbf{B}_k)^{-1}\mathbf{B}_k'$ be the associated orthogonal projection matrix and let $\mathbf{P}_0 := \mathbf{0}$. Define $\mathbf{v}_1, \dots, \mathbf{v}_K$ by

$$\mathbf{v}_k := \mathbf{b}_k - \mathbf{P}_{k-1}\mathbf{b}_k \tag{3.8}$$

Looking at the definition of the annihilator in (3.5), the idea behind the algorithm is clear—we are using the annihilator to project each successive $\mathbf{b}_k$ into the orthogonal complement of the span of the previous vectors $\mathbf{b}_1, \dots, \mathbf{b}_{k-1}$.

Exercises 3.6.22–3.6.24 step you through the process of verifying that $\{\mathbf{v}_1, \dots, \mathbf{v}_k\}$ is orthogonal with span equal to that of $\{\mathbf{b}_1, \dots, \mathbf{b}_k\}$ for $k = 1, \dots, K$. As part of the proof it is shown that no $\mathbf{v}_k$ equals $\mathbf{0}$, and hence we can define $\mathbf{u}_k := \mathbf{v}_k / \|\mathbf{v}_k\|$. The set $\{\mathbf{u}_1, \dots, \mathbf{u}_k\}$ is clearly orthonormal for each k, and its span is the same as $\{\mathbf{v}_1, \dots, \mathbf{v}_k\}$. (Why?) In other words, it has the properties stated in theorem 3.3.3.

3.3.4 Solving Systems via QR Decomposition

Let $\mathbf{X}$ be an $N \times K$ matrix with linearly independent columns $\mathbf{x}_1, \dots, \mathbf{x}_K$. The QR decomposition theorem on page 110 tells us that there exists an invertible upper

triangular $K \times K$ matrix $\mathbf{R}$ and an $N \times K$ matrix $\mathbf{Q}$ with orthonormal columns such that $\mathbf{X} = \mathbf{QR}$. Now we have theorem 3.3.3 in hand, we can give a construction for these matrices.

To begin, theorem 3.3.3 gives us existence of an orthonormal set $\{\mathbf{u}_1, \ldots, \mathbf{u}_K\}$ such that the span of $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$ equals that of $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ for $k = 1, \ldots, K$. In particular, $\mathbf{x}_k$ is in the span of $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$. Appealing to fact 3.1.4 on page 109, we can write

$$\begin{aligned}
\mathbf{x}_1 &= (\mathbf{x}_1'\mathbf{u}_1)\mathbf{u}_1 \\
\mathbf{x}_2 &= (\mathbf{x}_2'\mathbf{u}_1)\mathbf{u}_1 + (\mathbf{x}_2'\mathbf{u}_2)\mathbf{u}_2 \\
\mathbf{x}_3 &= (\mathbf{x}_3'\mathbf{u}_1)\mathbf{u}_1 + (\mathbf{x}_3'\mathbf{u}_2)\mathbf{u}_2 + (\mathbf{x}_3'\mathbf{u}_3)\mathbf{u}_3
\end{aligned}$$

and so on. Sticking to the 3×3 case to simplify expressions, we can stack these equations horizontally to get

$$\begin{pmatrix} | & | & | \\ \mathbf{x}_1 & \mathbf{x}_2 & \mathbf{x}_3 \\ | & | & | \end{pmatrix} = \begin{pmatrix} | & | & | \\ \mathbf{u}_1 & \mathbf{u}_2 & \mathbf{u}_3 \\ | & | & | \end{pmatrix} \begin{pmatrix} (\mathbf{x}_1'\mathbf{u}_1) & (\mathbf{x}_2'\mathbf{u}_1) & (\mathbf{x}_3'\mathbf{u}_1) \\ 0 & (\mathbf{x}_2'\mathbf{u}_2) & (\mathbf{x}_3'\mathbf{u}_2) \\ 0 & 0 & (\mathbf{x}_3'\mathbf{u}_3) \end{pmatrix}$$

or $\mathbf{X} = \mathbf{QR}$. This is our QR decomposition.

Given this decomposition $\mathbf{X} = \mathbf{QR}$, the least squares solution $\hat{\boldsymbol{\beta}}$ defined in (3.6) can also be written as $\hat{\boldsymbol{\beta}} = \mathbf{R}^{-1}\mathbf{Q}'\mathbf{y}$. Exercise 3.6.25 asks you to confirm this.

3.4 Projections in L_2

The main purpose of this section is to introduce conditional expectations and study their properties. The definition of conditional expectations given in elementary probability texts is often cumbersome to work with, and fails to provide the big picture. In advanced texts, there are several different approaches to presenting conditional expectations. The one I present here is less common than the plain vanila treatment, but it is, to my mind, by far the most intuitive. As you might expect given the location of this discussion, the presentation involves orthogonal projection.

3.4.1 The Space L_2

Sometimes we want to predict the value of a random variable y using another variable x. In this case we usually choose x such that y and x are expected to be close

under each realization of uncertainty. A natural measure of closeness for random variables is **mean squared error** (MSE), which is defined in this case as $\mathbb{E}\left[(x-y)^2\right]$. This same concept of distance between random variables came up previously in our linear prediction problem of §2.4.4, as well as in measuring convergence of sequences of random variables (see, e.g., page 75). Mean square distance and its variations constitute the most commonly used measures of deviation between random variables in probability theory.

One such variation is when we replace MSE with the root mean squared error (RMSE), which is, as the name suggests, the square root of the MSE. Since we'll be using it a lot, let's give the RMSE its own notation:

$$\|x-y\| := \sqrt{\mathbb{E}\left[(x-y)^2\right]}$$

This notion of distance is quite reminiscent of Euclidean distance between vectors. For example, it's constructed from a norm, in the sense that if we define the L_2 **norm** as

$$\|z\| := \sqrt{\mathbb{E}\left[z^2\right]} \tag{3.9}$$

then the RMSE between x and y is the norm of the random variable $x-y$. This is similar to the idea of Euclidean distance, where the distance between $\mathbf{x}$ and $\mathbf{y}$ is the norm of $\mathbf{x}-\mathbf{y}$. Moreover, the norm is obtained by taking the square root after "summing" over squared values, just like Euclidean norm.

Of course the value (3.9) is only finite if z has finite second moment. So for the remainder of this section we'll restrict attention to such random variables. The standard name of this set of random variables is L_2. That is,

$$L_2 := \{ \text{ all random variables } x \text{ with } \mathbb{E}\left[x^2\right] < \infty \}$$

The space L_2 with its norm (3.9) is important partly because it shares so many properties with Euclidean space, and hence geometric intuition carries over from the latter to L_2. We'll see many examples of this below. Here's a list of some of the many ways that L_2 is similar to Euclidean space

Fact 3.4.1. Let x and y be random variables in L_2 and let α and β be any scalars. The following statements are true:

1. $\|\alpha x\| = |\alpha|\|x\|$

2. $\|x\| \geq 0$ and $\|x\| = 0$ if and only if $x = 0$.[2]
3. $\|x + y\| \leq \|x\| + \|y\|$.

One important consequence of these results is that if $\alpha, \beta \in \mathbb{R}$ and $x, y \in L_2$, then $\alpha x + \beta y$ also has finite norm, and is therefore in L_2. More generally, any **linear combination**

$$\alpha_1 x_1 + \cdots \alpha_K x_K, \qquad \alpha_k \in \mathbb{R},\ x_k \in L_2$$

of L_2 random variables is again L_2. The following definitions mimic the Euclidean case: If $X := \{x_1, \ldots, x_K\}$ is a subset of L_2, then the set of finite linear combinations of elements of X is called the **span** of X, and denoted by $\text{span}(X)$:

$$\text{span}(X) := \left\{ \text{all random variables } \sum_{k=1}^{K} \alpha_k x_k \text{ such that } \boldsymbol{\alpha} := (\alpha_1, \ldots, \alpha_K) \in \mathbb{R}^K \right\}$$

A subset S of L_2 is called a **linear subspace** of L_2 if it is closed under addition and scalar multiplication. That is, for each $x, y \in S$ and $\alpha, \beta \in \mathbb{R}$, we have $\alpha x + \beta y \in S$.

Example 3.4.1. Let Z be the set of all zero mean random variables in L_2. The set Z is a linear subspace of L_2 because if $x, y \in Z$ and $\alpha, \beta \in \mathbb{R}$, then $\mathbb{E}[\alpha x + \beta y] = \alpha \mathbb{E}[x] + \beta \mathbb{E}[y] = 0$.

Example 3.4.2. Fix $p \in L_2$ and let S be all zero mean random variables uncorrelated with p. That is, $S := \{x \in Z : \text{cov}[x, p] = 0\}$. The set S is a linear subspace of L_2 because if $x, y \in S$ and $\alpha, \beta \in \mathbb{R}$, then

$$\text{cov}[\alpha x + \beta y, p] = \mathbb{E}[(\alpha x + \beta y)p] = \alpha \mathbb{E}[xp] + \beta \mathbb{E}[yp] = \alpha \,\text{cov}[x, p] + \beta \,\text{cov}[y, p] = 0$$

Following on from fact 3.4.1, we can draw another parallel between L_2 norm and the Euclidean norm. As we saw in §1.1.1, the Euclidean norm is defined in terms of the inner product on $\mathbb{R}^N$. If $\mathbf{x}$ and $\mathbf{y}$ are two vectors in $\mathbb{R}^N$, then the inner product is $\mathbf{x}'\mathbf{y}$, and the norm of vector $\mathbf{x}$ is $\|\mathbf{x}\| = \sqrt{\mathbf{x}'\mathbf{x}}$. Similarly, for random variables x and y, we define

$$\langle x, y \rangle := \text{inner product of } x \text{ and } y \ := \mathbb{E}[xy]$$

[2] Actually, while $\|\mathbf{x}\| = 0$ certainly implies that $\mathbf{x} = \mathbf{0}$ in Euclidean space, it isn't quite true that $\|x\| = 0$ implies that x is the zero random variable (i.e., $x(\omega) = 0$ for all $\omega \in \Omega$). However, we can say that if $\|x\| = 0$, then $x = 0$ with probability one. More formally, the set $E := \{\omega \in \Omega : |x(\omega)| > 0\}$ satisfies $\mathbb{P}(E) = 0$. In this sense, x differs from the zero random variable only in a trivial way. In the applications we consider this caveat never causes problems. If you do continuous time stochastic dynamics, however, you'll find that there's a lot of mathematical machinery built up to keep control of this issues.

As for the Euclidean case, we then have

$$\|x\| = \sqrt{\langle x, x \rangle} := \sqrt{\mathbb{E}\left[x^2\right]}$$

As in the Euclidean case, if the inner product of x and y is zero, then we say that x and y are **orthogonal**, and write $x \perp y$. This terminology is used frequently in econometrics (sometimes by people who aren't actually sure why the term "orthogonal" is used—which puts you one step ahead of them). This is because if either x or y is zero mean, then orthogonality of x and y is equivalent to $\operatorname{cov}[x, y] = 0$. Hence, for centered random variables, orthogonality is equivalent to lack of correlation.

The next fact shows how the inner product on L_2 has all the same essential properties as the inner product on Euclidean space (cf. fact 1.1.1 on page 4).

Fact 3.4.2. For any $\alpha, \beta \in \mathbb{R}$ and any $x, y, z \in L_2$, the following statements are true:

1. $\langle x, y \rangle = \langle y, x \rangle$
2. $\langle \alpha x, \beta y \rangle = \alpha\beta \langle x, y \rangle$
3. $\langle x, (\alpha y + \beta z) \rangle = \alpha \langle x, y \rangle + \beta \langle x, z \rangle$

This properties follow immediately from linearity of $\mathbb{E}$. The inequality

$$|\langle x, y \rangle| \leq \|x\| \|y\| \qquad (\text{all } x, y \in L_2)$$

also carries over from the Euclidean case (see fact 1.1.2 on page 5). In fact we've already seen this inequality—it's precisely the Cauchy-Schwarz inequality for random variables shown in fact 2.2.7 on page 60. Now you know why these two inequalities share the same name.

Given that inner products and norms in L_2 share so many properties with their Euclidean cousins, you might not be surprized to learn that a great variety results about Euclidean space carry over to L_2. Often we can just lift the proofs across with barely a change to notation. Since we know a lot about vectors and Euclidean space, this process yields many insights about random variables and their properties.

One result that carries across essentially unchanged is the orthogonal projection theorem. Let's now state the L_2 version. Given $y \in L_2$ and linear subspace $S \subset L_2$, we seek the closest element $\hat{y}$ of S to y. Closeness is in terms of L_2 norm, so $\hat{y}$ is the minimizer of $\|y - z\|$ over all $z \in S$. That is,

$$\hat{y} = \operatorname*{argmin}_{z \in S} \|y - z\| = \operatorname*{argmin}_{z \in S} \sqrt{\mathbb{E}\left[(y - z)^2\right]} \tag{3.10}$$

The next theorem mimics theorem 3.2.1 on page 112.

Theorem 3.4.1 (Orthogonal Projection Theorem IV). *Let $y \in L_2$ and let S be any nonempty linear subspace of L_2. The following statements are true:*[3]

1. *The optimization problem* (3.10) *has exactly one solution.*
2. *Vector $\hat{y} \in L_2$ is the unique solution to* (3.10) *if and only if $\hat{y} \in S$ and $y - \hat{y} \perp S$.*

The vector $\hat{y}$ in theorem 3.2.1 is called the **orthogonal projection of y onto S**. Holding S fixed, we can think of the operation

$$y \mapsto \text{ the orthogonal projection of } y \text{ onto } S$$

as a function from L_2 to L_2. The function will be denoted by P, so that, for each $y \in L_2$, the symbol Py represents image of y under P, which is the orthogonal projection $\hat{y}$. Just as in the Euclidean case this function P is linear, in the sense that

$$P(\alpha x + \beta y) = \alpha Px + \beta Py \tag{3.11}$$

for any $x, y \in L_2$ and scalars α and β. In addition,

Fact 3.4.3. Let S be any linear subspace of L_2, and let $P \colon L_2 \to L_2$ be the orthogonal projection onto S. For any $y \in L_2$, we have

2. $Py \in S$
3. $y - Py \perp S$
4. $\|y\|^2 = \|Py\|^2 + \|y - Py\|^2$
5. $\|Py\| \leq \|y\|$
6. $Py = y$ if and only if $y \in S$
7. If $\{u_1, \ldots, u_K\}$ is an orthonormal basis of S, then

$$Py = \sum_{k=1}^{K} \langle y, u_k \rangle u_k \tag{3.12}$$

[3]There are two small caveats I should mention. First, we actually require that S is a "closed" linear subspace of L_2, which means that if $\{x_n\} \subset S$, $x \in L_2$ and $\|x_n - x\| \to 0$, then $x \in S$. For the subspaces we consider here, this condition is always true. Second, when we talk about uniqueness in L_2, we do not distinguish between elements x and x' of L_2 such that $\mathbb{P}\{x = x'\} = 1$. A nice treatment of orthogonal projection in Hilbert spaces (of which L_2 is one example) is provided in Cheney (2001, chapter 2). Most other books covering Hilbert space will provide some discussion.

In the final item, the meaning of orthonormal basis is that $\langle u_j, u_k \rangle = 1$ if $j = k$ and zero otherwise, and that S is equal to the span of $\{u_1, \ldots, u_K\}$. See (3.3) on page 114 for comparison. More generally, these results parallel those given for projection on Euclidean space, and the proofs are essentially the same.

3.4.2 Application: Best Linear Prediction

There are many many important applications of orthogonal projection in L_2. We'll treat one of the most important ones in the next section—conditional expectations. For now let's stick to a simpler "warm up" applications. Our main objective is to revisit the problem of best linear predictors from §2.4.4, establishing the same set of results in a different way.

To begin, let

- $\mathbb{1} := \mathbb{1}_\Omega =$ the constant random variable that is always equal to 1
- $S_1 := \text{span}(\mathbb{1}) =$ the linear subspace of all constant random variables
- $P_1 :=$ be the orthogonal projection onto S_1
- $M_1 :=$ be the annihilator projection (cf. page 117) defined by $M_1 x = x - P_1 x$

Now let $x \in L_2$ with $\mathbb{E}[x] = \mu$ and $\text{var}[x] = \sigma_x^2$. Observe first that $P_1 x = \mu \mathbb{1}$. (We've written "$\mu\mathbb{1}$" instead of just "μ" to remind ourselves that we're thinking of a constant random variable, rather than just a constant.) To see this we need only confirm that $\mu\mathbb{1} \in S_1$, which is obvious, and that $\mathbb{E}[\mathbb{1}(x - \mu\mathbb{1})] = 0$. The second claim is also obvious, because $\mathbb{E}[\mathbb{1}(x - \mu\mathbb{1})] = \mathbb{E}[x - \mu] = \mathbb{E}[x] - \mu$.

Since $P_1 x = \mu\mathbb{1}$, we have $M_1 x = x - \mu\mathbb{1}$, or, written more conventionally, $M_1 x = x - \mu$. In other words, $M_1 x$ is the "de-meaned" or "centered" version of x. The norm of $M_1 x$ is

$$\|M_1 x\| = \sqrt{\mathbb{E}[(x - \mu)^2]} = \sigma_x$$

In other words, the norm of $M_1 x$ is the standard devation of x.

Next fix $x, y \in L_2$ and consider projecting y onto $S_2 := \text{span}(\mathbb{1}, x)$. That is, S_2 is equal to the set of random variables

$$\alpha + \beta x := \alpha\mathbb{1} + \beta x \quad \text{for scalars} \quad \alpha, \beta$$

Let P_2 and M_2 be the orthogonal projection and annihilator associated with S_2.

Perhaps the easiest way to project y onto S_2 is to find an orthonormal basis $\{u_1, u_2\}$ for S_2 and then apply (3.12). Getting our inspiration from the Gram-Schmidt orthogonalization procedure (page 119), we take

$$u_1 := \mathbb{1} \quad \text{and} \quad u_2 := \frac{x - \mu}{\sigma_x} = \frac{M_1 x}{\|M_1 x\|}$$

It's easy to check that $\langle u_1, u_2 \rangle = \mathbb{E}[u_1 u_2] = 0$ and $\|u_1\| = \|u_2\| = 1$, so this pair is indeed orthonormal. It's also straightforward to show that $\operatorname{span}(u_1, u_2) = S_2$, so $\{u_1, u_2\}$ is an orthonormal basis for S_2. Hence, by (3.12),

$$P_2 y = \langle y, u_1 \rangle u_1 + \langle y, u_2 \rangle u_2 = \mathbb{E}[y] + \frac{\operatorname{cov}[x, y]}{\operatorname{var}[x]}(x - \mathbb{E}[x])$$

The missing algebra between these two equations is not difficult to replicate. Alternatively, we can write

$$P_2 y = \alpha^* + \beta^* x \quad \text{where} \quad \beta^* := \frac{\operatorname{cov}[x, y]}{\operatorname{var}[x]} \quad \text{and} \quad \alpha^* := \mathbb{E}[y] - \beta^* \mathbb{E}[x]$$

These are the same solutions we obtained using calculus in §2.4.4.

3.4.3 Measurability

Now let's turn towards a really fundamental application of orthogonal projection in L_2. We'll set up these results in several steps. To begin, recall that, in the case of $\mathbb{R}^N$, orthogonal projection starts with a linear subspace S of $\mathbb{R}^N$. Once we have this subspace, we think about how to project onto it. The space S is obviously crucial because once we select S, we implicitly define the orthogonal projection mapping. So when I tell you that conditional expectation is characterized by orthogonal projection, you will understand that the first thing we need to do is identify the linear subspaces that we want to project onto.

The first step in this process is a definition at the very heart of probability theory: measurability. Let $x_1, \ldots, x_p$ be some collection of random variables, and let $\mathcal{G} := \{x_1, \ldots, x_p\}$. Thus, $\mathcal{G}$ is a set of random variables, often referred to in what follows as the **information set**. We will say that another random variable z is **$\mathcal{G}$-measurable** if there exists a (nonrandom) function $g\colon \mathbb{R}^p \to \mathbb{R}$ such that

$$z = g(x_1, \ldots, x_p)$$

Informally, what this means is that once the values of the random variables $x_1, \ldots, x_p$ have been realized, the variable z is completely determined (i.e., no longer random) and it's realized value can be calculated (assuming that we can calculate the functional form g). You might like to imagine it like this: Uncertainty is realized, in the sense that some ω is selected from the sample space Ω. Suppose that we don't get to view ω itself, but we do get to view certain random outcomes. For example, we might get to observe the realized values $x_1(\omega), \ldots, x_p(\omega)$. If z is $\mathcal{G}$-measurable, we can now calculate the realized value $z(\omega)$ of z, even without knowning ω, because we can compute $z(\omega) = g(x_1(\omega), \ldots, x_p(\omega))$.[4]

As a matter of notation, if $\mathcal{G} = \{x\}$ and y is $\mathcal{G}$-measurable, then we will also say that y is x-measurable.

Example 3.4.3. Let x and z be two random variables. If $z = 2x + 3$, then z is x-measurable. To see this formally, we can write $z = g(x)$ when $g(x) = 2x + 3$. Less formally, when x is realized, the value of z can be calculated.

Example 3.4.4. Let $x_1, \ldots, x_N$ be random variables and let $\bar{x}_N$ be their sample mean. If $\mathcal{G} = \{x_1, \ldots, x_N\}$, then $\bar{x}_N := N^{-1} \sum_{n=1}^{N} x_n$ is clearly $\mathcal{G}$-measurable.

Example 3.4.5. If x and y are independent and neither random variable is constant, then y is not x-measurable. Indeed, if y was x-measurable, then we would have $y = g(x)$ for some function g. This contradicts independence of x and y.

Example 3.4.6. Let x, y and z be three random variables with $z = x + y$. Suppose that x and y are independent. Then z is not x-measurable. Intuitively, even if we know the realized value of x, the realization of z cannot be computed until we know the realized value of y. Formally, if z is x-measurable then $z = g(x)$ for some function g. But then $y = g(x) - x$, so y is x-measurable. This contradicts independence of x and y.

Example 3.4.7. Let $y = \alpha$, where α is a constant. This degenerate random variable is $\mathcal{G}$-measurable for any information set $\mathcal{G}$, because y is already deterministic. For example, if $\mathcal{G} = \{x_1, \ldots, x_p\}$, then we can take $y = g(x_1, \ldots, x_p) = \alpha + \sum_{i=1}^{p} 0 x_i$.

If x and y are known given the information in $\mathcal{G}$, then a third random variable that depends on only on x and y is likewise known given $\mathcal{G}$. Hence $\mathcal{G}$-measurability is preserved under the taking of sums, products, etc. In particular,

[4] A technical note: In the definition of measurability above, where we speak of existence of the function g, it is additional required that the function g is "Borel measurable." For the purposes of this course, we can regard non-Borel measurable functions as a mere theoretical curiosity. As such, the distinction will be ignored. See Williams (1991) or any similar text for further details.

Fact 3.4.4. Let α, β be any scalars, and let x and y be random variables. If x and y are both $\mathcal{G}$-measurable, then $u := xy$ and $v := \alpha x + \beta y$ are also $\mathcal{G}$-measurable.

Let $\mathcal{G}$ and $\mathcal{H}$ be two information sets with $\mathcal{G} \subset \mathcal{H}$. In this case, if random variable z is $\mathcal{G}$ measurable, *then it is also $\mathcal{H}$-measurable*. This follows from our intuitive definition of measurability: If the value z is known once the variables in $\mathcal{G}$ are known, then it is certainly known when the extra information provided by $\mathcal{H}$ is available. The next example helps to clarify.

Example 3.4.8. Let x, y and z be three random variables, let $\mathcal{G} = \{x\}$, and let $\mathcal{H} = \{x, y\}$. Suppose that $z = 2x + 3$, so that z is $\mathcal{G}$-measurable. Then z is also $\mathcal{H}$-measurable. Informally, we can see that z is deterministic once the variables in $\mathcal{H}$ are realized. Formally, we can write $z = g(x, y)$, where $g(x, y) = 2x + 3 + 0y$. Hence z is also $\mathcal{H}$-measurable as claimed.

Let's note this idea as a fact:

Fact 3.4.5. If $\mathcal{G} \subset \mathcal{H}$ and z is $\mathcal{G}$-measurable, then z is $\mathcal{H}$-measurable.

We started off this section by talking about how conditional expectations were going to be projections onto linear subspaces, and we wanted to identify the subspaces. Let's clarify this now. Given $\mathcal{G} \subset L_2$, define

$$L_2(\mathcal{G}) := \{\text{the set of all } \mathcal{G}\text{-measurable random variables in } L_2\}$$

In view of fact 3.4.4, we have the following result:

Fact 3.4.6. For any $\mathcal{G} \subset L_2$, the set $L_2(\mathcal{G})$ is a linear subspace of L_2.

From fact 3.4.5 we see that, in the sense of set inclusion, the linear subspace is increasing with respect to the information set.

Fact 3.4.7. If $\mathcal{G} \subset \mathcal{H}$, then $L_2(\mathcal{G}) \subset L_2(\mathcal{H})$.

3.4.4 Conditional Expectation

Now it's time to define conditional expectations. Let $\mathcal{G} \subset L_2$ and y be some random variable in L_2. The **conditional expectation** of y given $\mathcal{G}$ is written as $\mathbb{E}[y \mid \mathcal{G}]$

or $\mathbb{E}^{\mathcal{G}}[y]$, and defined as the closest $\mathcal{G}$-measurable random variable to y.[5] More formally,

$$\mathbb{E}[y \mid \mathcal{G}] := \operatorname*{argmin}_{z \in L_2(\mathcal{G})} \|y - z\| \tag{3.13}$$

This definition makes a lot of sense. Our intuitive understanding of the conditional expectation $\mathbb{E}[y \mid \mathcal{G}]$ is that it is the best predictor of y given the information contained in $\mathcal{G}$. The definition in (3.13) says the same thing. It simultaneously restricts $\mathbb{E}[y \mid \mathcal{G}]$ to be $\mathcal{G}$-measurable, so we can actually compute it once the variables in $\mathcal{G}$ are realized, and selects $\mathbb{E}[y \mid \mathcal{G}]$ as the closest such variable to y in terms of RMSE.

While the definition makes sense, it still leaves many open questions. For example, there are many situations where minimizers don't exist, or, if the do exist, there are lots of them. So is our definition really a definition? Moreover, even assuming we do have a proper definition, how do we actually go about computing conditional expectations in practical situations? And what properties do conditional expectations have?

These look like tricky questions, but fortunately the orthogonal projection theorem comes to the rescue.

Comparing (3.13) and (3.10), we see that $y \mapsto \mathbb{E}[y \mid \mathcal{G}]$ is exactly the orthogonal projection function P in the special case where the subspace S is the $\mathcal{G}$-measurable functions $L_2(\mathcal{G})$.

Okay, so $\mathbb{E}[y \mid \mathcal{G}]$ is the orthogonal projection of y onto $L_2(\mathcal{G})$. That's kind of neat, but what does it actually tell us? Well, it tells us quite a lot. For starters, theorem 3.4.1 implies that $\mathbb{E}[y \mid \mathcal{G}]$ is always well defined and unique. Second, it gives us a useful characterization of $\mathbb{E}[y \mid \mathcal{G}]$, because we now know that $\mathbb{E}[y \mid \mathcal{G}]$ is the unique point in L_2 such that $\mathbb{E}[y \mid \mathcal{G}] \in L_2(\mathcal{G})$ and $y - \mathbb{E}[y \mid \mathcal{G}] \perp z$ for all $z \in L_2(\mathcal{G})$. Rewriting these conditions in a slightly different way, we can give an alternative (and equivalent) definition of conditional expectation: $\mathbb{E}[y \mid \mathcal{G}] \in L_2$ is the **conditional expectation** of y given $\mathcal{G}$ if

1. $\mathbb{E}[y \mid \mathcal{G}]$ is $\mathcal{G}$-measurable, and
2. $\mathbb{E}[\mathbb{E}[y \mid \mathcal{G}]\, z] = \mathbb{E}[yz]$ for all $\mathcal{G}$-measurable $z \in L_2$.

[5] I prefer the notation $\mathbb{E}^{\mathcal{G}}[y]$ to $\mathbb{E}[y \mid \mathcal{G}]$ because, as we will see, $\mathbb{E}^{\mathcal{G}}$ is a function (an orthogonal projection) from L_2 to L_2, and the former notation complements this view. However, the notation $\mathbb{E}[y \mid \mathcal{G}]$ is a bit more standard, so that's the one we'll use.

This definition seems a bit formidable, but it's not too hard to use. Before giving an application, let's bow to common notation and define

$$\mathbb{E}[y \mid x_1, \ldots, x_p] := \mathbb{E}[y \mid \mathcal{G}]$$

Also, let's note the following "obvious" fact:

Fact 3.4.8. Given $\{x_1, \ldots, x_p\}$ and y in L_2, there exists a function $g\colon \mathbb{R}^p \to \mathbb{R}$ such that $\mathbb{E}[y \mid x_1, \ldots, x_p] = g(x_1, \ldots, x_p)$.

This is obvious because, by definition, $\mathbb{E}[y \mid \mathcal{G}]$ is $\mathcal{G}$-measurable. At the same time, it's worth keeping in mind: A conditional expectation with respect to a collection of random variables is some function of those random variables.

Example 3.4.9. If x and w are independent and $y = x + w$, then $\mathbb{E}[y \mid x] = x + \mathbb{E}[w]$.

Let's check this using the second definition of conditional expectations given above. To check that $x + \mathbb{E}[w]$ is indeed the conditional expectation of y given $\mathcal{G} = \{x\}$, we need to show that $x + \mathbb{E}[w]$ is x-measurable and that $\mathbb{E}[(x + \mathbb{E}[w])\, z] = \mathbb{E}[yz]$ for all x-measurable z. The first claim is clearly true, because $x + \mathbb{E}[w]$ is a deterministic function of x. The second claim translates to the claim that

$$\mathbb{E}[(x + \mathbb{E}[w])g(x)] = \mathbb{E}[(x + w)g(x)] \tag{3.14}$$

for any function g. Verifying this equality is left as an exercise (exercise 3.6.27)

The next example shows that when x and y are linked by a conditional density (remember: densities don't always exist), then our definition of conditional expectation reduces to the one seen in elementary probability texts. The proof of the claim in the example is the topic of exercise 3.6.32.

Example 3.4.10. If x and y are random variables and $p(y \mid x)$ is the conditional density of y given x, then

$$\mathbb{E}[y \mid x] = \int t p(t \mid x) dt$$

There are some additional goodies we can harvest using the fact that conditional expectation is an orthogonal projection.

Fact 3.4.9. Let x and y be random variables in L_2, let α and β be scalars, and let $\mathcal{G}$ and $\mathcal{H}$ be subsets of L_2. The following properties hold.

1. Linearity: $\mathbb{E}[\alpha x + \beta y \mid \mathcal{G}] = \alpha \mathbb{E}[x \mid \mathcal{G}] + \beta \mathbb{E}[y \mid \mathcal{G}]$.
2. If $\mathcal{G} \subset \mathcal{H}$, then $\mathbb{E}[\mathbb{E}[y \mid \mathcal{H}] \mid \mathcal{G}] = \mathbb{E}[y \mid \mathcal{G}]$ and $\mathbb{E}[\mathbb{E}[y \mid \mathcal{G}]] = \mathbb{E}[y]$.
3. If y is independent of the variables in $\mathcal{G}$, then $\mathbb{E}[y \mid \mathcal{G}] = \mathbb{E}[y]$.
4. If y is $\mathcal{G}$-measurable, then $\mathbb{E}[y \mid \mathcal{G}] = y$.
5. If x is $\mathcal{G}$-measurable, then $\mathbb{E}[xy \mid \mathcal{G}] = x\mathbb{E}[y \mid \mathcal{G}]$.

Checking of these facts is mainly left to the exercises. Most are fairly straightforward. For example, consider the claim that if y is $\mathcal{G}$-measurable, then $\mathbb{E}[y \mid \mathcal{G}] = y$. In other words, we are saying that if $y \in L_2(\mathcal{G})$, then y is projected into itself. This is immediate from the last statement in theorem 3.4.1.

The fact that if $\mathcal{G} \subset \mathcal{H}$, then $\mathbb{E}[\mathbb{E}[y \mid \mathcal{H}] \mid \mathcal{G}] = \mathbb{E}[y \mid \mathcal{G}]$ is called the "tower" property of conditional expectations (by mathematicians), or the law of iterated expectations (by econometricians). The law follows from the property of orthogonal projections given in fact 3.2.2 on page 114: Projecting onto the bigger subspace $L_2(\mathcal{H})$ and from there onto $L_2(\mathcal{G})$ is the same as projecting directly onto the smaller subspace $L_2(\mathcal{G})$.

3.4.5 The Vector/Matrix Case

Conditional expectations of random matrices are defined using the notion of conditional expectations for scalar random variables. For example, given random matrices $\mathbf{X}$ and $\mathbf{Y}$, we set

$$\mathbb{E}[\mathbf{Y} \mid \mathbf{X}] := \begin{pmatrix} \mathbb{E}[y_{11} \mid \mathbf{X}] & \mathbb{E}[y_{12} \mid \mathbf{X}] & \cdots & \mathbb{E}[y_{1K} \mid \mathbf{X}] \\ \mathbb{E}[y_{21} \mid \mathbf{X}] & \mathbb{E}[y_{22} \mid \mathbf{X}] & \cdots & \mathbb{E}[y_{2K} \mid \mathbf{X}] \\ \vdots & \vdots & & \vdots \\ \mathbb{E}[y_{N1} \mid \mathbf{X}] & \mathbb{E}[y_{N2} \mid \mathbf{X}] & \cdots & \mathbb{E}[y_{NK} \mid \mathbf{X}] \end{pmatrix}$$

where

$$\mathbb{E}[y_{nk} \mid \mathbf{X}] := \mathbb{E}[y_{nk} \mid x_{11}, \ldots, x_{\ell m}, \ldots, x_{LM}]$$

We also define

$$\operatorname{cov}[\mathbf{x}, \mathbf{y} \mid \mathbf{Z}] := \mathbb{E}[\mathbf{x}\mathbf{y}' \mid \mathbf{Z}] - \mathbb{E}[\mathbf{x} \mid \mathbf{Z}]\mathbb{E}[\mathbf{y} \mid \mathbf{Z}]'$$

and

$$\operatorname{var}[\mathbf{x} \mid \mathbf{Z}] := \mathbb{E}[\mathbf{x}\mathbf{x}' \mid \mathbf{Z}] - \mathbb{E}[\mathbf{x} \mid \mathbf{Z}]\mathbb{E}[\mathbf{x} \mid \mathbf{Z}]'$$

Using the definitions, one can show that all of the results on conditional expectations in fact 3.4.9 continue to hold in the current setting, replacing scalars with vectors and matrices. We state necessary results for convenience:

Fact 3.4.10. Let $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ be random matrices, and let $\mathbf{A}$ and $\mathbf{B}$ be constant matrices. Assuming conformability of matrix operations, the following results hold:

1. $\mathbb{E}[\mathbf{Y} \mid \mathbf{Z}]' = \mathbb{E}[\mathbf{Y}' \mid \mathbf{Z}]$.
2. $\mathbb{E}[\mathbf{AX} + \mathbf{BY} \mid \mathbf{Z}] = \mathbf{A}\mathbb{E}[\mathbf{X} \mid \mathbf{Z}] + \mathbf{B}\mathbb{E}[\mathbf{Y} \mid \mathbf{Z}]$.
3. $\mathbb{E}[\mathbb{E}[\mathbf{Y} \mid \mathbf{X}]] = \mathbb{E}[\mathbf{Y}]$ and $\mathbb{E}[\mathbb{E}[\mathbf{Y} \mid \mathbf{X}, \mathbf{Z}] \mid \mathbf{X}] = \mathbb{E}[\mathbf{Y} \mid \mathbf{X}]$.
4. If $\mathbf{X}$ and $\mathbf{Y}$ are independent, then $\mathbb{E}[\mathbf{Y} \mid \mathbf{X}] = \mathbb{E}[\mathbf{Y}]$.
5. If g is a (nonrandom) function, so that $g(\mathbf{X})$ is a matrix depending only on $\mathbf{X}$, then
   - $\mathbb{E}[g(\mathbf{X}) \mid \mathbf{X}] = g(\mathbf{X})$
   - $\mathbb{E}[g(\mathbf{X})\,\mathbf{Y} \mid \mathbf{X}] = g(\mathbf{X})\mathbb{E}[\mathbf{Y} \mid \mathbf{X}]$
   - $\mathbb{E}[\mathbf{Y}\,g(\mathbf{X}) \mid \mathbf{X}] = \mathbb{E}[\mathbf{Y} \mid \mathbf{X}]\,g(\mathbf{X})$

3.4.6 An Exercise in Conditional Expectations

Let x and y be two random variables. We saw that $\mathbb{E}[y \mid x]$ is a function f of x such that $f(x)$ is the best predictor of y in terms of root mean squared error. Since monotone increasing transformations do not affect minimizers, f also minimizes the mean squared error. In other words, f solves

$$\min_{g \in G} \mathbb{E}[(y - g(x))^2] \tag{3.15}$$

where G is the set of functions from $\mathbb{R}$ to $\mathbb{R}$. From this definition of conditional expectations, we employed the orthogonal projection theorem to deduce various properties of conditional expectations. We can also reverse this process, showing directly that $f(x) := \mathbb{E}[y \mid x]$ solves (3.15), given the various properties of conditional expectations listed in fact 3.4.9. To begin, suppose that the properties in fact 3.4.9 hold, and fix an arbitrary $g \in G$. We have

$$\begin{aligned}(y - g(x))^2 &= (y - f(x) + f(x) - g(x))^2 \\ &= (y - f(x))^2 + 2(y - f(x))(f(x) - g(x)) + (f(x) - g(x))^2\end{aligned}$$

Let's consider the expectation of the cross-product term. From the law of iterated expectations (fact 3.4.9), we obtain

$$\mathbb{E}\left\{(y - f(x))(f(x) - g(x))\right\} = \mathbb{E}\left\{\mathbb{E}\left[(y - f(x))(f(x) - g(x)) \,|\, x\right]\right\} \tag{3.16}$$

We can re-write the term inside the curly brackets on the right-hand side of (3.16) as

$$(f(x) - g(x))\mathbb{E}\left[(y - f(x)) \,|\, x\right]$$

(Which part of fact 3.4.9 are we using here?) Regarding the second term in this product, we have (by which facts?) the result

$$\mathbb{E}\left[y - f(x) \,|\, x\right] = \mathbb{E}\left[y \,|\, x\right] - \mathbb{E}\left[f(x) \,|\, x\right] = \mathbb{E}\left[y \,|\, x\right] - f(x) = \mathbb{E}\left[y \,|\, x\right] - \mathbb{E}\left[y \,|\, x\right] = 0$$

We conclude that the expectation in (3.16) is $\mathbb{E}\left[0\right] = 0$. It then follows that

$$\begin{aligned}\mathbb{E}\left[(y - g(x))^2\right] &= \mathbb{E}\left[(y - f(x))^2 + 2(y - f(x))(f(x) - g(x)) + (f(x) - g(x))^2\right] \\ &= \mathbb{E}\left[(y - f(x))^2\right] + \mathbb{E}\left[(f(x) - g(x))^2\right]\end{aligned}$$

Since $(f(x) - g(x))^2 \geq 0$ we have $\mathbb{E}\left[(f(x) - g(x))^2\right] \geq 0$, and we conclude that

$$\mathbb{E}\left[(y - g(x))^2\right] \geq \mathbb{E}\left[(y - f(x))^2\right] := : \mathbb{E}\left[(y - \mathbb{E}\left[y \,|\, x\right])^2\right]$$

Since g was an arbitrary element of G, we conclude that

$$f = \operatorname*{argmin}_{g \in G} \mathbb{E}\left[(y - g(x))^2\right]$$

3.5 Further Reading

To be written.

3.6 Exercises

Ex. 3.6.1. Find two unit vectors (i.e., vectors with norm equal to one) that are orthogonal to $(1, -2)$.

Ex. 3.6.2. Prove the Pythagorean law (fact 3.1.2 on page 107). See fact 1.1.3 if you need a hint.

Ex. 3.6.3. Prove fact 3.1.3 on page 109.

Ex. 3.6.4. Let $\mathbf{Q}$ be an orthogonal matrix. Show that $\mathbf{Q}^{-1} = \mathbf{Q}'$ and $\det(\mathbf{Q}) \in \{-1, 1\}$ both hold.

Ex. 3.6.5. Use theorem 3.1.1 (page 110) to prove the following part of fact 1.4.13 (page 37): A symmetric matrix $\mathbf{A}$ is positive definite if and only if its eigenvalues are all positive.

Ex. 3.6.6. Prove theorem 3.2.3 using theorems 3.2.1–3.2.2.

Ex. 3.6.7. Prove fact 3.1.7: If $S \subset \mathbb{R}^N$, then $S \cap S^\perp = \{\mathbf{0}\}$.

Ex. 3.6.8. Prove fact 3.2.2.

Ex. 3.6.9. Let $\mathbf{x}$ and $\mathbf{y}$ be any two $N \times 1$ vectors.

1. Show that $\|\mathbf{x} + \mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 + 2\mathbf{x}'\mathbf{y}$
2. Explain the connection between this equality and the Pythagorean Law.

Ex. 3.6.10. Show that if $\mathbf{Q}$ is an $N \times N$ orthogonal matrix, then $\mathbf{Q}$ is an isometry on $\mathbb{R}^N$. That is, for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$, we have $\|\mathbf{Q}\mathbf{x} - \mathbf{Q}\mathbf{y}\| = \|\mathbf{x} - \mathbf{y}\|$.

Ex. 3.6.11. Let $\mathbf{P}$ be the orthogonal projection described in theorem 3.2.2 (page 113). Confirm that $\mathbf{P}$ is a linear function from $\mathbb{R}^N$ to $\mathbb{R}^N$, as defined in §1.2.1.

Ex. 3.6.12. Let $\mathbf{P}$ be an $N \times N$ matrix that is both symmetric and idempotent. Let $S := \text{rng}(\mathbf{P})$. Show that $\mathbf{P}$ is precisely the orthogonal projection mapping onto the linear space S. (In other words, for any given $\mathbf{y} \in \mathbb{R}^N$, the vector $\mathbf{P}\mathbf{y}$ is the closest point in S to $\mathbf{y}$.)

Ex. 3.6.13. In this exercise you are asked to prove the Cauchy-Schwarz inequality $|\mathbf{x}'\mathbf{y}| \leq \|\mathbf{x}\|\|\mathbf{y}\|$ from fact 1.1.2 on page page 5 via the orthogonal projection theorem. Let $\mathbf{y}$ and $\mathbf{x}$ be nonzero vectors in $\mathbb{R}^N$ (since if either equals zero then the inequality is trivial), and let $\text{span}(\mathbf{x})$ be all vectors of the form $\alpha\mathbf{x}$ for $\alpha \in \mathbb{R}$.

1. Letting $\mathbf{P}$ be the orthgonal projection onto $\text{span}(\mathbf{x})$, show that
$$\mathbf{P}\mathbf{y} = \frac{\mathbf{x}'\mathbf{y}}{\mathbf{x}'\mathbf{x}}\mathbf{x}$$
2. Using this expression and any relevant properties of orthgonal projections (see theorem 3.2.2 on page 113), confirm the Cauchy-Schwarz inequality.

Ex. 3.6.14. Consider theorem 3.3.2 on page 118. If $N = K$, what does $\hat{\boldsymbol{\beta}}$ reduce to? Interpret.

Ex. 3.6.15. Let $S := \{(x_1, x_2, x_3) \in \mathbb{R}^3 : x_3 = 0\}$, and let $\mathbf{y} := \mathbf{1} := (1, 1, 1)$. Using the orthogonal projection theorem, find the closest point in S to $\mathbf{y}$.

Ex. 3.6.16. Let $\mathbf{P}$ be the orthogonal projection described in theorem 3.2.2 (page 113). Is it true that $\mathbf{Px} \neq \mathbf{Py}$ whenever $\mathbf{x} \neq \mathbf{y}$? Why or why not?[6]

Ex. 3.6.17. Show that when $N \times K$ matrix $\mathbf{B}$ is full column rank, the matrix $\mathbf{B}'\mathbf{B}$ is nonsingular.[7]

Ex. 3.6.18. Prove the claim at the end of remark 3.3.1 on page 118. In particular, let $\mathbf{X}$ be $N \times K$ with linearly *dependent* columns, and let $\hat{\mathbf{y}}$ be the closest point to $\mathbf{y}$ in $\operatorname{span}(\mathbf{X})$, existence of which follows from the orthogonal projection theorem. Prove that there are infinitely many $\mathbf{b}$ such that $\hat{\mathbf{y}} = \mathbf{Xb}$.

Ex. 3.6.19. Show by direct computation that the projection matrix $\mathbf{P}$ and annihilator $\mathbf{M}$ in fact 3.3.1 are both symmetric and idempotent.

Ex. 3.6.20. Let $\mathbf{A}$ be an $N \times N$ matrix.

1. Show that if $\mathbf{I}_N - \mathbf{A}$ is idempotent, then $\mathbf{A}$ is idempotent.
2. Show that if $\mathbf{A}$ is both symmetric and idempotent, then the matrix $\mathbf{I}_N - 2\mathbf{A}$ is orthogonal.

Ex. 3.6.21. Verify fact 3.3.2 (i.e., $\mathbf{MX} = \mathbf{0}$) directly using matrix algebra.

Ex. 3.6.22. Taking all notation as in §3.3.3 and adopting the assumptions of theorem 3.3.3 on page 119, let $\mathbf{V}_k$ be the $N \times k$ matrix formed by columns $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ for each $k = 1, \ldots, K$. Show that $\operatorname{span}(\mathbf{V}_k) \subset \operatorname{span}(\mathbf{B}_k)$ for each k.

Ex. 3.6.23. Continuing on from exercise 3.6.22, show that $\{\mathbf{v}_1, \ldots, \mathbf{v}_K\}$ is an orthogonal set.

Ex. 3.6.24. Following on from exercises 3.6.22 and 3.6.23, show that $\operatorname{span}(\mathbf{V}_k) = \operatorname{span}(\mathbf{B}_k)$ for each k. [8]

[6] Hint: Sketch the graph and think about it visually.
[7] Hint: In view of fact 1.4.14, it suffices to show that $\mathbf{B}'\mathbf{B}$ is positive definite.
[8] Hint: Use the results of both of these exercises and fact 1.1.8 on page 14.

Ex. 3.6.25. Let $\mathbf{X}$ be an $N \times K$ matrix with linearly independent columns and QR factorization $\mathbf{X} = \mathbf{QR}$. (See §3.3.4.) Fix $\mathbf{y} \in \mathbb{R}^N$. Show that $\hat{\boldsymbol{\beta}}$ defined in (3.6) can also be written as $\hat{\boldsymbol{\beta}} = \mathbf{R}^{-1}\mathbf{Q}'\mathbf{y}$.

Ex. 3.6.26. Prove the Cauchy-Schwarz inequality for random variables, which was first stated as fact 2.2.7 on page 60. That is, shown that for $x, y \in L_2$ we have $|\mathbb{E}[xy]| \leq \sqrt{\mathbb{E}[x^2]\mathbb{E}[y^2]}$. Use the results on orthogonal projections in L_2 as found in §3.4.1. If you get stuck, follow the solution to exercise 3.6.13, adjusting from vectors to L_2 as required.

Ex. 3.6.27. Show that the equality in (3.14) holds when x and w are independent.

Ex. 3.6.28. In fact 3.4.9, it is stated that if y is independent of the variables in $\mathcal{G}$, then $\mathbb{E}[y \,|\, \mathcal{G}] = \mathbb{E}[y]$. Prove this using the (second) definition of the conditional expectation $\mathbb{E}[y \,|\, \mathcal{G}]$. To make the proof a bit simpler, you can take $\mathcal{G} = \{x\}$.

Ex. 3.6.29. Confirm the claim in fact 3.4.9 that if x is $\mathcal{G}$-measurable, then $\mathbb{E}[xy \,|\, \mathcal{G}] = x\mathbb{E}[y \,|\, \mathcal{G}]$.

Ex. 3.6.30. Let $\operatorname{var}[y \,|\, x] := \mathbb{E}[y^2 \,|\, x] - (\mathbb{E}[y \,|\, x])^2$. Show that

$$\operatorname{var}[y] = \mathbb{E}[\operatorname{var}[y \,|\, x]] + \operatorname{var}[\mathbb{E}[y \,|\, x]]$$

Ex. 3.6.31. Show that the conditional expectation of a constant α is α. In particular, using the results in fact 3.4.9 (page 130) as appropriate, show that if α is a constant and $\mathcal{G}$ is any information set, then $\mathbb{E}[\alpha \,|\, \mathcal{G}] = \alpha$.

Ex. 3.6.32. Prove the claim in example 3.4.10. (Warning: The proof is a little advanced and you should be comfortable manipulating double integrals.)

3.6.1 Solutions to Selected Exercises

Solution to Exercise 3.6.3. Let $O = \{\mathbf{x}_1, \ldots, \mathbf{x}_K\} \subset \mathbb{R}^N$ be an orthogonal set that does not contain $\mathbf{0}$. Let $\alpha_1, \ldots, \alpha_K$ be such that $\sum_{k=1}^K \alpha_k \mathbf{x}_k = \mathbf{0}$. We claim that $\alpha_j = 0$ for any j. To see that this is so, fix j and take the inner product of both sides of $\sum_{k=1}^K \alpha_k \mathbf{x}_k = \mathbf{0}$ with respect to $\mathbf{x}_j$ to obtain $\alpha_j \|\mathbf{x}_j\|^2 = 0$. Since $\mathbf{x}_j \neq \mathbf{0}$, we conclude that $\alpha_j = 0$. The proof is done. □

Solution to Exercise 3.6.4. Let $\mathbf{Q}$ be an orthogonal matrix with columns $\mathbf{u}_1, \ldots, \mathbf{u}_N$. By the definition of matrix multiplication, the m, n-th element of $\mathbf{Q}'\mathbf{Q}$ is $\mathbf{u}_m'\mathbf{u}_n$, which

is 1 if $m = n$ and zero otherwise. Hence $\mathbf{Q}'\mathbf{Q} = \mathbf{I}$. It follows from fact 1.3.3 on page 24 that $\mathbf{Q}'$ is the inverse of $\mathbf{Q}$.

To see that $\det(\mathbf{Q}) \in \{-1, 1\}$, apply the results of fact 1.3.5 (page 25) and fact 1.4.4 (page 29) to the equality $\mathbf{Q}'\mathbf{Q} = \mathbf{I}$ to obtain $\det(\mathbf{Q})^2 = 1$. The claim follows. □

Solution to Exercise 3.6.5. Suppose that $\mathbf{A}$ is symmetric with eigenvalues $\lambda_1, \ldots, \lambda_N$. By theorem 3.1.1 we can decompose it as $\mathbf{A} = \mathbf{Q}\mathbf{D}\mathbf{Q}'$ where $\mathbf{D}$ is the diagonal matrix formed from eigenvalues and $\mathbf{Q}$ is an orthogonal matrix. Fixing $\mathbf{x} \in \mathbb{R}^N$ and letting $\mathbf{y} := \mathbf{Q}'\mathbf{x}$, we have

$$\mathbf{x}'\mathbf{A}\mathbf{x} = (\mathbf{Q}'\mathbf{x})'\mathbf{D}(\mathbf{Q}'\mathbf{x}) = \mathbf{y}'\mathbf{D}\mathbf{y} = \lambda_1 y_1^2 + \cdots + \lambda_N y_N^2 \tag{3.17}$$

Suppose that all eigenvalues are positive. Take $\mathbf{x}$ to be nonzero. The vector $\mathbf{y}$ must be nonzero (why?), and it follows from (3.17) that $\mathbf{x}'\mathbf{A}\mathbf{x} > 0$. Hence $\mathbf{A}$ is positive definite as claimed.

Conversely, suppose that $\mathbf{A}$ is positive definite. Fix $n \leq N$ and set $\mathbf{x} = \mathbf{Q}\mathbf{e}_n$. Evidently $\mathbf{x}$ is nonzero (why?). Hence $\mathbf{x}'\mathbf{A}\mathbf{x} > 0$. Since $\mathbf{Q}'$ is the inverse of $\mathbf{Q}$, it follows that

$$\lambda_n = \mathbf{e}_n'\mathbf{D}\mathbf{e}_n = (\mathbf{Q}'\mathbf{x})'\mathbf{D}\mathbf{Q}'\mathbf{x} = \mathbf{x}'\mathbf{Q}\mathbf{D}\mathbf{Q}'\mathbf{x} = \mathbf{x}'\mathbf{A}\mathbf{x} > 0$$

Since n was arbitrary, all eigenvalues are positive. □

Solution to Exercise 3.6.7. Let $S \subset \mathbb{R}^N$. We aim to show that $S \cap S^\perp = \{\mathbf{0}\}$. Fix $\mathbf{a} \in S \cap S^\perp$. Since $\mathbf{a} \in S^\perp$, we know that $\mathbf{a}'\mathbf{s} = 0$ for any $\mathbf{s} \in S$. Since $\mathbf{a} \in S$, we have in particular, $\mathbf{a}'\mathbf{a} = \|\mathbf{a}\|^2 = 0$. As we saw in fact 1.1.2, the only such vector is $\mathbf{0}$. □

Solution to Exercise 3.6.10. Fixing $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$ and letting $\mathbf{z} := \mathbf{x} - \mathbf{y}$ we have

$$\|\mathbf{Q}\mathbf{x} - \mathbf{Q}\mathbf{y}\|^2 = \|\mathbf{Q}\mathbf{z}\|^2 = (\mathbf{Q}\mathbf{z})'\mathbf{Q}\mathbf{z} = \mathbf{z}'\mathbf{Q}'\mathbf{Q}\mathbf{z} = \mathbf{z}'\mathbf{z} = \|\mathbf{z}\|^2 = \|\mathbf{x} - \mathbf{y}\|^2 \qquad \square$$

Solution to Exercise 3.6.11. Fix $\alpha, \beta \in \mathbb{R}$ and $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$. The claim is that

$$\mathbf{P}(\alpha\mathbf{x} + \beta\mathbf{y}) = \alpha\mathbf{P}\mathbf{x} + \beta\mathbf{P}\mathbf{y}$$

To verify this equality, we need to show that the right-hand side is the orthogonal projection of $\alpha\mathbf{x} + \beta\mathbf{y}$ onto S. Going back to theorem 3.2.1, we need to show that (i) $\alpha\mathbf{P}\mathbf{x} + \beta\mathbf{P}\mathbf{y} \in S$ and (ii) for any $\mathbf{z} \in S$, we have

$$(\alpha\mathbf{x} + \beta\mathbf{y} - (\alpha\mathbf{P}\mathbf{x} + \beta\mathbf{P}\mathbf{y}))'\mathbf{z} = 0$$

Here (i) is immediate, because $\mathbf{Px}$ and $\mathbf{Py}$ are in S by definition; and, moreover S is a linear subspace. To see that (ii) holds, just note that

$$(\alpha\mathbf{x} + \beta\mathbf{y} - (\alpha\mathbf{Px} + \beta\mathbf{Py}))'\mathbf{z} = \alpha(\mathbf{x} - \mathbf{Px})'\mathbf{z} + \beta(\mathbf{y} - \mathbf{Py})'\mathbf{z}$$

By definition, the projections of $\mathbf{x}$ and $\mathbf{y}$ are orthogonal to S, so we have $(\mathbf{x} - \mathbf{Px})'\mathbf{z} = (\mathbf{y} - \mathbf{Py})'\mathbf{z} = 0$. We are done. □

Solution to Exercise 3.6.13. Regarding part 1, the expression for $\mathbf{Py}$ given in exercise 3.6.13 can also be written as $\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y}$. Since $\mathbf{x}$ is a basis for $\operatorname{span}(\mathbf{x})$, the validity of this expression as the projection onto $\operatorname{span}(\mathbf{x})$ follows immediately from theorem 3.3.1. Regarding part 2, recall that orthogonal projections contract norms, so that, in particular, $\|\mathbf{Py}\| \leq \|\mathbf{y}\|$ must hold. Using our expression for $\mathbf{Py}$ from part 1 and rearranging gives the desired bound $|\mathbf{x}'\mathbf{y}| \leq \|\mathbf{x}\|\|\mathbf{y}\|$. □

Solution to Exercise 3.6.14. If $N = K$, then, in view of the full column rank assumption and theorem 1.3.3 on page 23, the matrix $\mathbf{X}$ is nonsingular. By fact 1.4.4 on page 29, $\mathbf{X}'$ is likewise nonsingular. Applying the usual rule for inverse of products (fact 1.3.4 on page 24), we have

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{X}^{-1}(\mathbf{X}')^{-1}\mathbf{X}'\mathbf{y} = \mathbf{X}^{-1}\mathbf{y}$$

This is of course the standard solution to the system $\mathbf{X}\boldsymbol{\beta} = \mathbf{y}$. □

Solution to Exercise 3.6.15. Let $\mathbf{x} = (x_1, x_2, x_3)$ be the closest point in S to $\mathbf{y}$. Note that $\mathbf{e}_1 \in S$ and $\mathbf{e}_2 \in S$. By the orthogonal projection theorem we have (i) $\mathbf{x} \in S$, and (ii) $\mathbf{y} - \mathbf{x} \perp S$. From (i) we have $x_3 = 0$. From (ii) we have

$$\langle \mathbf{y} - \mathbf{x}, \mathbf{e}_1 \rangle = 0 \quad \text{and} \quad \langle \mathbf{y} - \mathbf{x}, \mathbf{e}_2 \rangle = 0$$

These equations can be expressed more simply as $1 - x_1 = 0$ and $1 - x_2 = 0$. We conclude that $\mathbf{x} = (1, 1, 0)$. □

Solution to Exercise 3.6.16. It is false to say that $\mathbf{Px} \neq \mathbf{Py}$ whenever $\mathbf{x} \neq \mathbf{y}$: We can find examples of vectors $\mathbf{x}$ and $\mathbf{y}$ such that $\mathbf{x} \neq \mathbf{y}$ but $\mathbf{Px} = \mathbf{Py}$. Indeed, if we fix any $\mathbf{y}$ and then set $\mathbf{x} = \mathbf{Py} + \alpha\mathbf{My}$ for some constant α, you should be able to confirm that $\mathbf{Px} = \mathbf{Py}$, and also that $\mathbf{x} \neq \mathbf{y}$ when $\alpha \neq 1$. □

Solution to Exercise 3.6.17. Let $\mathbf{A} = \mathbf{B}'\mathbf{B}$. It suffices to show that $\mathbf{A}$ is positive definite, since this implies that its determinant is strictly positive, and any matrix with nonzero determinant is nonsingular. To see that $\mathbf{A}$ is positive definite, pick any $\mathbf{b} \neq \mathbf{0}$. We must show that $\mathbf{b}'\mathbf{A}\mathbf{b} > 0$. To see this, observe that

$$\mathbf{b}'\mathbf{A}\mathbf{b} = \mathbf{b}'\mathbf{B}'\mathbf{B}\mathbf{b} = (\mathbf{B}\mathbf{b})'\mathbf{B}\mathbf{b} = \|\mathbf{B}\mathbf{b}\|^2$$

By the properties of norms, this last term is zero only when $\mathbf{B}\mathbf{b} = \mathbf{0}$. But this is not true, because $\mathbf{b} \neq \mathbf{0}$ and $\mathbf{B}$ is full column rank (see fact 1.1.6). □

Solution to Exercise 3.6.18. By the definition of orthogonal projection we have $\hat{\mathbf{y}} \in \operatorname{span}(\mathbf{X})$, and hence there exists a vector $\mathbf{b}$ such that $\hat{\mathbf{y}} = \mathbf{X}\mathbf{b}$. Since $\mathbf{X}$ has linearly dependent columns, there exists a nonzero vector $\mathbf{a}$ such that $\mathbf{X}\mathbf{a} = \mathbf{0}$. Hence $\mathbf{X}\lambda\mathbf{a} = \mathbf{0}$ for all $\lambda \in \mathbb{R}$. For each such λ we have $\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} = \mathbf{X}\mathbf{b} + \mathbf{X}\lambda\mathbf{a} = \mathbf{X}(\mathbf{b} + \lambda\mathbf{a})$. □

Solution to Exercise 3.6.21. We have $\mathbf{M}\mathbf{X} = \mathbf{X} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} = \mathbf{0}$. □

Solution to Exercise 3.6.22. To see that $\operatorname{span}(\mathbf{V}_k) \subset \operatorname{span}(\mathbf{B}_k)$ for each k, observe first that (3.8) can be rewritten as $\mathbf{v}_k = \mathbf{b}_k - \mathbf{B}_{k-1}\mathbf{x}$ for suitable choice of $\mathbf{x}$. Hence $\mathbf{v}_k \in \operatorname{span}(\mathbf{B}_k)$. Since spans increase as we add more elements, it follows that $\mathbf{v}_j \in \operatorname{span}(\mathbf{B}_k)$ for $j \leq k$. Therefore $\operatorname{span}(\mathbf{V}_k) \subset \operatorname{span}(\mathbf{B}_k)$. □

Solution to Exercise 3.6.23. To show that $\{\mathbf{v}_1, \ldots, \mathbf{v}_K\}$ is an orthogonal set, we must show that arbitrary distinct elements are orthogonal, for which it suffices to check that $\mathbf{v}_k \perp \mathbf{v}_j$ whenever $j < k$. To see this, fix any such j and k. By construction, $\mathbf{v}_k$ lies in the orthogonal complement of $\operatorname{span}(\mathbf{B}_{k-1})$. On the other hand, as shown in the solution to exercise 3.6.22, we have $\mathbf{v}_j \in \operatorname{span}(\mathbf{B}_{k-1})$. Hence $\mathbf{v}_k \perp \mathbf{v}_j$. □

Solution to Exercise 3.6.24. We wish to confirm that $\operatorname{span}(\mathbf{V}_k) = \operatorname{span}(\mathbf{B}_k)$ holds. In exercise 3.6.22 we showed that $\operatorname{span}(\mathbf{V}_k) \subset \operatorname{span}(\mathbf{B}_k)$. As a result, it is enough to check that the columns of $\mathbf{V}_k$ are linearly independent (see, e.g., fact 1.1.8 on page 14). As we've just shown the columns of $\mathbf{V}_k$ are mutually orthogonal (see exercise 3.6.23), it suffices to show that none of them are zero (fact 3.1.3 on page 109). This is easy enough, for if $\mathbf{v}_k = \mathbf{0}$ then, by (3.8), we have $\mathbf{b}_k \in \operatorname{span}(\mathbf{B}_{k-1})$, which contradicts linear independence. □

Solution to Exercise 3.6.25. Let $\mathbf{X}$, $\mathbf{Q}$, $\mathbf{R}$ and $\mathbf{y} \in \mathbb{R}^N$ be as in the statement of the exercise. The claim is that $\hat{\boldsymbol{\beta}}$ defined in (3.6) is equal to $\tilde{\boldsymbol{\beta}} := \mathbf{R}^{-1}\mathbf{Q}'\mathbf{y}$. To show this, in

view of linear independence of the columns of $\mathbf{X}$, it suffices to show that $\mathbf{X}\tilde{\boldsymbol{\beta}} = \mathbf{X}\hat{\boldsymbol{\beta}}$, or

$$\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{Q}\mathbf{R}\mathbf{R}^{-1}\mathbf{Q}'\mathbf{y}$$

After simplifying, we see it suffices to show that $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{Q}\mathbf{Q}'$. Since $\mathbf{X}$ and $\mathbf{Q}$ have the same column space, this follows from theorem 3.3.1 on page 116. □

Solution to Exercise 3.6.27. Let g be any function from $\mathbb{R} \to \mathbb{R}$. Given independence of x and w (and applying fact 2.4.2 on page 71), we have

$$\begin{aligned}\mathbb{E}\left[(x + \mathbb{E}\left[w\right])g(x)\right] &= \mathbb{E}\left[xg(x)\right] + \mathbb{E}\left[w\right]\mathbb{E}\left[g(x)\right] \\ &= \mathbb{E}\left[xg(x)\right] + \mathbb{E}\left[wg(x)\right] \\ &= \mathbb{E}\left[(x + w)g(x)\right]\end{aligned}$$

This confirms (3.14). □

Solution to Exercise 3.6.28. Let y be independent of x. From the (second) definition of conditional expectation, to show that $\mathbb{E}\left[y \mid x\right] = \mathbb{E}\left[y\right]$ we need to show that

1. $\mathbb{E}\left[y\right]$ is $\mathcal{G}$-measurable, and
2. $\mathbb{E}\left[\mathbb{E}\left[y\right]g(x)\right] = \mathbb{E}\left[yg(x)\right]$ for any function $g\colon \mathbb{R} \to \mathbb{R}$.

Part 1 is immediate, because $\mathbb{E}\left[y\right]$ is constant (see example 3.4.7 on page 127). Regarding part 2, if g is any function, then by facts 2.4.1 and 2.4.2 (see page 71) we have $\mathbb{E}\left[yg(x)\right] = \mathbb{E}\left[y\right]\mathbb{E}\left[g(x)\right]$. By linearity of expectations, $\mathbb{E}\left[y\right]\mathbb{E}\left[g(x)\right] = \mathbb{E}\left[\mathbb{E}\left[y\right]g(x)\right]$. □

Solution to Exercise 3.6.29. We need to show that if x is $\mathcal{G}$-measurable, then $\mathbb{E}\left[xy \mid \mathcal{G}\right] = x\mathbb{E}\left[y \mid \mathcal{G}\right]$. To confirm this, we must show that

1. $x\mathbb{E}\left[y \mid \mathcal{G}\right]$ is $\mathcal{G}$-measurable, and
2. $\mathbb{E}\left[x\mathbb{E}\left[y \mid \mathcal{G}\right]z\right] = \mathbb{E}\left[xyz\right]$ for any $z \in L_2(\mathcal{G})$.

Regarding part 1, $\mathbb{E}\left[y \mid \mathcal{G}\right]$ is $\mathcal{G}$-measurable by definition, and x is $\mathcal{G}$-measurable by assumption, so $x\mathbb{E}\left[y \mid \mathcal{G}\right]$ is $\mathcal{G}$-measurable by fact 3.4.4 on page 128. Regarding part 2, fix $z \in L_2(\mathcal{G})$, and let $u := xz$. Since $x \in L_2(\mathcal{G})$, we have $u \in L_2(\mathcal{G})$. We need to show that

$$\mathbb{E}\left[\mathbb{E}\left[y \mid \mathcal{G}\right]u\right] = \mathbb{E}\left[yu\right]$$

Since $u \in L_2(\mathcal{G})$, this is immediate from the definition of $\mathbb{E}\left[y \mid \mathcal{G}\right]$. □

Solution to Exercise 3.6.31. By fact 3.4.9 (page 130), we know that if α is $\mathcal{G}$-measurable, then $\mathbb{E}[\alpha \mid \mathcal{G}] = \alpha$. Example 3.4.7 on page 127 tells us that α is indeed $\mathcal{G}$-measurable. □

Solution to Exercise 3.6.32. As in example 3.4.10, let x and y be random variables where $p(y \mid x)$ is the conditional density of y given x. Let $g(x) := \int t p(t \mid x) dt$. The claim is that $\mathbb{E}[y \mid x] = g(x)$. To prove this, we need to show that $g(x)$ is x-measurable, and that

$$\mathbb{E}[g(x)h(x)] = \mathbb{E}[yh(x)] \quad \text{for any function} \quad h \colon \mathbb{R} \to \mathbb{R} \tag{3.18}$$

The first claim is obvious. Regarding (3.18), let h be any such function. Using the notation in (2.22) on page 70, we can write

$$\begin{aligned}
\mathbb{E}[g(x)h(x)] &= \mathbb{E}\left[\int t p(t \mid x) dt \; h(x)\right] \\
&= \int \int t p(t \mid s) dt \; h(s) p(s) ds \\
&= \int \int t \frac{p(s,t)}{p(s)} dt \; h(s) p(s) ds \\
&= \int \int t \, h(s) p(s,t) dt ds
\end{aligned}$$

This is equal to the right-hand side of (3.18), and the proof is done. □

Part I

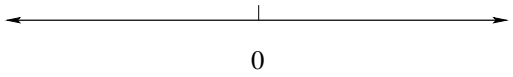
Appendices

Chapter 4

Appendix A: Analysis

4.1 Sets

In the course we often refer to the **real numbers**. This set is denoted by $\mathbb{R}$, and we understand it to contain "all of the numbers." $\mathbb{R}$ can be visualized as the "continuous" real line:



It contains both the rational and the irrational numbers.

What's "real" about the real numbers? Well, "real" is in contrast to "imaginary," where the latter refers to the set of imaginary numbers. Actually, the imaginary numbers are no more imaginary (or less real) than any other kind of numbers, but we don't need to talk any more about this.

$\mathbb{R}$ is an example of a **set**. A set is a collection of objects viewed as a whole. (In this case the objects are numbers.) Other examples of sets are the set of all rectangles in the plane, or the set of all monkeys in Japan.

If A is a set, then the statement $x \in A$ means that x is contained in (alternatively, is an element of) A. If B is another set, then $A \subset B$ means that any element of A is also an element of B, and we say that A is a **subset** of B. The statement $A = B$ means that A and B contain the same elements (each element of A is an element of B and

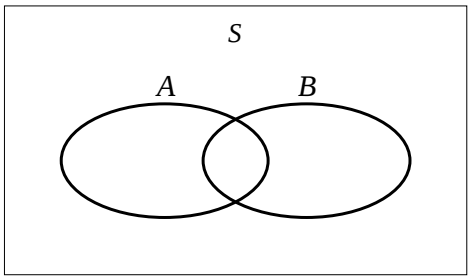


Figure 4.1: Sets A and B in S

vice versa). For example, if $\mathbb{I}$ is the irrational numbers,[1] then $\mathbb{I} \subset \mathbb{R}$. Also, $0 \in \mathbb{R}$, $\pi \in \mathbb{R}$, $-3 \in \mathbb{R}$, $e \in \mathbb{R}$, and so on.

Commonly used subsets of $\mathbb{R}$ include the intervals. For arbitrary a and b in $\mathbb{R}$, the **open inverval** (a, b) is defined as

$$(a, b) := \{x \in \mathbb{R} : a < x < b\}$$

while the **closed inverval** $[a, b]$ is defined as

$$[a, b] := \{x \in \mathbb{R} : a \leq x \leq b\}$$

We also use half open intervals such as $[a, b) := \{x \in \mathbb{R} : a \leq x < b\}$, half lines such as $(-\infty, b) = \{x \in \mathbb{R} : x < b\}$, and so on.

Let S be a set and let A and B be two subsets of S, as illustrated in figure 4.1. The **union** of A and B is the set of elements of S that are in A or B or both:

$$A \cup B := \{x \in S : x \in A \text{ or } x \in B\}$$

Here and below, "or" is used in the mathematical sense. It means "and/or". The **intersection** of A and B is the set of all elements of S that are in both A and B:

$$A \cap B := \{x \in S : x \in A \text{ and } x \in B\}$$

The set $A \setminus B$ is all points in A that are not points in B:

$$A \setminus B := \{x \in S : x \in A \text{ and } x \notin B\}$$

The **complement** of A is the set of elements of S that are not contained in A:

$$A^c := S \setminus A :=: \{x \in S : x \notin A\}$$

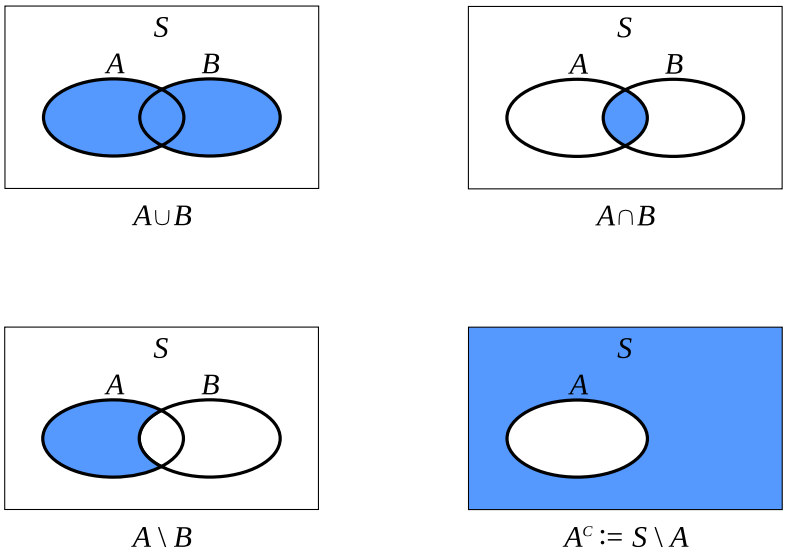


Figure 4.2: Unions, intersection and complements

Here $x \notin A$ means that x is not an element of A. Figure 4.2 illustrate these definitions.

For example, since $\mathbb{R}$ consists of the irrationals $\mathbb{I}$ and the rationals $\mathbb{Q}$, we have

$$\mathbb{Q} \subset \mathbb{R},\ \mathbb{I} \subset \mathbb{R},\ \mathbb{Q} \cup \mathbb{I} = \mathbb{R},\ \mathbb{Q}^c = \mathbb{I},\ \text{etc.}$$

Also,

$$\mathbb{N} := \{1, 2, 3, \ldots\} \subset \mathbb{Q} \subset \mathbb{R}$$

The **empty set** is, unsurprisingly, the set containing no elements. It is denoted by $\emptyset$. If the intersection of A and B equals $\emptyset$, then A and B are said to be **disjoint**.

The next fact lists some well known rules for set theoretic operations.

Fact 4.1.1. Let A and B be subsets of S. The following statements are true:

1. $A \cup B = B \cup A$ and $A \cap B = B \cap A$.
2. $(A \cup B)^c = B^c \cap A^c$ and $(A \cap B)^c = B^c \cup A^c$.

[1]The **irrationals** are those numbers such as π and $\sqrt{2}$ that cannot be expressed as fractions of whole numbers.

3. $A \setminus B = A \cap B^c$.

4. $(A^c)^c = A$.

4.2 Functions

There are two fundamental primitives in mathematics: sets and functions.[2] A brief discussion of sets is given in §4.1. Here we recall some basic definitions concerning functions.

Given arbitrary sets A and B, a **function** or **map** f from A to B is a rule that associates to each element a of A one and only one element of B. This element of B is usually called the **image of a under f**, and written $f(a)$. If we write $f\colon A \to B$, this means that f is a function from A to B.

Example 4.2.1. Think of the hands on an old school clock. If we know it's morning, then each position of the two hands is associated with one and only one time. If we don't know it's morning, one position of the hands is associated with two possible times, in am and pm. The relationship is no longer functional.

Consider figure 4.3. Bottom right is not a function because the middle point on the left-hand side is associated with two different points (images). Bottom left is not a function because the top point on the left-hand side is not associated with any image. From the definition, this is not allowed. Top right and top left are both functions.

If $f\colon A \to B$ and $g\colon B \to C$ then the function $h\colon A \to C$ defined by $h(a) = g(f(a))$ is called the **composition** of g and f, and written as $g \circ f$. For example, if $f\colon \mathbb{R} \to \mathbb{R}$ is defined by $f(x) = e^x := \exp(x)$ and $g\colon \mathbb{R} \to \mathbb{R}$ is defined by $g(x) = x^2$, then

$$(g \circ f)(x) = g(f(x)) = \exp(2x)$$

If a and b are points such that $f(a) = b$, then a is called a **preimage** of b. As shown in figure 4.3, the set of preimages of a can be empty, a singleton or contain multiple values. If $f\colon A \to B$ then the set A is called the **domain** of f. The set of points

$$\{b \in B : f(a) = b \text{ for some } a \in A\}$$

[2] Actually functions can be represented as a special type of set containing ordered pairs, and hence in pure mathematics cannot claim to be as foundational as sets. However, for our purposes, it will be fine to think of a function as a primitive in its own right, defined as a certain kind of "rule."

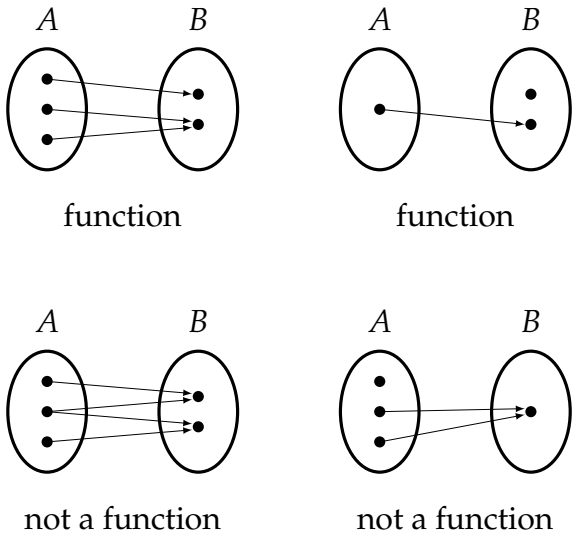


Figure 4.3: Functions and non-functions

is called the **range** of f, and written as $\text{rng}(f)$. Thus b is in the range of f if it has at least one preimage.

Figure 4.4 graphs a one-dimensional function $f\colon [0,1] \to \mathbb{R}$. The red interval represents the range of f. Also shown is the preimage x of a point $b \in \text{rng}(f)$.

A vast multitude of mathematical problems come down to finding the x such that $f(x) = b$ for given function f and constant (or vector, or any other object) b. There are two things that can go wrong here. One is that such an x might not be uniquely defined; in other words, there are multiple preimages of b under f. The other potential problem is lack of existence. This happens if b lies outside the range of f. Figure 4.5 gives an example of failure of uniqueness. Both x_1 and x_2 solve $f(x) = b$.

We use some additional language to keep track of when these problems will occur. A function $f\colon A \to B$ is called **one-to-one** if $f(a) = f(a')$ implies that $a = a'$. For example, top right in figure 4.3 is one-to-one while top left is not (because there exist distinct points a and a' with $f(a) = f(a')$). If $f\colon A \to B$ and f is one-to-one, then the equation $f(x) = b$ has *at most* one solution. (Why?)

A function $f\colon A \to B$ is called **onto** if $\text{rng}(f) = B$; this is if every $b \in B$ has at lease one preimage. For example, top left in figure 4.3 is onto but top left is not. If $f\colon A \to B$ and f is onto, then the equation $f(x) = b$ has *at least* one solution. (Why?)

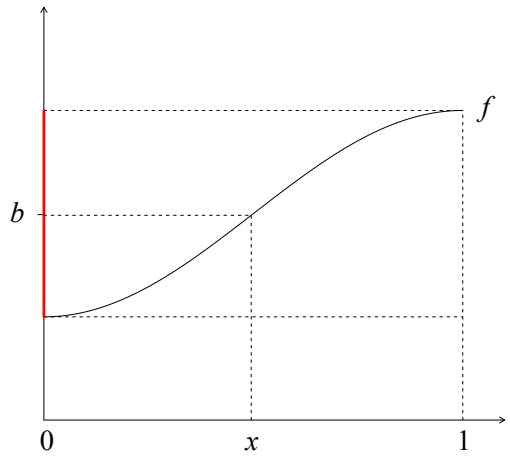


Figure 4.4: Preimage of b under f

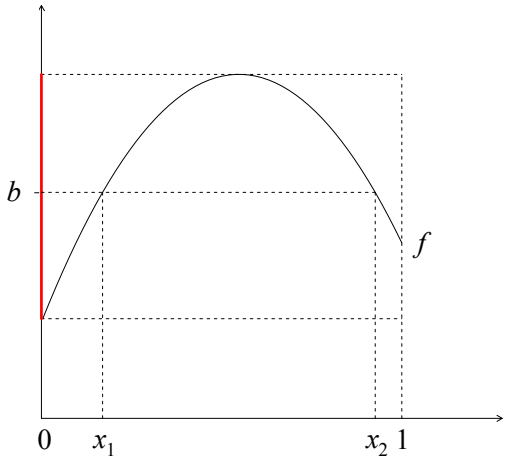


Figure 4.5: Multiple solutions

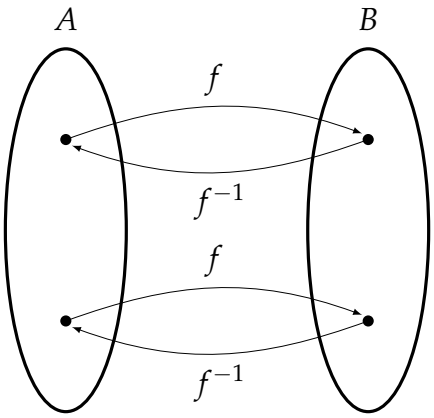


Figure 4.6: f is a bijection

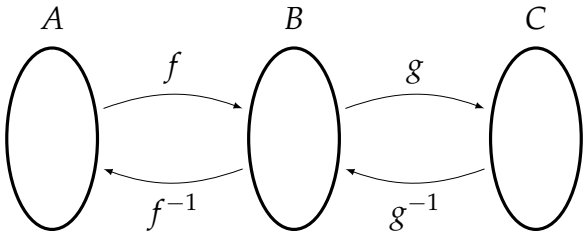


Figure 4.7: The inverse of $g \circ f$ is $f^{-1} \circ g^{-1}$

Finally, $f\colon A \to B$ is called a **bijection** if it is both one-to-one and onto. Bijections are sometimes called one-to-one correspondences. The nice thing about bijections, of course, is that the equation $f(x) = b$ always has exactly one solution. We can therefore define the **inverse function** to f, denoted by f^{-1}. In particular, f^{-1} is the map from $B \to A$ such that $f^{-1}(b)$ is the unique a with $f(a) = b$.

Fact 4.2.1. Let $f\colon A \to B$ and $g\colon B \to C$ be bijections.

1. f^{-1} is a bijection and its inverse is f
2. $f^{-1}(f(a)) = a$ for all $a \in A$
3. $f(f^{-1}(b)) = b$ for all $b \in B$
4. $g \circ f$ is a bijection from A to C and $(g \circ f)^{-1} = f^{-1} \circ g^{-1}$.

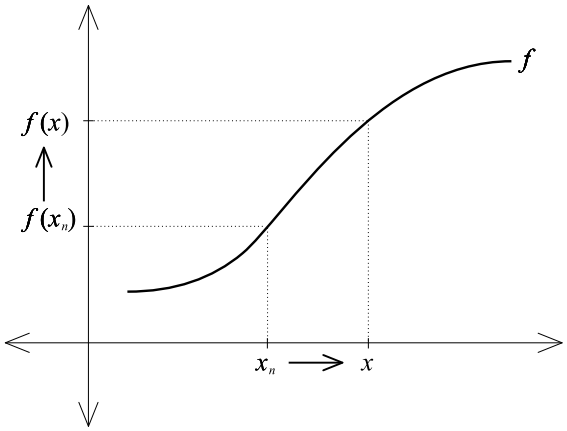


Figure 4.8: Continuity

4.2.1 Convergence and Continuity

Let $\{x_n\}_{n=1}^{\infty}$ be a sequence of real numbers. (For each $n = 1, 2, \ldots$ we have a corresponding $x_n \in \mathbb{R}$.) We say that x_n converges to 0 if, given any neighborhood of 0, the sequence points are eventually in that neighborhood. More formally (we won't use the formal definition, so feel free to skip this), given any $\epsilon > 0$, there exists an $N \in \mathbb{N}$ such that $|x_n| < \epsilon$ whenever $n \geq N$. Symbolically, $x_n \to 0$.

Now let $\{\mathbf{x}_n\}_{n=1}^{\infty}$ be a sequence of vectors in $\mathbb{R}^N$. We say that **$\mathbf{x}_n$ converges to $\mathbf{x} \in \mathbb{R}$** if $\|\mathbf{x}_n - \mathbf{x}\| \to 0$. Symbolically, $\mathbf{x}_n \to \mathbf{x}$. This is the fundamental notion of convergence in $\mathbb{R}^N$. Whole branches of mathematics are built on this idea.

Let $A \subset \mathbb{R}^N$ and $B \subset \mathbb{R}^M$. A function $f\colon A \to B$ is called **continuous at $\mathbf{x}$** if $f(\mathbf{x}_n) \to f(\mathbf{x})$ whenever $\mathbf{x}_n \to \mathbf{x}$, and **continuous** if it is continuous at $\mathbf{x}$ for all $\mathbf{x} \in A$. Figure 4.8 illustrates.

4.3 Real-Valued Functions

For any set A, if $f\colon A \to \mathbb{R}$ then f is called a **real-valued function**. If f and g are real-valued functions, then $f + g$ is defined by $(f + g)(x) = f(x) + g(x)$, while αf is defined by $(\alpha f)(x) = \alpha f(x)$. A **maximizer** of f on A is a point $a^* \in A$ such that

$$f(a^*) \geq f(a) \text{ for all } a \in A$$

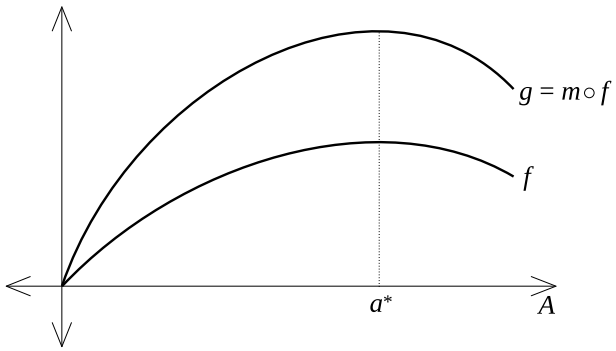


Figure 4.9: Monotone transforms preserve maximizers

The value $f(a^*)$ is called the **maximum** of f on A. A **minimizer** of f on A is a point $b^* \in A$ such that

$$f(b^*) \geq f(a) \text{ for all } a \in A$$

The value $f(b^*)$ is called the **minimum** of f on A.

Monotone increasing transformations of functions do not affect maximizers. To see this, let $f\colon \mathbb{R} \to \mathbb{R}$ and let m be a **monotone increasing function**, in the sense that if $x \leq x'$, then $m(x) \leq m(x')$, and let g be the function defined by $g(a) = m(f(a))$. Our claim is this:

Any maximizer of f on A is also a maximizer of g on A.

It's easy to see why this is the case. Let $a \in A$. Since a^* is a maximizer of f, it must be the case that $f(a) \leq f(a^*)$. Since m is monotone increasing, this implies that $m(f(a)) \leq m(f(a^*))$. Given that a was chosen arbitrarily, we have now shown that

$$g(a^*) \geq g(a) \text{ for all } a \in A$$

In other words, a^* is a maximizer of g on A.

Before finishing this topic, let's recall the notions of supremum and infimum. To illustrate, consider the function $f\colon (0,1) \to (0,1)$ defined by $f(x) = x$. It should be clear that f has no maximiser on $(0,1)$: given any $a^* \in (0,1)$, we can always choose another point $a^{**} \in (0,1)$ such that $a^{**} = f(a^{**}) > f(a^*) = a^*$. No maximizer exists and the optimization problem $\max_{x\in(0,1)} f(x)$ has no solution.

To get around this kind of problem, we often the notion of supremum instead. If A is a set, then the **supremum** $s := \sup A$ is the unique number s such that $a \leq s$ for

every $a \in A$, and, moreover, there exists a sequence $\{x_n\} \subset A$ such that $x_n \to s$. For example, 1 is the supremum of both $(0,1)$ and $[0,1]$. The **infimum** $i := \inf A$ is the unique number i such that $a \geq i$ for every $a \in A$, and, moreover, there exists a sequence $\{x_n\} \subset A$ such that $x_n \to i$. For example, 0 is the supremum of both $(0,1)$ and $[0,1]$.

One can show that the supremum and infimum of any bounded set A exist, and any set A when the values $-\infty$ and ∞ are admitted as a possible infima and supremum.

Returning to our original example with $f(x) = x$, while $\max_{x \in (0,1)} f(x)$ is not well defined, $\sup_{x \in (0,1)} f(x) := \sup\{f(x) : x \in (0,1)\} = \sup(0,1) = 1$.

Bibliography

[1] Amemiya, T. (1994): *Introduction to Statistics and Econometrics,* Harvard UP.

[2] Bishop, C.M. (2006): *Pattern Recognition and Machine Learning*, Springer.

[3] Casella, G. and R. L. Berger (1990): *Statistical Inference,* Duxbury Press, CA.

[4] Cheney, W. (2001): *Analysis for Applied Mathematics*, Springer.

[5] Dasgupta, A. (2008): *Asymptotic Theory of Statistics and Probability,* Springer.

[6] Davidson, R. (2000): *Econometric Theory,* Blackwell Publishing.

[7] Davidson, R., and MacKinnon, J. G. (1993): *Estimation and Inference in Econometrics*, OUP.

[8] Davidson, R., and MacKinnon, J. G. (2009): *Econometric Theory and Methods,* OUP.

[9] Devroye, Luc and Gabor Lugosi (2001): "Combinatorial Methods in Density Estimation" Springer-Verlag, New York.

[10] Doeblin, W. (1938): "Exposé de la theorie des chaîns simples constantes de Markov à un nombre fini d'états," *Rev. Math. Union Interbalkanique,* 2, 77–105.

[11] Durrett, R. (1996): *Probability: Theory and Examples*, Duxbury Press.

[12] Evans, G. and S. Honkapohja (2005): "An Interview with Thomas Sargent," *Macroeconomic Dynamics*, 9, 561–583.

[13] Freedman, D. A. (2009): *Statistical Models: Theory and Practice,* Cambridge UP.

[14] Greene, W. H. (1993): *Econometric Analysis,* Prentice-Hall, New Jersey.

[15] Hall, R. E. (1978): "Stochastic implications of the life cycle-permanent income hypothesis," *Journal of Political Economy* 86 (6), 971–87.

[16] Hamilton, J. D. (1994). *Time Series Analysis,* Princeton: Princeton university press.

[17] Hayashi, F. (2000): *Econometrics,* Princeton UP.

[18] Hoerl, A. E. and R. W. Kennard (1970): "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics,* 12 (1), 55–67.

[19] Jones, G. L. (2004): "On the Markov chain central limit theorem," *Probability Surveys,* 1, 5-1, 299–320.

[20] Kennedy, P. (2003): *A Guide to Econometrics.* MIT press.

[21] Marcus, M. and H. Minc (1965): *Introduction to Linear Algebra.* Dover Publications, N.Y.

[22] Olley, S. and A. Pakes (1996): "The Dynamics of Productivity in the Telecommunications Equipment Industry," *Econometrica,* 64 (6), 1263-97.

[23] Negri, I and Y. Nishiyama (2010): "Review on Goodness of Fit Tests for Ergodic Diffusion processes by Different Sampling Schemes," *Economic Notes,* 39, 91–106.

[24] Stachurski, J. and V. Martin (2008): "Computing the Distributions of Economic Models via Simulation," *Econometrica,* 76 (2), 443–450.

[25] Stachurski, J. (2009): *Economic Dynamics: Theory and Computation*, MIT Press.

[26] Van der Vaart, A. W. (2000): *Asymptotic statistics*, Cambridge university press.

[27] Vapnik, V. N. (2000): *The Nature of Statistical Learning Theory,* Springer-Verlag, New York.

[28] Wasserman, L. (2004): *All of Statistics,* Springer, New York.

[29] Williams, D. (1991): *Probabiliity with Martingales*, Cambridge Mathematical Textbooks.

[30] Wooldridge, J. M. (2010): *Econometric analysis of cross section and panel data,* MIT press.

Index