Chapter 1

RANDOM VARIABLES

1 Introduction

Coupling means the joint construction of two or more random variables (or processes), usually in order to deduce properties of the individual variables or gain insight into distributional similarities or relations between them. In this chapter and the next the method is introduced through a series of basic elementary examples. The arguments are carried out in full detail at an undergraduate level, suppressing measure-theoretic language. Advanced readers should be able to fill in any missing measure-theoretic notation or find it at the beginning of Chapter 3, where we return to the definition of coupling.

Let us spend a few lines on terminology before turning to the examples. A *copy* or a *representation* of a random variable X is a random variable $\hat{X}$ with the same distribution as X. Denote this by

$$\hat{X} \stackrel{D}{=} X.$$

A *coupling* of a collection of random variables X_i, $i \in \mathbb{I}$, [where $\mathbb{I}$ is some index set] is a family of random variables $(\hat{X}_i : i \in \mathbb{I})$ such that

$$\hat{X}_i \stackrel{D}{=} X_i, \quad i \in \mathbb{I}.$$

Note that only the individual $\hat{X}_i$ are copies of the individual X_i, while the whole family $(\hat{X}_i : i \in \mathbb{I})$ is typically not a copy of the family $(X_i : i \in \mathbb{I})$. In other words, the joint distribution of the $\hat{X}_i$ need not be the same as that of

the X_i. In fact, the X_i need not even have a (specified) joint distribution. On the other hand, we write $(\hat{X}_i : i \in \mathbb{I})$ in parentheses to stress that the $\hat{X}_i$ have a joint distribution. A trivial but often useful coupling is the *independence* coupling consisting of independent copies of the X_i.

Thus a coupling has fixed marginal distributions (the distributions of the individual X_i), and the trick is to find a dependence structure (joint distribution) that fits one's purposes.

2 The i.i.d. Coupling – Positive Correlation

A *self-coupling* of a random variable X is a family $(\hat{X}_i : i \in \mathbb{I})$ where each $\hat{X}_i$ is a copy of X. A trivial (and not so useful) self-coupling is the one with all the $\hat{X}_i$ identical. Another trivial self-coupling is the *i.i.d. coupling* consisting of independent copies of X. As an example of an efficient use of the i.i.d. coupling we shall prove the following result.

> For every random variable X and nondecreasing bounded functions f and g, the random variables $f(X)$ and $g(X)$ are positively correlated, that is,
>
> $$\mathrm{Cov}[f(X), g(X)] \geqslant 0. \tag{2.1}$$

In order to prove this claim let X' be an independent copy of X [thus (X, X') is an i.i.d. coupling of X]. The additivity of covariances yields

$$\begin{aligned}\mathrm{Cov}[f(X) - f(X'), g(X) - g(X')] &= \mathrm{Cov}[f(X), g(X)] \\ &\quad - \mathrm{Cov}[f(X), g(X')] - \mathrm{Cov}[f(X'), g(X)] + \mathrm{Cov}[f(X'), g(X')].\end{aligned}$$

Since X and X' are independent, the middle terms on the right are zero, and since X and X' have the same distribution, the remaining terms on the right are identical. Thus

$$\mathrm{Cov}[f(X), g(X)] = \tfrac{1}{2}\,\mathrm{Cov}[f(X) - f(X'), g(X) - g(X')].$$

Since the mean of both $f(X) - f(X')$ and $g(X) - g(X')$ is zero, we have

$$\begin{aligned}&\mathrm{Cov}[f(X) - f(X'), g(X) - g(X')] \\ &\quad = \mathbf{E}[(f(X) - f(X'))(g(X) - g(X'))],\end{aligned}$$

which is positive, since f and g nondecreasing implies that

$$f(x) - f(y) \text{ and } g(x) - g(y) \text{ are either both } \geqslant 0 \text{ or both } \leqslant 0.$$

Thus (2.1) holds.

3 Quantile Coupling – Stochastic Domination

In this section we produce a coupling that turns so-called stochastic domination into ordinary (pointwise) domination. Another application can be found in Section 8. See also Section 9.

3.1 The Coupling

Consider a random variable X with distribution function F, that is,

$$\mathbf{P}(X \leqslant x) = F(x), \quad x \in \mathbb{R}.$$

Let F^{-1} be the *generalized inverse* of F (or *quantile function*) defined by

$$F^{-1}(u) = \inf\{x \in \mathbb{R} : F(x) \geqslant u\}, \quad u \in [0,1].$$

Note that if F is continuous and strictly increasing, then F^{-1} is the ordinary inverse of F (see Figure 3.1).

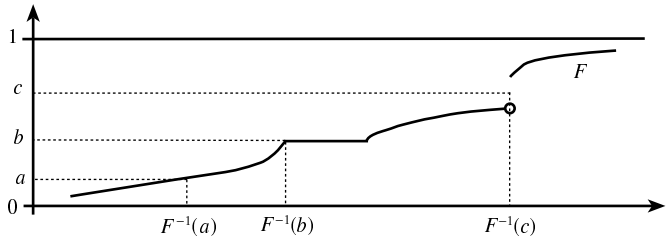

FIGURE 3.1. The generalized inverse F^{-1}.

Let U be uniform on $[0,1]$ (this is short for saying that U is a random variable that is uniformly distributed on $[0,1]$). Then the random variable

$$\hat{X} = F^{-1}(U)$$

is a copy of X, since [note that $F^{-1}(u) \leqslant x$ if and only if $u \leqslant F(x)$]

$$\mathbf{P}(\hat{X} \leqslant x) = \mathbf{P}(F^{-1}(U) \leqslant x) = \mathbf{P}(U \leqslant F(x)) = F(x), \quad x \in \mathbb{R}.$$

Thus letting F run over the class of all distribution functions (using the same U) yields a coupling of all differently distributed random variables. Call it the *quantile* coupling.

Since F^{-1} is nondecreasing, we have, according to Section 2, that the quantile coupling consists of positively correlated random variables. We might even think of this coupling as a maximal dependence coupling because knowing the value of only one of its variables, namely the value of U itself, gives us the value of all the others.

3.2 Application – Stochastic Domination

Let X and X' be two random variables with distribution functions F and G, respectively. If there is a coupling $(\hat{X}, \hat{X}')$ of X and X' such that $\hat{X}$ is *pointwise dominated* by $\hat{X}'$, that is,

$$\hat{X} \leqslant \hat{X}',$$

then $\{\hat{X} \leqslant x\} \supseteq \{\hat{X}' \leqslant x\}$, which implies $\mathbf{P}(\hat{X} \leqslant x) \geqslant \mathbf{P}(\hat{X}' \leqslant x)$ and thus

$$F(x) \geqslant G(x), \quad x \in \mathbb{R}. \tag{3.1}$$

If (3.1) holds, then X is said to be *stochastically dominated* (or *dominated in distribution*) by X'. Denote this by

$$X \overset{D}{\leqslant} X'.$$

We shall now show that the quantile coupling turns stochastic domination back into pointwise domination: due to (3.1), $G(x) \geqslant u$ implies $F(x) \geqslant u$ and thus

$$\{x \in \mathbb{R} : G(x) \geqslant u\} \subseteq \{x \in \mathbb{R} : F(x) \geqslant u\}$$

and thus $F^{-1}(u) \leqslant G^{-1}(u)$, which yields $F^{-1}(U) \leqslant G^{-1}(U)$ [see Figure 3.2].

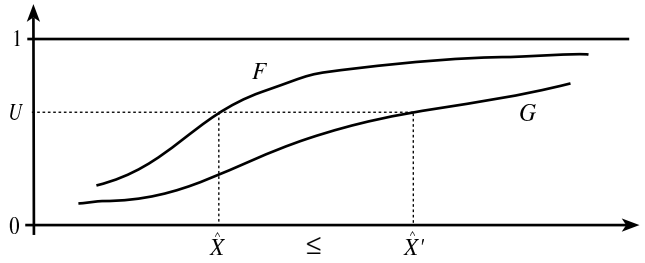

FIGURE 3.2. Turning stochastic domination into pointwise domination.

We have established the following result.

Theorem 3.1. *Let X and X' be random variables. Then*

$$X \overset{D}{\leqslant} X'$$

if and only if there is a coupling $(\hat{X}, \hat{X}')$ of X and X' such that

$$\hat{X} \leqslant \hat{X}'.$$

3.3 What For?

The direct usefulness of Theorem 3.1 is mainly due to the fact that it is easier to carry out arguments using pointwise domination than stochastic domination. As an illustration of this we shall prove the following result.

Corollary 3.1. *Let X_1, X_2, X_1', and X_2' be random variables such that*

X_1 *and* X_2 *are independent,*

X_1' *and* X_2' *are independent,*

$X_1 \overset{D}{\leqslant} X_1'$ *and* $X_2 \overset{D}{\leqslant} X_2'$.

Then

$$X_1 + X_2 \overset{D}{\leqslant} X_1' + X_2'. \tag{3.2}$$

Proof. Let $(\hat{X}_1, \hat{X}_1')$ be a coupling of X_1 and X_1' such that $\hat{X}_1 \leqslant \hat{X}_1'$ and let $(\hat{X}_2, \hat{X}_2')$ be a coupling of X_2 and X_2' such that $\hat{X}_2 \leqslant \hat{X}_2'$. Let $(\hat{X}_1, \hat{X}_1')$ and $(\hat{X}_2, \hat{X}_2')$ be independent. Then $(\hat{X}_1 + \hat{X}_2, \hat{X}_1' + \hat{X}_2')$ is a coupling of $X_1 + X_2$ and $X_1' + X_2'$, and

$$\hat{X}_1 + \hat{X}_2 \leqslant \hat{X}_1' + \hat{X}_2'.$$

This implies (3.2). □

A more substantial example of obtaining a distributional result through a pointwise argument by way of coupling can be found in Section 9, where we use the quantile coupling to obtain the distributional version of dominated convergence from the standard pointwise version.

3.4 On the General Coupling Idea

Theorem 3.1 and Corollary 3.1, illustrate two general points about coupling. Firstly, a coupling characterization of a distributional property deepens our understanding of that property: according to Theorem 3.1, stochastic domination is simply the distributional form of pointwise domination. Secondly, the coupling characterization can also be directly useful because it is easier to argue pointwise (as in the proof of Corollary 3.1) than in distribution.

3.5 Variations on the Quantile Coupling

Let U be uniform on $[0, 1]$ and define

$$\hat{X} = F^{-1}(U), \quad \hat{X}' = G^{-1}(1 - U).$$

Then $(\hat{X}, \hat{X}')$ is still a coupling of X and X', since $1 - U$ is also uniform on $[0, 1]$. Now $G^{-1}(1 - u)$ is nonincreasing in u, and an obvious modification

of the last three lines in Section 2 yields that $\hat{X}$ and $\hat{X}'$ are negatively correlated.

More generally, if we put

$$\hat{X} = F^{-1}((a + bU) \bmod 1), \quad \hat{X}' = G^{-1}((c + dU) \bmod 1),$$

where $a, c \in \mathbb{R}$ and $b, d = \pm 1$, and $x \bmod 1$ means the fractional part of x,

$$x \bmod 1 = x - [x],$$

then $(\hat{X}, \hat{X}')$ is a coupling of X and X'. Here we could even allow a, b, c, and d to be random variables that are independent of U.

If we take $G = F$, these modifications of the quantile coupling yield a nontrivial self-coupling of X.

The quantile approach is used heavily in simulation to generate random variables with specified distributions.

3.6 Comment

Suppose $X \overset{D}{\leqslant} X'$ and apply Theorem 3.1 to obtain a coupling $(\hat{X}, \hat{X}')$ such that $\hat{X} \leqslant \hat{X}'$. Then for each bounded nondecreasing function g we have $g(\hat{X}) \leqslant g(\hat{X}')$ and thus

$$\mathbf{E}[g(X)] \leqslant \mathbf{E}[g(X')]. \tag{3.3}$$

Conversely, suppose (3.3) holds for all bounded nondecreasing functions g. Fix an $x \in \mathbb{R}$ and take $g = 1_{(x,\infty)}$ to obtain from (3.3) that

$$\mathbf{P}(X > x) \leqslant \mathbf{P}(X' > x), \quad x \in \mathbb{R}.$$

Thus $X \overset{D}{\leqslant} X'$ if and only if (3.3) holds for all bounded nondecreasing g. This is taken as the definition of stochastic domination in higher dimensions and, more generally, in partially ordered spaces. Theorem 3.1 can in fact be extended to partially ordered Polish spaces; cf. Lindvall (1992), Chapter IV.1 (Strassen's theorem).

4 Coupling Event – Maximal Coupling

Let X_i, $i \in \mathbb{I}$, be a collection of discrete or continuous random variables. We shall construct a coupling such that the variables coincide maximally. We first treat the discrete case and start by establishing an upper bound on the coincidence probability. In Section 5 we give an application of this coupling.

4.1 The Coupling Event Inequality – Discrete Variables

Suppose $(\hat{X}_i : i \in \mathbb{I})$ is a coupling of X_i, $i \in \mathbb{I}$, and let C be an event such that if C occurs, then all the $\hat{X}_i$ coincide, that is,

$$C \subseteq \{\hat{X}_i = \hat{X}_j \text{ for all } i, j \in \mathbb{I}\}.$$

Call such an event a *coupling event*.

Consider first the discrete case: let all the X_i take values in a finite or countable set E and denote the *probability mass functions* by p_i, that is, for $x \in E$,

$$\mathbf{P}(X_i = x) = p_i(x).$$

For all $i, j \in \mathbb{I}$ and $x \in E$ we have

$$\mathbf{P}(\hat{X}_i = x, C) = \mathbf{P}(\hat{X}_j = x, C) \leqslant p_j(x)$$

and thus for all $i \in \mathbb{I}$ and $x \in E$

$$\mathbf{P}(\hat{X}_i = x, C) \leqslant \inf_{j \in \mathbb{I}} p_j(x).$$

Summing over $x \in E$ yields the following basic *coupling event inequality*.

Theorem 4.1. *If C is a coupling event of a coupling of discrete random variables X_i, $i \in \mathbb{I}$, taking values in a finite or countable set E, then*

$$\mathbf{P}(C) \leqslant \sum_{x \in E} \inf_{i \in \mathbb{I}} p_i(x). \tag{4.1}$$

4.2 Maximal Coupling – Discrete Variables

We shall now construct a coupling with a coupling event C such that (4.1) holds with identity. Call such a coupling *maximal* and C a *maximal* coupling event. Put

$$c := \sum_{x \in E} \inf_{i \in \mathbb{I}} p_i(x) \qquad \text{(the maximal coupling probability)}.$$

If $c = 0$, take the $\hat{X}_i$ independent and $C = \varnothing$. If $c = 1$, take the $\hat{X}_i$ identical and $C = \Omega =$ the set of all outcomes. If $0 < c < 1$, let us mix these couplings as follows. Let

$$I, V, \text{ and } W_i, i \in \mathbb{I}, \text{ be independent random variables}$$

such that

$$I \text{ is 0-1 valued with } \mathbf{P}(I = 1) = c,$$

$$\mathbf{P}(V = x) = \inf_{i \in \mathbb{I}} p_i(x)/c, \qquad x \in E.$$

$$\mathbf{P}(W_i = x) = (p_i(x) - c\mathbf{P}(V = x))/(1 - c), \quad x \in E.$$

Define, for each $i \in \mathbb{I}$,

$$\hat{X}_i = \begin{cases} V & \text{if } I = 1, \\ W_i & \text{if } I = 0. \end{cases} \tag{4.2}$$

Then

$$\begin{aligned} \mathbf{P}(\hat{X}_i = x) &= \mathbf{P}(V = x)\mathbf{P}(I = 1) + \mathbf{P}(W_i = x)\mathbf{P}(I = 0) \\ &= \mathbf{P}(X_i = x). \end{aligned}$$

Moreover, $C = \{I = 1\}$ is a coupling event and $\mathbf{P}(C)$ has the desired value c. We have established the following result.

Theorem 4.2. *Suppose X_i, $i \in \mathbb{I}$, are discrete random variables taking values in a finite or countable set E. Then there exists a maximal coupling, that is, a coupling with coupling event C such that*

$$\mathbf{P}(C) = \sum_{x \in E} \inf_{i \in \mathbb{I}} p_i(x).$$

4.3 The Coupling Event Inequality – Continuous Variables

Now let the X_i be continuous random variables with densities f_i, that is, for intervals A

$$\mathbf{P}(X_i \in A) = \int_A f_i \qquad (\text{which is short for } \int_A f_i(x)\,dx).$$

It is a little harder to establish the coupling event inequality in this case, and we shall make the simplifying assumption that the X_i are either finitely or countably many, that is,

$$\mathbb{I} = \{1, \ldots, n\} \qquad \text{or} \qquad \mathbb{I} = \{1, 2, \ldots\}.$$

Suppose $(\hat{X}_i : i \in \mathbb{I})$ is a coupling of $X_i, i \in \mathbb{I}$, and C is a coupling event. Then, for intervals (Borel sets) A and $i, j \in \mathbb{I}$,

$$\mathbf{P}(\hat{X}_i \in A, C) = \mathbf{P}(\hat{X}_j \in A, C) \leqslant \int_A f_j. \tag{4.3}$$

Consider first the finite case $\mathbb{I} = \{1, \ldots, n\}$ and define a partition of $\mathbb{R}$ by

$$A_1 = \{x \in \mathbb{R} : f_1(x) = \inf_{1 \leqslant j \leqslant n} f_j(x)\}$$

and recursively for $1 < k \leqslant n$

$$A_k = \{x \in \mathbb{R} : f_k(x) = \inf_{1 \leqslant j \leqslant n} f_j(x)\} \setminus (A_1 \cup \cdots \cup A_{k-1}).$$

Then (4.3) yields the inequality in

$$\mathbf{P}(\hat{X}_i \in A \cap A_k, C) \leqslant \int_{A\cap A_k} f_j = \int_{A\cap A_k} \inf_{1\leqslant j\leqslant n} f_k, \tag{4.4}$$

while the equality follows from the definition of A_k. Sum over $k \in \mathbb{I}$ to obtain, in the finite case, that

$$\mathbf{P}(\hat{X}_i \in A, C) \leqslant \int_A \inf_{j\in\mathbb{I}} f_j, \quad i \in \mathbb{I}. \tag{4.5}$$

In the countable case $\mathbb{I} = \{1, 2, \dots\}$ fix $n < \infty$ to obtain that (4.4) still holds for $i, k \leqslant n$. This yields (4.5) with $\inf_{j\in\mathbb{I}} f_j$ replaced by $\inf_{1\leqslant j\leqslant n} f_j$. Sending $n \to \infty$ yields (4.5), since $\inf_{1\leqslant j\leqslant n} f_j$ decreases to $\inf_{j\in\mathbb{I}} f_j$.

Take $A = \mathbb{R}$ in (4.5) to obtain the following *coupling event inequality*.

Theorem 4.3. *If C is a coupling event of a coupling of continuous random variables with densities $f_1, f_2, \dots$ (or $f_1, \dots, f_n$), then*

$$\mathbf{P}(C) \leqslant \int \inf_i f_i. \tag{4.6}$$

4.4 Maximal Coupling – Continuous Variables

Call a coupling and event achieving identity in (4.6) *maximal*. The construction in Section 4.2 extends with an obvious modification to the continuous case. Put

$$c := \int \inf_{i\in\mathbb{I}} f_i \qquad \text{(the maximal coupling probability)}.$$

If $c = 0$, take the $\hat{X}_i$ independent and $C = \varnothing$. If $c = 1$, take the $\hat{X}_i$ identical and $C = \Omega$. If $0 < c < 1$, mix these couplings as follows. Let

I, V, and W_i, $i \in \mathbb{I}$, be independent random variables

such that

I is 0-1 valued with $\mathbf{P}(I = 1) = c$,

V has density $\inf_{i\in\mathbb{I}} f_i/c$,

W_i has density $(f_i - \inf_{j\in\mathbb{I}} f_j)/(1 - c)$.

Define $\hat{X}_i$ by (4.2). Then $(\hat{X}_i : i \in \mathbb{I})$ is a coupling of the X_i, since for intervals A,

$$\begin{aligned}\mathbf{P}(\hat{X}_i \in A) &= \mathbf{P}(V \in A)\mathbf{P}(I = 1) + \mathbf{P}(W_i \in A)\mathbf{P}(I = 0)\\ &= \mathbf{P}(X_i \in A).\end{aligned}$$

Moreover, $C = \{I = 1\}$ is a coupling event, and $\mathbf{P}(C)$ has the desired value. We have established the following result.

Theorem 4.4. *Suppose $X_1, X_2, \ldots$ (or $X_1, \ldots, X_n$) are continuous random variables with densities $f_1, f_2, \ldots$ (or $f_1, \ldots, f_n$). Then there exists a maximal coupling, that is, a coupling with coupling event C such that*

$$\mathbf{P}(C) = \int \inf_i f_i.$$

4.5 Comments

It is often natural to take

$$C = \{\hat{X}_i = \hat{X}_j \text{ for all } i, j \in \mathbb{I}\}.$$

By definition, any coupling event of $(\hat{X}_i : i \in I)$ is contained in this set, and thus the maximal couplings in Theorems 4.2 and 4.4 are also maximal with this choice of C.

In particular, for two discrete random variables X and X' there exists a coupling $(\hat{X}, \hat{X}')$ such that, with $\wedge$ denoting minimum,

$$\mathbf{P}(\hat{X} = \hat{X}') = \sum_x \mathbf{P}(X = x) \wedge \mathbf{P}(X' = x) \tag{4.7}$$

and for two continuous random variables X and X' with densities f and f' there exists a coupling $(\hat{X}, \hat{X}')$ such that

$$\mathbf{P}(\hat{X} = \hat{X}') = \int f \wedge f' \qquad \text{(see Figure 4.1).} \tag{4.8}$$

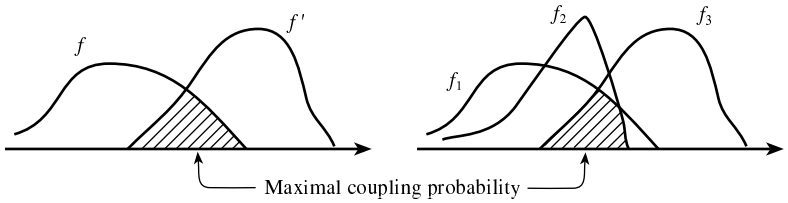

FIGURE 4.1. The maximal coupling probability.

Call these couplings *maximal* (without a reference to a particular coupling event).

In simulation a maximal coupling of continuous random variables X and X' with densities f and g can be generated as follows. Choose a point uniformly at random under the f-curve and let its x-coordinate be a realization of $\hat{X}$. If the point happens to be under the g-curve, let its x-coordinate also be a realization of $\hat{X}'$. If not, choose a new point uniformly at random under

the g-curve and above the f-curve and let its x-coordinate be a realization of $\hat{X}'$.

This simulation procedure extends to a collection $X_1, \ldots, X_n$ of random variables with densities $f_1, \ldots, f_n$ as follows. Choose a point uniformly at random under f_1, consider the densities under which this point falls, and let its x-coordinates be the realizations of the corresponding variables. Then pick a point uniformly at random above these densities and under one of the remaining densities, consider the densities under which the point falls, and let its x-coordinates be the realizations of the corresponding variables. Repeat this until no density remains. This yields a coupling $(\hat{X}_1, \ldots, \hat{X}_n)$ such that all subcollections $(\hat{X}_{n_1}, \ldots, \hat{X}_{n_k})$ are maximal couplings. In fact, repeating this ad infinitum yields a coupling of a countable collection of continuous random variables such that each subcollection is a maximal coupling.

We shall refer to the representation (4.2) of X_i as a *splitting* representation.

In Chapter 3 (Section 7) we extend the results of this section to arbitrary collections of random elements.

5 Poisson Approximation – Total Variation

The following well-known approximation

$$\mathrm{Bin}(n,p) \approx \mathrm{Poi}(np) \tag{5.1}$$

can be established and made precise by coupling.

5.1 Approximating a 0-1 Variable

Let X be a 0-1 variable with $\mathbf{P}(X = 1) = p$ where $0 \leqslant p \leqslant 1$ and let X' be Poisson p. Let $(\hat{X}, \hat{X}')$ be a maximal coupling of X and X'. In order to determine the maximal coupling probability $\mathbf{P}(\hat{X} = \hat{X}')$, recall that for all real x it holds that $1 + x \leqslant e^x$, which yields

$$\mathbf{P}(X = 0) = 1 - p \leqslant e^{-p} = \mathbf{P}(X' = 0),$$

and note that

$$\mathbf{P}(X = 1) = p \geqslant pe^{-p} = \mathbf{P}(X' = 1).$$

This and (4.7) yields

$$\begin{aligned}\mathbf{P}(\hat{X} = \hat{X}') &= \mathbf{P}(X = 0) \wedge \mathbf{P}(X' = 0) + \mathbf{P}(X = 1) \wedge \mathbf{P}(X' = 1) \\ &= 1 - p + pe^{-p}.\end{aligned}$$

Since $e^{-p} \geqslant 1 - p$, this implies that $\mathbf{P}(\hat{X} = \hat{X}') \geqslant 1 - p^2$ and thus

$$\mathbf{P}(\hat{X} \neq \hat{X}') \leqslant p^2. \tag{5.2}$$

5.2 Sums of Independent 0-1 Variables

Let $X_1, \ldots, X_n$ be independent 0-1 variables with $\mathbf{P}(X_i = 1) = p_i$, where $0 \leqslant p_i \leqslant 1$. Put

$$X = X_1 + \cdots + X_n.$$

Let $X_1', \ldots, X_n'$ be independent Poisson variables, X_i' with parameter p_i. Recall that

$$X' := X_1' + \cdots + X_n' \text{ is Poisson } p_1 + \cdots + p_n.$$

Let $(\hat{X}_1, \hat{X}_1'), \ldots, (\hat{X}_n, \hat{X}_n')$ be independent pairs such that for each i, $(\hat{X}_i, \hat{X}_i')$ is a maximal coupling of X_i and X_i'. Put

$$\hat{X} = \hat{X}_1 + \cdots + \hat{X}_n \quad \text{and} \quad \hat{X}' = \hat{X}_1' + \cdots + \hat{X}_n'.$$

Then $(\hat{X}, \hat{X}')$ is a coupling of X and X', and

$$\mathbf{P}(\hat{X} \neq \hat{X}') \leqslant \mathbf{P}(\hat{X}_i \neq \hat{X}_i' \text{ for some } i) \leqslant \sum_{1 \leqslant i \leqslant n} \mathbf{P}(\hat{X}_i \neq \hat{X}_i').$$

Applying (5.2) yields

$$\mathbf{P}(\hat{X} \neq \hat{X}') \leqslant \sum_{1 \leqslant i \leqslant n} p_i^2. \tag{5.3}$$

If we take $p_i = p$, then X is binomial (n, p), and thus we have the following clear and intuitively appealing random variable formulation of (5.1):

$$\text{Bin}(n, p) \text{ differs from } \text{Poi}(np) \text{ with probability at most } np^2.$$

In order to use the above coupling to formulate (5.1) in terms of total variation distance between distributions we take an excursion into that topic for the next two subsections.

5.3 Total Variation – Definition and Identities

Let X and X' be random variables with distributions λ and μ, that is, for each (Borel) set A,

$$\lambda(A) = \mathbf{P}(X \in A) \qquad \text{and} \qquad \mu(A) = \mathbf{P}(X' \in A).$$

The *total variation* distance between λ and μ is simply twice the supremum distance

$$\|\lambda - \mu\| := 2 \sup_A |\lambda(A) - \mu(A)|. \tag{5.4}$$

The reason for multiplying by 2 and using the phrase 'total variation' is the following. Suppose X and X' are discrete with probability mass functions p and q, or continuous with densities f and g. Then twice the supremum of $\lambda - \mu$ equals the actual total variation (the *total* of the variation) of $p - q$, or $f - g$, namely

$$\|\lambda - \mu\| = \sum_x |p(x) - q(x)| \quad \text{or} \quad \|\lambda - \mu\| = \int |f - g|. \tag{5.5}$$

We shall establish (5.5) and two other useful identities:

Theorem 5.1. *If X and X' are discrete with probability mass functions p and q, or continuous with densities f and g, then* (5.5) *holds and*

$$\|\lambda - \mu\| = 2\sum_x (p(x) - q(x))^+ \quad \textit{or } \|\lambda - \mu\| = 2\int (f - g)^+, \tag{5.6}$$

$$\|\lambda - \mu\| = 2 - 2\sum_x p(x) \wedge q(x) \ \textit{ or } \|\lambda - \mu\| = 2 - 2\int f \wedge g. \tag{5.7}$$

Here we have used the following standard notation: for real numbers a and b let

$$a^+ = a \vee 0, \qquad \text{where} \quad a \vee b = \text{ maximum of } a \text{ and } b,$$
$$a^- = -(a \wedge 0), \quad \text{where} \quad a \wedge b = \text{ minimum of } a \text{ and } b.$$

PROOF. We shall carry out the proof of Proposition 5.1 in the discrete case, the continuous case is analogous. It is clear that for sets A,

$$\lambda(A) - \mu(A) \leqslant \sum_x (p(x) - q(x))^+$$

and that equality holds if we take $A = \{x : p(x) > q(x)\}$. Thus

$$\sup_A (\lambda(A) - \mu(A)) = \sum_x (p(x) - q(x))^+, \tag{5.8}$$

and similarly,

$$\sup_A (\mu(A) - \lambda(A)) = \sum_x (p(x) - q(x))^-. \tag{5.9}$$

From $\sum_x p(x) = 1 = \sum_x q(x)$ it follows that

$$\sum_x (p(x) - q(x))^+ = \sum_x (p(x) - q(x))^-. \tag{5.10}$$

Combining (5.8), (5.9), and (5.10) yields

$$\sup_A |\lambda(A) - \mu(A)| = \sum_x (p(x) - q(x))^+,$$

and thus (5.6) holds. From

$$|p - q| = (p - q)^+ + (p - q)^-$$

together with (5.6) and (5.10) we obtain (5.5). Finally,

$$(p - q)^+ = p - p \wedge q$$

together with (5.6) and $\sum_x p(x) = 1$ yields (5.7). □

5.4 Total Variation and Coupling

Let $(\hat{X}, \hat{X}')$ be a coupling of two random variables X and X', and let C be a coupling event. Since C implies that $\hat{X} = \hat{X}'$, we have for (Borel) sets A,

$$\mathbf{P}(\hat{X} \in A, C) = \mathbf{P}(\hat{X}' \in A, C)$$

and thus

$$\begin{aligned} \mathbf{P}(X \in A) - \mathbf{P}(X' \in A) &= \mathbf{P}(\hat{X} \in A) - \mathbf{P}(\hat{X}' \in A) \\ &= \mathbf{P}(\hat{X} \in A, C^c) - \mathbf{P}(\hat{X}' \in A, C^c) \\ &\leqslant \mathbf{P}(C^c). \end{aligned}$$

Apply (5.4) to obtain the *coupling event inequality*

$$\|\mathbf{P}(X \in \cdot) - \mathbf{P}(X' \in \cdot)\| \leqslant 2\mathbf{P}(C^c). \tag{5.11}$$

From (5.7) we see that in the discrete and continuous cases this is just a total variation formulation of Theorems 4.1 and 4.3 specialized to two variables. We also see that the coupling is maximal if and only if identity holds in (5.11). Thus when $(\hat{X}, \hat{X}')$ is a maximal coupling, we have

$$\|\mathbf{P}(X \in \cdot) - \mathbf{P}(X' \in \cdot)\| = 2\mathbf{P}(\hat{X} \neq \hat{X}'). \tag{5.12}$$

5.5 Back to the Poisson Approximation

Combining (5.3) and the coupling event inequality in the form (5.11) [and with $C = \{\hat{X} = \hat{X}'\}$] yields

$$\|\mathbf{P}(X \in \cdot) - \mathrm{Poi}(p_1 + \cdots + p_n)\| \leqslant 2 \sum_{i=1}^n p_i^2.$$

In particular, if $p_i = p$, then X is binomial (n, p), and thus we have the following precise formulation of (5.1):

$$\| \operatorname{Bin}(n,p) - \operatorname{Poi}(np) \| \leqslant 2np^2.$$

If a Poisson parameter c is given and $n \geqslant c$, then taking $p = c/n$ yields

$$\| \operatorname{Bin}(n,c/n) - \operatorname{Poi}(c) \| \leqslant 2c^2/n.$$

Sending n to infinity yields in particular, with $\overset{tv}{\to}$ denoting convergence in total variation,

$$\operatorname{Bin}(n,c/n) \overset{tv}{\to} \operatorname{Poi}(c), \quad n \to \infty, \tag{5.13}$$

which further implies

$$\binom{n}{x} (c/n)^x (1-c/n)^{n-x} \to e^{-c} \frac{c^x}{x!} \quad \text{as } n \to \infty, \quad x \in \mathbb{Z}_+, \tag{5.14}$$

where $\mathbb{Z}_+$ are the nonnegative integers.

5.6 Comment

The above results can be much sharpened and extended; see Barbour, Holst, and Janson (1992). We just mention here Le Cam's theorem: with X as in Section 5.2,

$$\| \mathbf{P}(X \in \cdot) - \operatorname{Poi}(p_1 + \cdots + p_n) \| \leqslant 2 \max_{1 \leqslant i \leqslant n} p_i,$$

and in particular,

$$\| \operatorname{Bin}(n,p) - \operatorname{Poi}(np) \| \leqslant 2p.$$

6 Convergence of Discrete Random Variables

Let $X_1, \ldots, X_\infty$ be discrete random variables taking values in a finite or countable set E. We shall first show that convergence in total variation, like (5.13), is (somewhat surprisingly) equivalent to the apparently weaker pointwise convergence of probability mass functions, like (5.14). We shall then show that these distributional modes of convergence can be turned by coupling into a convergence where the random variables actually hit the limit and stay there.

6.1 Mass Function Convergence $\Leftrightarrow$ Total Variation Convergence

Suppose

$$\mathbf{P}(X_n = x) \to \mathbf{P}(X_\infty = x) \quad \text{as } n \to \infty \quad \text{for each } x \in E. \tag{6.1}$$

Then $(\mathbf{P}(X_\infty = x) - \mathbf{P}(X_n = x))^+ \to 0$, and since

$$(\mathbf{P}(X_\infty = x) - \mathbf{P}(X_n = x))^+ \leqslant \mathbf{P}(X_\infty = x)$$

and

$$\sum_{x \in E} \mathbf{P}(X_\infty = x) = 1 < \infty,$$

we have by dominated convergence that

$$\sum_{x \in E} (\mathbf{P}(X_\infty = x) - \mathbf{P}(X_n = x))^+ \to 0, \qquad n \to \infty.$$

Now (5.6) yields convergence in total variation:

$$X_n \overset{tv}{\to} X_\infty, \qquad n \to \infty. \tag{6.2}$$

Conversely, it is clear that (6.2) implies (6.1). Thus (6.1) and (6.2) are equivalent.

6.2 Hitting the Limit

We now show that if (6.1) holds, then there exists a coupling $(\hat{X}_1, \ldots, \hat{X}_\infty)$ of $X_1, \ldots, X_\infty$ and a finite random integer K such that

$$\hat{X}_n = \hat{X}_\infty, \qquad n \geqslant K. \tag{6.3}$$

We obtain this by elaborating on the maximal coupling construction in Section 4.2. Note that (6.1) implies, for all $x \in E$,

$$q_n(x) := \inf_{n \leqslant k < \infty} \mathbf{P}(X_k = x) \uparrow \mathbf{P}(X_\infty = x) \quad \text{as } n \to \infty. \tag{6.4}$$

Put $q_0 \equiv 0$ and let $K, V_1, V_2, \ldots, W_1, W_2, \ldots$ be independent random variables such that for $1 \leqslant n < \infty$ and $x \in E$

$$\mathbf{P}(K = n) = \sum_{x \in E} q_n(x) - \sum_{x \in E} q_{n-1}(x),$$

$$\mathbf{P}(V_n = x) = \begin{cases} (q_n(x) - q_{n-1}(x))/\mathbf{P}(K = n) & \text{if } \mathbf{P}(K = n) > 0, \\ \text{arbitrary} & \text{if } \mathbf{P}(K = n) = 0, \end{cases}$$

$$\mathbf{P}(W_n = x) = \begin{cases} (\mathbf{P}(X_n = x) - q_n(x))/\mathbf{P}(K > n) & \text{if } \mathbf{P}(K > n) > 0, \\ \text{arbitrary} & \text{if } \mathbf{P}(K > n) = 0. \end{cases}$$

The random variable K is finite, since by dominated convergence and (6.4)

$$\mathbf{P}(K \leqslant n) = \sum_{x \in E} q_n(x) \uparrow \sum_{x \in E} \mathbf{P}(X_\infty = x) = 1 \text{ as } n \to \infty.$$

Define, for $1 \leqslant n \leqslant \infty$,

$$\hat{X}_n = \begin{cases} V_K & \text{if } n \geqslant K, \\ W_n & \text{if } n < K. \end{cases} \tag{6.5}$$

This is a coupling of $X_1, \ldots, X_\infty$, since for $1 \leqslant n < \infty$ and each $x \in E$

$$\begin{aligned} \mathbf{P}(\hat{X}_n = x) &= \sum_{1 \leqslant k \leqslant n} \mathbf{P}(V_k = x)\mathbf{P}(K = k) + \mathbf{P}(W_n = x)\mathbf{P}(K > n) \\ &= \sum_{1 \leqslant k \leqslant n} (q_k(x) - q_{k-1}(x)) + (\mathbf{P}(X_n = x) - q_n(x)) \\ &= \mathbf{P}(X_n = x), \end{aligned}$$

while $\hat{X}_\infty = V_K$, and thus [due to (6.4)] for each $x \in E$

$$\mathbf{P}(\hat{X}_\infty = x) = \sum_{1 \leqslant k < \infty} (q_k(x) - q_{k-1}(x)) = \mathbf{P}(X_\infty = x).$$

Clearly, (6.3) holds.

6.3 Converse

Conversely, suppose (6.3) holds. Then $\{K \leqslant n\}$ is a coupling event of the coupling $(\hat{X}_n, \hat{X}_\infty)$ of X_n and X_∞. Applying the coupling event inequality (5.11) and the finiteness of K yields

$$\|\mathbf{P}(X_n \in \cdot) - \mathbf{P}(X_\infty \in \cdot)\| \leqslant 2\mathbf{P}(K > n) \to 0, \qquad n \to \infty,$$

which implies (6.1).

Since (6.1) in turn implies (6.4), which was in fact the condition under which we established (6.3), we have established the following equivalences.

Theorem 6.1. *Let $X_1, \ldots, X_\infty$ be discrete random variables taking values in a finite or countable set E. Then the three claims*

$$\lim_{n \to \infty} \mathbf{P}(X_n = x) = \mathbf{P}(X_\infty = x), \quad x \in E, \qquad [\textit{this is } (6.1)]$$

$$X_n \xrightarrow{tv} X_\infty, \quad n \to \infty, \qquad [\textit{this is } (6.2)]$$

$$\liminf_{n \to \infty} \mathbf{P}(X_n = x) = \mathbf{P}(X_\infty = x), \quad x \in E, \qquad [\textit{this is } (6.4)]$$

are equivalent and hold if and only if there exists a coupling $(\hat{X}_1, \ldots, \hat{X}_\infty)$ of $X_1, \ldots, X_\infty$ and a finite random integer K such that

$$\hat{X}_n = \hat{X}_\infty, \quad n \geqslant K, \qquad [\textit{this is } (6.3)].$$

7 Continuous Variables – Hitting the Limit

Let $X_1, \ldots, X_\infty$ be continuous random variables with densities $f_1, \ldots, f_\infty$. How should Theorem 6.1 be extended to this case? This section is structured as the previous one and gives the answer at the end.

7.1 Density Convergence ⇒ Total Variation Convergence

Replacing probability mass functions by densities in the argument in Section 6.1 yields that the following analogue of (6.1):

$$\begin{aligned} &\text{the densities } f_1, \ldots, f_\infty \text{ can be chosen so that} \\ &f_n(x) \to f_\infty(x) \quad \text{as } n \to \infty \quad \text{for each } x \in \mathbb{R}, \end{aligned} \tag{7.1}$$

implies convergence in total variation,

$$X_n \xrightarrow{tv} X_\infty, \quad n \to \infty. \tag{7.2}$$

However, the converse is not as obvious. In fact, it is no longer true, as we shall see in a while.

7.2 Hitting the Limit

We shall now show that the condition (7.1) is sufficient to hit the limit, that is, if (7.1) holds, then there exists a coupling $(\hat{X}_1, \ldots, \hat{X}_\infty)$ of $X_1, \ldots, X_\infty$ and a finite random integer K such that

$$\hat{X}_n = \hat{X}_\infty, \quad n \geqslant K. \tag{7.3}$$

This follows by a coupling construction analogous to the one in Section 6.2. Let us go through the essential part of it again. Put

$$g_0 \equiv 0 \quad \text{and for } n \geqslant 1 \quad g_n \equiv \inf_{n \leqslant k < \infty} f_k.$$

Let $K, V_1, V_2, \ldots, W_1, W_2, \ldots$ be independent random variables such that for $1 \leqslant n < \infty$,

$$K \text{ is integer valued and } \mathbf{P}(K = n) = \int g_n - \int g_{n-1},$$

$$V_n \text{ has } \begin{cases} \text{density } (g_n - g_{n-1})/\mathbf{P}(K = n) & \text{if } \mathbf{P}(K = n) > 0, \\ \text{arbitrary density} & \text{if } \mathbf{P}(K = n) = 0, \end{cases}$$

$$W_n \text{ has } \begin{cases} \text{density } (f_n - g_n)/\mathbf{P}(K > n) & \text{if } \mathbf{P}(K > n) > 0, \\ \text{arbitrary density} & \text{if } \mathbf{P}(K > n) = 0. \end{cases}$$

Defining $\hat{X}_n$ as at (6.5) yields the desired result, since (7.1) implies that $g_n \uparrow f_\infty$ as $n \to \infty$.

7.3 Converse?

Note that (7.3) was actually established under $g_n \uparrow f_\infty$, that is, under

$$\liminf_{n\to\infty} f_n \quad \text{is a density of } X_\infty, \tag{7.4}$$

which is weaker than (7.1). We shall now show that (7.4) is the correct condition, that is, that (7.4) is implied by (7.3).

Suppose there is a coupling and a finite K such that (7.3) holds. Then $\{K \leqslant n\}$ is a coupling event of the coupling $(\hat{X}_n, \ldots, \hat{X}_\infty)$ of $X_n, \ldots, X_\infty$, and (4.5) yields for (Borel) sets A,

$$\mathbf{P}(\hat{X}_\infty \in A, K \leqslant n) \leqslant \int_A g_n, \quad 1 \leqslant n < \infty.$$

Now, g_n increases to $\liminf_{n\to\infty} f_n$, and thus [by monotone convergence and since $K < \infty$]

$$\mathbf{P}(\hat{X}_\infty \in A) \leqslant \int_A \liminf_{n\to\infty} f_n. \tag{7.5}$$

Also, since $\int g_n \leqslant \int f_n = 1$, we have $\int \liminf_{n\to\infty} f_n \leqslant 1$. Thus, for each (Borel) set A,

$$\begin{aligned} 1 &= \mathbf{P}(\hat{X}_\infty \in A) + \mathbf{P}(\hat{X}_\infty \in A^c) \\ &\leqslant \int_A \liminf_{n\to\infty} f_n + \int_{A^c} \liminf_{n\to\infty} f_n \leqslant 1. \end{aligned}$$

This cannot hold unless (7.5) holds with identity. Thus (7.4) holds.

7.4 Pointwise Convergence of Densities Is Too Strong

We have established the equivalence of (7.3) and (7.4). For discrete variables there were two more equivalences, which both break down in the continuous case. We start with (7.1) and (7.4): certainly, (7.1) implies (7.4), and the following example shows that (7.1) is, in fact, strictly stronger than (7.4).

EXAMPLE 7.1. Let the random variables $X_1, \ldots, X_\infty$ be $[0,1)$ valued and have densities $f_1, \ldots, f_\infty$ defined on $[0,1)$ as follows:

$$f_1(x) = f_\infty(x) = 1, \qquad x \in [0,1);$$

and for $n = 2^m + k$ where $m \geqslant 1$ and $0 \leqslant k < 2^m$ [each $n > 1$ can be written uniquely in this way] put (see Figure 7.1 on the next page)

$$f_n(x) = \begin{cases} 2, & x \in [k2^{-m}, (k+1)2^{-m}), \\ 2 - (1 - 2^{-m})^{-1}, & x \notin [k2^{-m}, (k+1)2^{-m}). \end{cases}$$

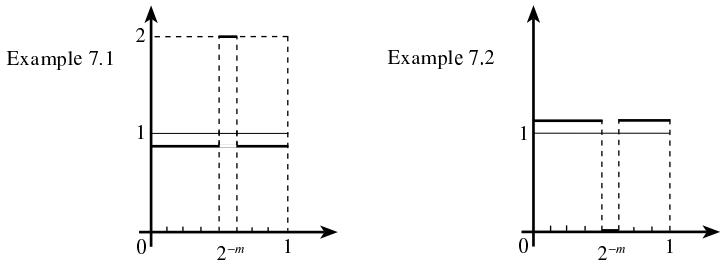

FIGURE 7.1. The functions f_n when $n = 12$ ($m = 3$ and $k = 4$).

Then for each $x \in [0,1)$ there are infinitely many n such that $f_n(x) = 2$, and thus

$$\limsup_{n\to\infty} f_n(x) = 2 \neq 1 = f_\infty(x), \quad x \in [0,1).$$

Hence (7.1) does not hold.

On the other hand, for each $x \in [0,1)$,

$$2 - (1 - 2^{-m})^{-1} \leqslant g_n(x) < 1.$$

This yields, as $n \to \infty$,

$$g_n(x) \to 1 = f_\infty(x), \quad x \in [0,1),$$

and thus (7.4) holds.

7.5 Total Variation Convergence Is Too Weak

Finally, consider (7.4) and (7.2). In Section 7.2 we showed that (7.4) implies (7.3) and in Section 7.3 that (7.3) implies (7.2). Thus (7.4) implies (7.2), and the following example shows that (7.4) is, in fact, strictly stronger than (7.2).

EXAMPLE 7.2. Let the random variables $X_1, \ldots, X_\infty$ be $[0,1)$ valued and have densities $f_1, \ldots, f_\infty$ defined on $[0,1)$ as follows:

$$f_\infty(x) = 1, \quad x \in [0,1);$$

and for $n = 2^m + k$ where $m \geqslant 0$ and $0 \leqslant k < 2^m$ put (see Figure 7.1)

$$f_n(x) = \begin{cases} 0, & x \in [k2^{-m}, (k+1)2^{-m}), \\ (1 - 2^{-m})^{-1}, & x \notin [k2^{-m}, (k+1)2^{-m}). \end{cases}$$

Then, due to (5.6),

$$\begin{aligned} \|\mathbf{P}(X_n \in \cdot) - \mathbf{P}(X_\infty \in \cdot)\| &= 2 \int (f_\infty - f_n)^+ \\ &= 2 \cdot 2^{-m} \to 0, \quad n \to \infty. \end{aligned}$$

Hence (7.2) holds. On the other hand, for each $x \in [0,1)$ there are infinitely many n such that $f_n(x) = 0$, which yields

$$\liminf_{n\to\infty} f_n(x) = 0, \quad x \in [0,1),$$

and thus (7.4) does not hold.

7.6 What Has Been Achieved?

In Sections 7.2–7.5 we have established the following result.

Theorem 7.1. *If $X_1, \ldots, X_\infty$ are continuous random variables with densities $f_1, \ldots, f_\infty$, then*

$$\lim_{n\to\infty} f_n \text{ is a density of } X_\infty \qquad [\textit{this is } (7.1)]$$

is strictly stronger than

$$\liminf_{n\to\infty} f_n \text{ is a density of } X_\infty, \qquad [\textit{this is } (7.4)]$$

which is strictly stronger than

$$X_n \xrightarrow{tv} X_\infty, \quad n \to \infty, \qquad [\textit{this is } (7.2)].$$

Moreover, (7.4) *holds if and only if there exists a coupling $(\hat{X}_1, \ldots, \hat{X}_\infty)$ of $X_1, \ldots, X_\infty$ and a finite random integer K such that*

$$\hat{X}_n = \hat{X}_\infty, \quad n \geqslant K, \qquad [\textit{this is } (7.3)].$$

In Chapter 3 (Section 9) we shall extend this coupling result to general random elements.

8 Convergence in Distribution and Pointwise

Let $X_1, \ldots, X_\infty$ be random variables with distribution functions $F_1, \ldots, F_\infty$. The X_n tend *pointwise* (or *surely*, or *realizationwise*) to X_∞ if

$$X_n \to X_\infty, \quad n \to \infty, \tag{8.1}$$

which is short for

$$X_n(\omega) \to X_\infty(\omega), \quad n \to \infty, \quad \text{for all outcomes } \omega.$$

This means that the X_n close in on the limit without necessarily hitting it as in (7.3). In order to compare (8.1) to (7.3) note that (8.1) can be rewritten as follows: for each $\varepsilon > 0$ there is a finite random integer K_ε such that

$$|X_n - X_\infty| \leqslant \varepsilon, \quad n \geqslant K_\varepsilon, \qquad \text{(see Figure 8.1).} \tag{8.2}$$

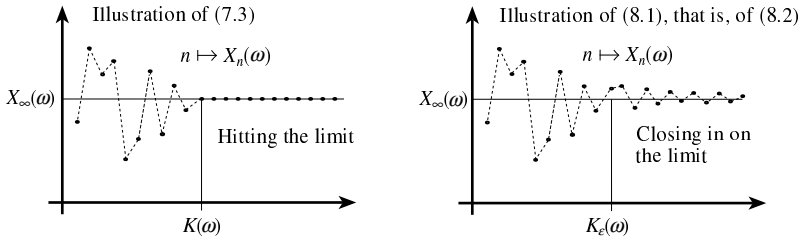

FIGURE 8.1. Comparison of (7.3) and (8.1).

In this section we shall dig out the distributional form of pointwise convergence. The result, once more, is stated at the end of the section.

8.1 Total Variation Convergence Is Too Strong

The distributional condition we are looking for should be implied by pointwise convergence. This excludes convergence in total variation, as can be seen by the following example. Put $X_n = 1/n$ and $X_\infty = 0$. Then certainly (8.1) holds, but X_n does not tend to X_∞ in total variation, since clearly

$$\|\mathbf{P}(X_n \in \cdot) - \mathbf{P}(X_\infty \in \cdot)\| = 2 \not\to 0.$$

Even the much weaker condition

$$F_n(x) \to F_\infty(x), \quad n \to \infty, \quad x \in \mathbb{R}, \tag{8.3}$$

is too strong, since in our example

$$F_n(0) = 0 \not\to 1 = F_\infty(0).$$

8.2 Pointwise Convergence $\Rightarrow$ Convergence in Distribution

The distributional form of pointwise convergence turns out to be the following slight weakening of (8.3):

$$F_n(x) \to F_\infty(x), \quad n \to \infty, \quad \text{for all } x \text{ where } F_\infty \text{ is continuous.} \tag{8.4}$$

This is called *convergence in distribution* and is denoted by

$$X_n \xrightarrow{D} X_\infty, \quad n \to \infty.$$

In order to see that pointwise convergence implies convergence in distribution assume that (8.1) holds and apply its equivalent form (8.2) to obtain that for all $x \in \mathbb{R}$ and $\varepsilon > 0$

$$\begin{aligned} F_n(x) &= \mathbf{P}(X_n \leqslant x, K_\varepsilon \leqslant n) + \mathbf{P}(X_n \leqslant x, K_\varepsilon > n) \\ &\leqslant \mathbf{P}(X_\infty \leqslant x + \varepsilon) + \mathbf{P}(K_\varepsilon > n) \\ &\to F_\infty(x + \varepsilon), \quad n \to \infty, \end{aligned}$$

and

$$\begin{aligned} F_n(x) &\geqslant \mathbf{P}(X_n \leqslant x, K_\varepsilon \leqslant n) \\ &\geqslant \mathbf{P}(X_\infty \leqslant x - \varepsilon, K_\varepsilon \leqslant n) \\ &\to F_\infty(x - \varepsilon), \quad n \to \infty. \end{aligned}$$

Thus for all $x \in \mathbb{R}$ and $\varepsilon > 0$

$$F_\infty(x - \varepsilon) \leqslant \liminf_{n\to\infty} F_n(x) \leqslant \limsup_{n\to\infty} F_n(x) \leqslant F_\infty(x + \varepsilon),$$

and sending ε to 0 shows that (8.4) holds.

8.3 Turning Distributional Convergence into Pointwise

We shall now use the quantile coupling (Section 3.1) to reverse the above implication, that is, turn convergence in distribution into pointwise convergence (see Figure 8.2).

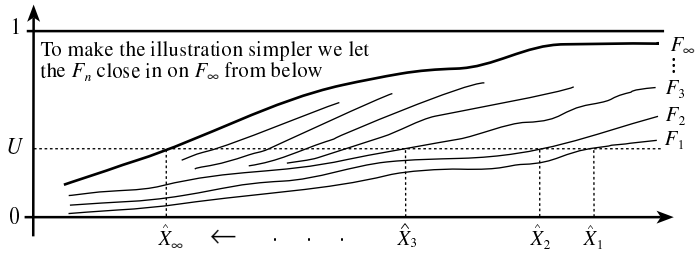

FIGURE 8.2. Turning convergence in distribution into pointwise convergence.

First we need the following fact.

Lemma 8.1. *For a nondecreasing real function f, the set of points where f is not continuous is either finite or countable.*

PROOF. That f is not continuous at u is equivalent to the left-hand limit being less than the right hand limit, $f(u-) < f(u+)$. To each such u we can associate a rational number that lies in the interval $[f(u-), f(u+))$. These intervals are disjoint (because f is nondecreasing), and thus we have established a one-to-one correspondence between $\{u \in \mathbb{R} : f(u-) < f(u+)\}$ and a subset of the rational numbers. Since the rationals are countable, the set $\{u \in \mathbb{R} : f(u-) < f(u+)\}$ is either finite or countable. □

Recall that the generalized inverse of a distribution function F is

$$F^{-1}(u) = \inf\{x \in \mathbb{R} : F(x) \geqslant u\}, \quad u \in [0,1].$$

Clearly, F^{-1} is nondecreasing,

$$F(F^{-1}(u)-) \leqslant u \leqslant F(F^{-1}(u)), \quad u \in [0,1], \tag{8.5}$$

and

$$\left.\begin{array}{l} F^{-1} \text{ is continuous at } u \\ \text{and } F(x-) \leqslant u \leqslant F(x) \end{array}\right\} \Rightarrow F^{-1}(u) = x. \tag{8.6}$$

Since F_∞^{-1} is nondecreasing, the set of points where F_∞^{-1} is not continuous is finite or countable. Thus there is a random variable U that is uniform on $[0,1]$ and takes values in the set of points at which F_∞^{-1} is continuous. Use this U to define the quantile coupling

$$\hat{X}_n = F_n^{-1}(U), \quad 0 \leqslant n \leqslant \infty.$$

We shall show that (8.4) implies

$$F_n^{-1}(u) \to F_\infty^{-1}(u), \ n \to \infty, \ \text{for all } u \text{ where } F_\infty^{-1} \text{ is continuous}, \tag{8.7}$$

which yields the desired result that $\hat{X}_n \to \hat{X}_\infty$ as $n \to \infty$.

8.4 Establishing That (8.4) Implies (8.7)

Fix $u \in [0,1]$ and put

$$x = \liminf_{n \to \infty} F_n^{-1}(u).$$

Since F_∞ is nondecreasing and thus is discontinuous at only finitely or countably many points, we can fix an arbitrarily small $\varepsilon > 0$ such that F_∞ is continuous at both $x - \varepsilon$ and $x + \varepsilon$. Let n_k, $k \geqslant 1$, be a sequence of integers such that

$$x - \varepsilon < F_{n_k}^{-1}(u) \leqslant x + \varepsilon, \quad k \geqslant 1.$$

Applying (8.5) yields

$$F_{n_k}(x-\varepsilon) \leqslant u \leqslant F_{n_k}(x+\varepsilon), \quad k \geqslant 1.$$

Send k to infinity and use (8.4) and the choice of ε to deduce

$$F_\infty(x-\varepsilon) \leqslant u \leqslant F_\infty(x+\varepsilon).$$

Then send ε to zero to obtain

$$F_\infty(x-) \leqslant u \leqslant F_\infty(x). \tag{8.8}$$

Replace x by

$$y = \limsup_{n\to\infty} F_n^{-1}(u)$$

in the above argument to obtain

$$F_\infty(y-) \leqslant u \leqslant F_\infty(y). \tag{8.9}$$

If F_∞^{-1} is continuous at u, we obtain from (8.6), (8.8), and (8.9) that

$$F_\infty^{-1}(u) = x = \liminf_{n\to\infty} F_n^{-1}(u),$$

$$F_\infty^{-1}(u) = y = \limsup_{n\to\infty} F_n^{-1}(u).$$

Thus (8.7) holds.

8.5 What Has Been Achieved?

In Sections 8.2–8.4 we have established the following result.

Theorem 8.1. *Let $X_1, \ldots, X_\infty$ be random variables. Then*

$$X_n \xrightarrow{D} X_\infty, \quad n \to \infty,$$

if and only if there exists a coupling $(\hat{X}_1, \ldots, \hat{X}_\infty)$ of $X_1, \ldots, X_\infty$ such that

$$\hat{X}_n \to \hat{X}_\infty, \quad n \to \infty.$$

We finally mention that the definition (8.4) of convergence in distribution can be seen to be equivalent to

$$\mathbf{E}[f(X_n)] \to \mathbf{E}[f(X_\infty)], \quad n \to \infty,$$

for all bounded continuous functions f. This is taken to be the definition of convergence in distribution for random elements in metric spaces. In Chapter 3 (Section 10) we extend Theorem 8.1 to random elements in a separable metric space.

9 Quantile Coupling – Dominated Convergence

The pointwise version of the *dominated convergence* theorem [see Ash (1972)] states that if $X_1, X_2, \ldots, X_\infty, X$ are random variables such that

$$|X_n| \leqslant X, \quad 1 \leqslant n < \infty,$$
$$\mathbf{E}[X] < \infty \quad \text{and} \quad X_n \to X_\infty \text{ as } n \to \infty,$$

then

$$\mathbf{E}[|X_\infty|] < \infty \quad \text{and} \quad \mathbf{E}[X_n] \to \mathbf{E}[X_\infty] \text{ as } n \to \infty. \tag{9.1}$$

Using the quantile coupling it is straightforward to extend this result to the following distributional form.

Theorem 9.1. *If $X_1, X_2, \ldots, X_\infty, X$ are random variables such that*

$$|X_n| \overset{D}{\leqslant} X, \quad 1 \leqslant n < \infty,$$
$$\mathbf{E}[X] < \infty \quad \textit{and} \quad X_n \overset{D}{\to} X_\infty \textit{ as } n \to \infty,$$

then (9.1) *holds.*

Proof. Under the assumptions of the theorem we have

$$X_n^+ \overset{D}{\leqslant} X \quad \text{and} \quad X_n^+ \overset{D}{\to} X_\infty^+ \quad \text{as } n \to \infty,$$
$$X_n^- \overset{D}{\leqslant} X \quad \text{and} \quad X_n^- \overset{D}{\to} X_\infty^- \quad \text{as } n \to \infty.$$

Apply the quantile coupling in Sections 3 and 8 to turn these distributional relations into pointwise ones, that is, to obtain copies of X_n^+ and X_n^- that are pointwise dominated by a copy of X (Section 3) and converge pointwise to copies of X_∞^+ and X_∞^-, respectively (Section 8). By the pointwise version of dominated convergence, this together with $\mathbf{E}[X] < \infty$ implies

$$\mathbf{E}[X_\infty^+] < \infty \quad \text{and} \quad \mathbf{E}[X_n^+] \to \mathbf{E}[X_\infty^+] \quad \text{as } n \to \infty,$$
$$\mathbf{E}[X_\infty^-] < \infty \quad \text{and} \quad \mathbf{E}[X_n^-] \to \mathbf{E}[X_\infty^-] \quad \text{as } n \to \infty.$$

Thus $\mathbf{E}[|X_\infty|] = \mathbf{E}[X_\infty^+] + \mathbf{E}[X_\infty^-] < \infty$ and

$$\begin{aligned}\mathbf{E}[X_n] &= \mathbf{E}[X_n^+] - \mathbf{E}[X_n^-] \\ &\to \mathbf{E}[X_\infty^+] - \mathbf{E}[X_\infty^-] = \mathbf{E}[X_\infty] \text{ as } n \to \infty,\end{aligned}$$

and the proof is complete. □

In Chapter 2 we shall need the following extension to a continuous index.

Corollary 9.1. *If X_t, $t \in [0,\infty]$, and X are random variables such that*

$$|X_t| \overset{D}{\leqslant} X, \quad t \in [0,\infty),$$

$$\mathbf{E}[X] < \infty \quad \textit{and} \quad X_t \overset{D}{\to} X_\infty \textit{ as } t \to \infty,$$

then

$$\mathbf{E}[|X_\infty|] < \infty \quad \textit{and} \quad \mathbf{E}[X_t] \to \mathbf{E}[X_\infty] \textit{ as } t \to \infty.$$

PROOF. A collection of real numbers like $\mathbf{E}[X_t]$, $t \in [0,\infty)$, tends to a limit if and only if it tends to this limit along all subsequences $\mathbf{E}[X_{t(n)}]$, $n \geqslant 1$, where the $t(n)$ increase to ∞ as $n \to \infty$. Apply Theorem 9.1 to $X_{t(n)}$ to obtain $\mathbf{E}[X_{t(n)}] \to \mathbf{E}[X_\infty]$ as $n \to \infty$. □

10 Impossible Coupling – Quantum Physics

We end this first chapter on a rather different note: the coupling aspect of a (the?) problem in quantum physics.

10.1 A Surprising Experimental Result

The following experiment has been carried out. Some material (calcium, carefully excited by laser) sends off particles (photons) in pairs, one particle to the left and the other to the right. Measuring devices are placed on each side of this material with measurements made when particles pass through. What is being measured is the so-called polarization of the particle, which can be either 1 or -1 and depends on the angle in the plane orthogonal to the direction of movement.

0. When the measuring devices are aligned to measure polarization in the same direction, say 0°, the same measurement is always recorded on both sides.

1. When the left device is tilted 30° and the right device is kept at the initial 0° position, then the measurements agree $\frac{3}{4}$ of the time.

2. When the left device is rotated back to its 0° position and the right device is tilted $-30°$ (that is, 30° in the opposite direction), then the measurements also agree $\frac{3}{4}$ of the time.

3. When the left device is again tilted 30° and the right device is kept at its new $-30°$ position (that is, the total relative rotation is 60°), then the measurements agree $\frac{1}{4}$ of the time.

10.2 Why Surprising?

On the basis of the above empirical facts it is now natural to build the following model. Consider a particular pair of photons, and set

X = the polarization of the left particle in the 0° direction

= the polarization of the right particle in the 0° direction,

Y = the polarization of the left particle in the 30° direction,

Z = the polarization of the right particle in the $-30°$ direction;

see Figure 10.1.

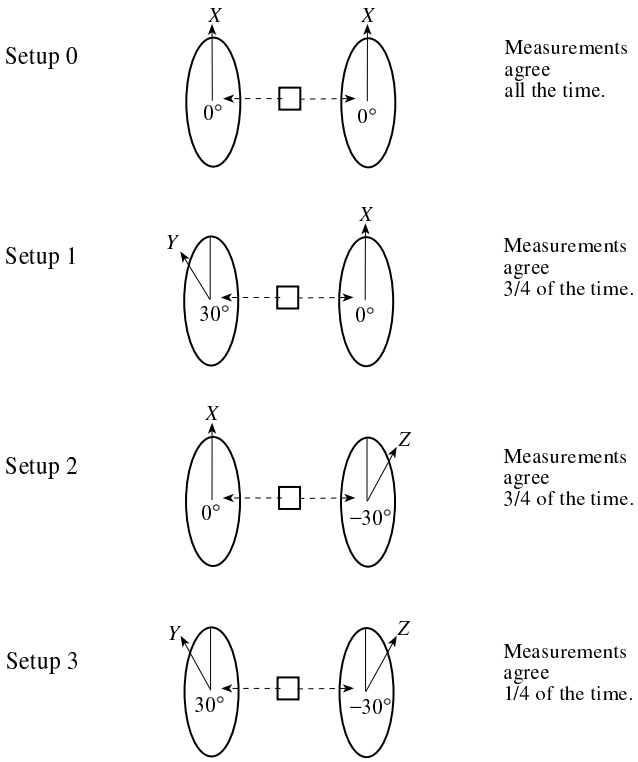

FIGURE 10.1. The experimental setups.

Interpreting the relative frequencies as probabilities, we have

$$\mathbf{P}(Y = X) = \mathbf{P}(X = Z) = \tfrac{3}{4}, \tag{10.2}$$

and

$$\mathbf{P}(Y = Z) = \tfrac{1}{4}. \tag{10.3}$$

By basic rules of probability,

$$\begin{aligned}\mathbf{P}(Y = Z) &\geqslant \mathbf{P}(Y = Z, X = Z)\\ &= \mathbf{P}(Y = X, X = Z)\\ &= \mathbf{P}(Y = X) - \mathbf{P}(Y = X, X \neq Z)\\ &\geqslant \mathbf{P}(Y = X) - \mathbf{P}(X \neq Z)\\ &= \mathbf{P}(Y = X) + \mathbf{P}(X = Z) - 1,\end{aligned}$$

that is,

$$\mathbf{P}(Y = Z) \geqslant \mathbf{P}(Y = X) + \mathbf{P}(X = Z) - 1. \tag{10.4}$$

Combine this, (10.2), and (10.3) to obtain the following contradiction:

$$\tfrac{1}{4} \geqslant \tfrac{1}{2}.$$

This contradiction is derived in an ordinary probabilistic way from straightforward empirical facts: real life *seems* to contradict probability theory...

10.3 Predicted by Quantum Theory

Because of this apparent contradiction it is all the more annoying (for probabilists) that the empirical results are in fact predicted by quantum mechanics, which calculates the probabilities as follows:

$$\mathbf{P}(Y = X) = \mathbf{P}(X = Z) = \cos^2 30^\circ = 1 - \sin^2 30^\circ = 1 - (\tfrac{1}{2})^2 = \tfrac{3}{4},$$
$$\mathbf{P}(Y = Z) = \cos^2 60^\circ = (\tfrac{1}{2})^2 = \tfrac{1}{4}.$$

10.4 No Contradiction at the Level of Observation

Note that X, Y, and Z refer to polarization as intrinsic properties of the particles, thought of as existing simultaneously without interaction with the macro world (without being measured). If we instead stay at the level of observation (measurement), then it turns out that the contradiction disappears.

It is clear that we are dealing with three experimental setups (leaving out the one with the measuring devices aligned).

First consider Setup 1: the case when the left device is tilted 30° and the right device is kept at the initial 0° position. Put

X_1 = *observed* polarization of the right particle in the 0° direction,

Y_1 = *observed* polarization of the left particle in the 30° direction.

In addition to the measurements agreeing $\frac{3}{4}$ of the time it has been recorded that -1 and 1 are observed in equal proportions on both sides. Specify the complete joint distribution of X_1 and Y_1 as

$$\mathbf{P}(X_1 = -1, Y_1 = -1) = \mathbf{P}(X_1 = 1, Y_1 = 1) = \tfrac{3}{8},$$

$$\mathbf{P}(X_1 = -1, Y_1 = 1) = \mathbf{P}(X_1 = 1, Y_1 = -1) = \tfrac{1}{8}.$$

This is in accordance with the relative frequencies since

$$\begin{aligned}\mathbf{P}(Y_1 = -1) &= \mathbf{P}(X_1 = -1) \\ &= \mathbf{P}(X_1 = -1, Y_1 = -1) + \mathbf{P}(X_1 = -1, Y_1 = 1) \\ &= \tfrac{1}{2},\end{aligned}$$

$$\begin{aligned}\mathbf{P}(X_1 = Y_1) &= \mathbf{P}(X_1 = -1, Y_1 = -1) + \mathbf{P}(X_1 = 1, Y_1 = 1) \\ &= \tfrac{3}{4}.\end{aligned}$$

Now consider Setup 2: the case when the left device is at the $0°$ position and the right device is tilted $-30°$. Put

$X_2 =$ *observed* polarization of the left particle in the $0°$ direction,

$Z_2 =$ *observed* polarization of the right particle in the $-30°$ direction.

Letting (X_2, Z_2) have the same distribution as (X_1, Y_1) again yields probabilities in accordance with the relative frequencies,

$$\mathbf{P}(Y_2 = -1) = \mathbf{P}(X_2 = -1) = \tfrac{1}{2} \quad \text{and} \quad \mathbf{P}(X_2 = Y_2) = \tfrac{3}{4}.$$

Finally, consider Setup 3: the case when the left device is tilted $30°$ and the right device $-30°$. Put

$Y_3 =$ *observed* polarization of the left particle in the $30°$ direction,

$Z_3 =$ *observed* polarization of the right particle in the $-30°$ direction.

The measurements now agree only $\frac{1}{4}$ of the time, but it has still been recorded that -1 and 1 are observed in equal proportions on both sides. Specify the complete joint distribution of Y_3 and Z_3 as

$$\mathbf{P}(Y_3 = -1, Z_3 = -1) = \mathbf{P}(Y_3 = 1, Z_3 = 1) = \tfrac{1}{8},$$

$$\mathbf{P}(Y_3 = -1, Z_3 = 1) = \mathbf{P}(Y_3 = 1, Z_3 = -1) = \tfrac{3}{8}.$$

This is in accordance with the relative frequencies,

$$\mathbf{P}(Y_3 = -1) = \mathbf{P}(Z_3 = -1) = \tfrac{1}{2} \quad \text{and} \quad \mathbf{P}(Y_3 = Z_3) = \tfrac{1}{4}.$$

We have managed to account for all three experiments, and thus the contradiction is not at the level of observation. The contradiction appears when we assume that each particle has a polarization in a direction where we do not make a measurement.

10.5 What Has This to Do with Coupling?

We have created three pairs (X_1, Y_1), (X_2, Z_2), and (Y_3, Z_3). What we proved in Section 10.2 is that there is no coupling of these pairs such that the X-variables agree, the Y-variables agree, and the Z-variables agree. More precisely, there is no jointly distributed triple (X, Y, Z) such that

$$(X, Y) \stackrel{D}{=} (X_1, Y_1), \quad (X, Z) \stackrel{D}{=} (X_2, Z_2), \quad (Y, Z) \stackrel{D}{=} (Y_3, Z_3).$$

That is, although reality *seems* to be able to construct a coupling, we can't.

10.6 Does Probability Not Suffice in the Micro World?

It is one of the implications of quantum theory that polarization cannot be measured simultaneously in all three directions; only one measurement on each particle is possible. The reason we have measurements in pairs is that we have two particles. The above contradiction further suggests that polarization exists in the micro world only through interaction with the macro world (only by being measured).

Is there then nothing, no reality, behind the observations? Or does probability not suffice to describe it?

One school of thought claims that *classical* probability (that is, Kolmogorov's axioms) is too narrow. It should be replaced by *quantum* probability (an axiom system more general than Kolmogorov's) in a similar way as Newton's theory had to be replaced by Einstein's in physics. Applying quantum probability there is no longer a contradiction to be derived from the assumption that polarization exists in all three directions. See Kümmerer and Maassen (1998) and Accardi (1998) for such viewpoints.

Note that there are finitely many possible outcomes in each individual experiment, so the contradiction does not appear to have to do with countable additivity. Since Kolmogorov's axioms otherwise reflect properties of relative frequencies, it is hard to swallow that they should not apply. And so it is not surprising that there are other attempts to get rid of the contradiction. See Maudlin (1994) and Gill (1998, 1999) for the following point of view.

Behind the attempt in Section 10.2 to create a model are several implicit assumptions. One assumption is that *measuring* the polarization in a particular direction does not affect the polarization in the other directions. In other words, an interplay between the micro and macro worlds is not allowed. Allowing a *local* interplay is not a serious crime against physical ideas, but it turns out that a *nonlocal* interplay is needed to get rid of the contradiction. Nonlocal means that the experimental setup on the left, for instance, affects the polarization of a particle measured on the right. This is not easy to accept, but for an Einsteinian realist this is easier to accept than having to discard Kolmogorov's axioms, which is too close to discarding $2 + 2 = 4$.

10.7 What Does This Teach Us About Coupling?

The above excursion into the quantum experience shows that we have to be careful when assuming existence of couplings. For empirical or intuitive reasons joint distributions may appear to exist when they do not. In Chapter 3 (Sections 3 through 5) we shall consider some safe methods for constructing couplings. The next chapter, however, is devoted to the classical triumphs of the coupling method.