Chapter 2

The representation theorem

Let \((X, d)\) be a metric space, equipped with a \(\sigma\)-algebra \(\mathcal{A}\). Let \(X, X_1, X_2, \ldots\) be \(X\)-valued stochastic variables, all \(\mathcal{A}\)-measurable. If \(X_n \overset{\text{a.e.}}{\to} X\) it is easy to see that \(X_n \overset{D}{\to} X\). If \(f : X \to \mathbb{R}\) is bounded, \(\mathcal{A}\)-measurable and continuous, it holds that \(f(X_n) \overset{\text{a.e.}}{\to} f(X)\) bounded by the constant \(\|f\|\) (the uniform norm) and thus
\[
\int f(X_n) \, dP \to \int f(X) \, dP.
\]
from the dominated convergence theorem.

Clearly, the opposite result is not true in general. Convergence in distribution is a statement about the sequence of marginal distributions of the variables involved, while almost sure convergence is a statement about the joint distribution. To make an extreme case: if \(X, X_1, X_2, \ldots\) all have the same marginal distribution, then it is trivially true \(X_n \overset{D}{\to} X\) irrespective of the joint distribution. But is is easy to construct joint distributions where the variables do not converge almost surely - we may for instance consider the case where \(X_1, X_2, \ldots\) are independent.

However, as we shall see, if \(X_n \overset{D}{\to} X\) it is in most cases possible to replace \(X_1, X_2, \ldots\) with another sequence \(X'_1, X'_2, \ldots\) of \(X\)-valued stochastic variables such that \(X'_n\) has the same distribution as \(X_n\) for every \(n\), and such that \(X'_n \overset{\text{a.e.}}{\to} X'\). Shifting attention to the \(X'_n\)-sequence allows us to think in terms om almost sure convergence, and this trivializes many arguments. If needed, the conclusion can usually be translated
back to the original sequence with a minimum of effort. An example of this type of translation is given in example 2.6, after the proof of the representation theorem.

2.1 An existence theorem for independent variables

In the sequel, we will need to invent new stochastic variables, which are independent of the variables already in play. Hence we need a general result guaranteeing that such variables always exist.

**Theorem 2.1** Let \((\mathcal{Y}_n, \mathbb{K}_n, \mu_n)\) be a sequence of probability spaces. There exists a probability space \((\Omega, \mathbb{F}, P)\) and stochastic variables \(Y_n : \Omega \to \mathcal{Y}_n\) for \(n \in \mathbb{N}\) with the property that

\[
Y_1(P) = \mu_1, \quad Y_2(P) = \mu_2, \ldots
\]

and such that \(Y_1, Y_2, \ldots\) are independent.

**Proof:** We will not give a full proof, but merely sketch the construction. Let \(\Omega\) be the sequence space

\[
\Omega = \mathcal{Y}_1 \times \mathcal{Y}_2 \times \ldots
\]

consisting of sequences \((y_1, y_2, \ldots)\) where \(y_1 \in \mathcal{Y}_1, y_2 \in \mathcal{Y}_2\) and so on. Consider the natural projections \(\rho_n : \Omega \to \mathcal{Y}_1 \times \ldots \times \mathcal{Y}_n\) given by

\[
\rho_n(y_1, y_2, \ldots) = (y_1, \ldots, y_n)
\]

Let \(\mathbb{F}_n\) be the \(\sigma\)-algebra on \(\Omega\) generated by \(\rho_n\) (with the obvious convention that \(\mathcal{Y}_1 \times \ldots \times \mathcal{Y}_n\) is equipped with \(\mathbb{K}_1 \otimes \ldots \otimes \mathbb{K}_n\)). It is easy to show that there is a unique probability measure \(P_n\) on \((\Omega, \mathbb{F}_n)\) defined by the requirement that

\[
P_n\left(\rho_n^{-1}(A)\right) = \mu_1 \otimes \ldots \otimes \mu_n(A) \quad \text{for } A \in \mathbb{K}_1 \otimes \ldots \otimes \mathbb{K}_n
\]

It is also easy to see that \(\mathbb{F}_n \subset \mathbb{F}_{n+1}\) for every \(n\) and that

\[
P_{n+1}(B) = P_n(B) \quad \text{for } B \in \mathbb{F}_n
\]

Hence we can knit the \(P_n\)-measures together to a finitely additive set function \(P\) defined on the algebra

\[
\mathbb{F}_\infty = \bigcup_{n=1}^{\infty} \mathbb{F}_n
\]
2.1. An existence theorem for independent variables

using the convention that

\[ P(B) = \mu_n(B) \quad \text{whenever } B \in \mathcal{F}_n \]

As an absolutely non-trivial application of Caratheodory’s extension theorem it can be shown that \( \mu \) can be extended to a proper probability measure on \( \mathcal{F} = \sigma(\mathcal{F}_n) \).

Taken this for granted, we define stochastic variables \( Y_n : \Omega \rightarrow \mathcal{Y}_n \) as the coordinate projections,

\[ Y_n(y_1, y_2, \ldots) = y_n. \]

It is easy to see that

\[ (Y_1, \ldots, Y_n)(P) = \mu_1 \otimes \cdots \otimes \mu_n \]

showing that \( Y_1, \ldots, Y_n \) are independent and have the desired marginal distributions.

\( \square \)

The application of Caratheorody’s extension theorem in the proof of theorem 2.1 is similar in spirit to Kolmogorov’s well-known consistency theorem, dealing with the existence of measures on infinite products of the real line. But note that the problem here is intrinsically different: Kolmogorov’s theorem relies on a topological property of the real line (compactifiability) to show that the continuity condition of Caratheodory’s theorem is satisfied. The present problem has no topological structure whatsoever, as we started out from arbitrary probability spaces, and hence the continuity condition has to be established by a different route. The argument is quite intricate, and the result is usually known as Tulcea’s theorem. The technique can be applied to quite a broad class of existence problems for measures, but only on countable products (as opposed to Kolmogorov’s theorem which easily accommodates uncountable products).

Theorem 2.1 allows to introduce new independent stochastic variables at our hearts desire. But a technical remark is necessary: we will frequently have stochastic variables \( X_1, X_2, \ldots \) already, and we will want to combine the new variables \( Y_1, Y_2, \ldots \) with the old variables, using various pointwise constructions. The old variables are defined on a background space \((\Omega, \mathcal{F}, P)\). The new variables, comming from theorem 2.1, are most likely defined on a different background space \((\Omega', \mathcal{F}', P')\). We solve this problem by switching to \((\Omega \times \Omega', \mathcal{F} \otimes \mathcal{F}', P \otimes P')\) on which it is possible to find both a copy of the \( X_n \)-sequence and a copy of the \( Y_n \)-sequence, independent of each other. This type of modification of the background space is usually done silently, and all stochastic variables are assumed to be defined on the same background space.
Chapter 2. The representation theorem

2.2 Open covers

Let \((X, d)\) be a metric space, and let \(A \subset X\). A collection \((G_i)_{i \in I}\) of open subsets of \(X\) is called an open cover of \(A\) if

\[
A \subset \bigcup_{i \in I} G_i
\]

If \(I_0 \subset I\) and if \((G_i)_{i \in I_0}\) is also a cover of \(A\), we refer to \((G_i)_{i \in I_0}\) as a subcover of the original open cover.

**Theorem 2.2 (Lindelöf)** Let \((X, d)\) be a metric space. A subset \(A \subset X\) is separable if and only if every open cover of \(A\) has a countable subcover.

**Proof:** Suppose that every open cover of \(A\) has a countable subcover. We can for any \(r > 0\) apply this to the open cover of balls \(\{B(x, r) \mid x \in A\}\). In particular, for each \(n \in \mathbb{N}\) we can find a sequence \((x_{nm})_{m \in \mathbb{N}}\) of points in \(A\) such that

\[
A \subset \bigcup_{m=1}^{\infty} B\left(x_{nm}, \frac{1}{n}\right)
\]

It easily follows that \((x_{nm})_{n,m \in \mathbb{N}}\) is a countable dense subset of \(A\).

Conversely, suppose \(A\) is separable with a dense countable subset \(A_0\), and let \((G_i)_{i \in I}\) be an open cover. Consider

\[
J = \left\{ B(x, r) \mid x \in A_0, r \in \mathbb{Q}^+, B(x, r) \subset G_i \text{ for some } i \in I \right\}
\]

Clearly \(J\) is countable. For each ball \(B(x, r) \in J\) we may pick an index \(i \in I\) such that \(B(x, r) \subset G_i\). Let \((G_j)_{j \in J}\) be the collection of open sets chosen this way. We will prove that \(A \subset \bigcup_{j \in J} G_j\), establishing that \((G_j)_{j \in J}\) is a subcover\(^1\).

Take \(x \in A\). As \(A\) is covered by \((G_i)_{i \in I}\), we can find \(i \in I\) such that \(x \in G_i\). As \(G_i\) is open, we can find \(r \in \mathbb{Q}^+\) such that \(B(x, r) \subset G_i\). Using that \(A_0\) is dense in \(A\), we can find

\[
x_0 \in A_0 \cap B\left(x, \frac{r}{2}\right)
\]

\(^1\)Strictly speaking, \(J\) is not a subset of \(I\), but as every \(G_j\)-set is a member of the original family \((G_i)_{i \in I}\), the term ‘subcover’ is still used reasonably.
2.3. The representation theorem

It is easy to see that
\[ x \in B\left(x_0, \frac{r}{2}\right) \subset B(x, r) \subset G_i \]
In particular, \( B(x_0, r/2) \in J \). So there is a \( G_j \) such that
\[ x \in B\left(x_0, \frac{r}{2}\right) \subset G_j \]
and the claim is proved. \( \square \)

We mention in passing that a well-known characterization of a compact subset of \( X \) is that every open cover has a finite subcover.

2.3 The representation theorem

Recall that a finite partition of a space \( X \) is a collection of subsets \( B_1, \ldots, B_n \) of \( X \) such that
\[ B_i \cap B_j = \emptyset \quad \text{for } i \neq j, \quad \bigcup_{i=1}^{n} B_i = X \]

**Lemma 2.3** Let \((X, d)\) be a metric space equipped with a \( \sigma \)-algebra \( \mathcal{A} \). Let \( X \) be an \( X \)-valued stochastic variable, such that \( P(X \in C) = 1 \) for a separable set \( C \in \mathcal{A} \) consisting of EH-regular points. For any \( \varepsilon > 0 \) there is a partition \( B_0, B_1, \ldots, B_k \) of \( \mathcal{A} \)-measurable subsets of \( X \) such that
1) \( \text{diam}(B_i) \leq 2\varepsilon \) for \( i = 1, \ldots, k \)
2) \( P(X \in \partial B_i) = 0 \) for \( i = 0, 1, \ldots, k \)
3) \( P(X \in B_0) < \varepsilon \)

*Without loss of generality it may be assumed that \( P(X \in B_i) > 0 \) for \( i = 1, \ldots, k \).*

**Proof:** Let \( \varepsilon > 0 \) be given. Consider a point \( x \in C \). Since \( x \) is EH-regular, and thus DP-regular, we can find an \( \mathcal{A} \)-measurable, continuous bump function \( f_x \) satisfying
\[ \{x\} \prec f \prec B(x, \varepsilon) \]
Consider the map $\phi : (0, 1) \to \mathbb{R}$ given by
\[
\phi(x) = P(f(X) > s) \quad \text{for} \ s \in (0, 1)
\]
This is decreasing, in particular it can only have countably many points of discontinuity. Take $s \in (0, 1)$ as a a point of continuity of $\phi$, and let
\[
A_x = \{ y \in X \mid f(y) > s \}
\]
Since $f$ is continous, we see that $A_x$ is an open and $\mathcal{A}$-measurable. It is easy to see that
\[
A_x \subset \{ y \in X \mid f_x(y) > 0 \} \subset B(x, \varepsilon)
\]
so in particular $\text{diam}(A_x) \leq 2\varepsilon$. It is also easy to see that
\[
P(X \in \partial A_x) \leq P(f(X) = s) = 0
\]
since $s$ is a point of continuity for $\phi$.

Obviously, $(A_x)_{x \in C}$ is an open cover of the separable set $C$. Hence Lindelöf’s theorem allows us to find countably many points $x_1, x_2, \ldots, \in C$ such that
\[
C \subset \bigcup_{n=1}^{\infty} A_{x_n}
\]
Let
\[
B_n = A_{x_n} \setminus \bigcup_{i=1}^{n-1} A_{x_i}
\]
be the result of the natural disjointification process. As
\[
\partial B_n \subset \bigcup_{i=1}^{n} \partial A_{x_i}
\]
it is clear that $P(X \in \partial B_n) = 0$ for every $n$. And since $B_n$ is constructed as a subset of $A_{x_n}$, it holds that $\text{diam}(B_n) \leq \text{diam}(A_{x_n}) \leq 2\varepsilon$.

As a property of the disjointification process, it holds that $\bigcup_{n=1}^{\infty} B_n = \bigcup_{n=1}^{\infty} A_{x_n}$. In particular $(B_n)_{n \in \mathbb{N}}$ is a disjoint cover of $C$, and thus
\[
\sum_{n=1}^{\infty} P(X \in B_n) = 1
\]
This allows us to find \( k \in \mathbb{N} \) such that

\[
\sum_{i=1}^{k} P(X \in B_i) \geq 1 - \varepsilon
\]

If we let

\[
B_0 = \bigcup_{i=k+1}^{\infty} B_i
\]

we are done. Note that we can always assume that \( P(X \in B_i) > 0 \) for \( i = 1, \ldots, k \). If not, we can simply transfer \( B_i \) to \( B_0 \). We do not know if \( P(X \in B_0) \) is non-zero, though.

\[\square\]

**Lemma 2.4** Let \((X, d)\) be a metric space equipped with a \( \sigma \)-algebra \( \mathcal{A} \). Let \( X \) be an \( X \)-valued stochastic variable, such that \( P(X \in C) = 1 \) for a separable set \( C \in \mathcal{A} \) consisting of EH-regular points. Suppose that for some \( \varepsilon > 0 \) there is given a partition \( B_0, B_1, \ldots, B_k \) of \( \mathcal{A} \)-measurable subsets of \( X \) such that

1) \( \text{diam}(B_i) \leq 2\varepsilon \) for \( i = 1, \ldots, k \)
2) \( P(X \in \partial B_i) = 0 \) for \( i = 0, 1, \ldots, k \)
3) \( P(X \in B_0) < \varepsilon \)

and assume that \( P(X \in B_i) > 0 \) for \( i = 1, \ldots, k \). Let \( Y_1, Y_2, \ldots \) be a sequence of \( X \)-valued stochastic variable with the property that\( P(Y_n \in B_i) \geq (1 - \varepsilon) P(X \in B_i) \) for \( i = 0, 1, \ldots, k \), \( n \in \mathbb{N} \)

Then there exists stochastic variables \( Y'_1, Y'_2, \ldots \) such that \( Y'_n \) has the same distribution as \( Y_n \) for every \( n \in \mathbb{N} \) and such that

\[
P\left( \bigcup_{n=1}^{\infty} (|X - Y'_n| > 2\varepsilon) \right) < 2\varepsilon
\]

(2.1)

**Proof:** Initially, we will work with a single stochastic variable \( Y \) satisfying that \( P(Y \in B_i) \geq (1 - \varepsilon) P(X \in B_i) \) for \( i = 0, 1, \ldots, k \)

From the positivity of \( P(X \in B_i) \) it follows that \( P(Y \in B_i) > 0 \) for \( i = 1, \ldots, k \). Introduce \( X \)-valued stochastic variables \( Z_0, Z_1, \ldots, Z_n \), independent of each other and of \( X \), such that

\[
P(Z_i \in A) = \frac{P(Y \in A \cap B_i)}{P(Y \in B_i)} \quad \text{for} \ A \in \mathcal{A}
\]
This specification of distribution makes sense for \( i = 1, \ldots, n \) and possibly also for \( i = 0 \) - but if \( P(Y \in B_0) = 0 \), we just let \( Z_0 \) have any distribution on \( X \), as the exact choice turns out to be irrelevant.

Furthermore, let \( W \) and \( V \) be two real-valued stochastic variables, independent of \( Z_0, \ldots, Z_k \), of \( X \) and of each other. We assume that \( W \) has values in \( \{0, 1, \ldots, k\} \), where the point probabilities \( P(W = i) \) are yet to be specified. And we assume that \( V \) is a Bernoulli variable with

\[
P(V = 1) = 1 - \varepsilon, \quad P(V = 0) = \varepsilon
\]

We define a new stochastic variable \( Y' \) by choosing one of the \( Z_i \)'s according to the following rule: if \( V = 1 \), we let the position of \( X \) determine the choice of \( Z_i \), meaning that \( Y' = Z_i \) if \( X \in B_i \). If \( V = 0 \), we instead let \( W \) determine the choice of \( Z_i \), meaning that \( Y' = Z_i \) if \( W = i \). Written out in a formula, we let

\[
Y' = \begin{cases} 
Z_0 & \text{on } (V = 1, X \in B_0) \cup (V = 0, W = 0) \\
Z_1 & \text{on } (V = 1, X \in B_1) \cup (V = 0, W = 1) \\
& \vdots \\
Z_k & \text{on } (V = 1, X \in B_k) \cup (V = 0, W = k)
\end{cases}
\]

Observe that on \( (V = 1, X \notin B_0) \) is holds that \( |X - Y'| \leq 2\varepsilon \). Hence

\[
P(|X - Y'| > 2\varepsilon) \leq P(V = 0) + P(X \in B_0) < 2\varepsilon
\]

The distribution of \( Y' \) is easily found:

\[
P(Y' \in A) = P(Y' \in A, V = 1) + P(Y' \in A, V = 0)
\]

\[
= \sum_{i=0}^{k} P(Y' \in A, V = 1, X \in B_i) + P(Y' \in A, V = 0, W = i)
\]

\[
= \sum_{i=0}^{k} P(Z_i \in A, V = 1, X \in B_i) + P(Z_i \in A, V = 0, W = i)
\]

\[
= \sum_{i=0}^{k} P(Z_i \in A) P(V = 1) P(X \in B_i) + P(Z_i \in A) P(V = 0) P(W = i)
\]

\[
= \sum_{i=0}^{k} P(Z_i \in A) \left( (1 - \varepsilon) P(X \in B_i) + \varepsilon P(W = i) \right)
\]
2.3. The representation theorem

As the distribution of $Z_i$ is equal to the conditional distribution of $Y$ given $(Y \in B_i)$, it is easy to see that $Y'$ has the same distribution as $Y$ if

$$(1 - \varepsilon) P(X \in B_i) + \varepsilon P(W = i) = P(Y \in B_i) \quad \text{for } i = 0, 1, \ldots, k$$

And this holds true if

$$P(W = i) = \frac{P(Y \in B_i) - (1 - \varepsilon) P(X \in B_i)}{\varepsilon} \quad \text{for } i = 0, 1, \ldots, k$$

It is an easy calculation to show that these point probabilities indeed sum to 1, so it is a valid specification of the distribution of $W$.

This concludes the construction for a single variable $Y$, obtaining physical closeness form closeness in distribution (of the specified form). For the sequence version, we can easily repeat the construction for every $Y_1, Y_2, \ldots$ to obtain variables $Y'_1, Y'_2, \ldots$ such that $Y'_n$ has the same distribution as $Y_n$ and such that

$$P(|X - Y'_n| > 2\varepsilon) < 2\varepsilon \quad \text{for every } n \in \mathbb{N}$$

This conclusion is however considerably weaker than (2.1). In order to obtain the stronger conclusion, note the we need not repeat the entire construction for every $Y'_n$. We have to construct new $Z_i$’s, and we have to construct a new $W$, but we can reuse the Bernoulli variable $V$ for every $Y'_n$! Doing that, we obtain that

$$\bigcup_{n=1}^{\infty} (|X - Y'_n| > 2\varepsilon) \subset (V = 0) \cup (X \in B_0)$$

and this larger event has probability less that $2\varepsilon$.

\[\square\]

**Theorem 2.5 (The representation theorem)** Let $(X, d)$ be a metric space equipped with a $\sigma$-algebra $\mathcal{A}$. Let $X, X_1, X_2, \ldots$ be $X$-valued stochastic variables such that $X_n \overset{D}{\rightarrow} X$. If there is a separable set $C \in \mathcal{A}$ consisting of EH-regular points such that $P(X \in C) = 1$, then there is another sequence $X'_1, X'_2, \ldots$ such that $X'_n$ and $X_n$ have the same distribution for every $n$, and such that $X'_n \overset{a.s.}{\rightarrow} X$.

**Proof:** Take $\varepsilon = 2^{-m}$ and find the corresponding partition $B_0, B_1, \ldots, B_k$ from lemma 2.3. Since $P(X \in \partial B_i) = 0$ we can use the convergence lemma to see that

$$P(X_n \in B_i) \rightarrow P(X \in B_i) \quad \text{for } n \rightarrow \infty$$
as explained p. 12. In particular we can find $N_m \in \mathbb{N}$ such that
\[ P(X_n \in B_i) \geq (1 - 2^{-m}) P(X \in B_i) \quad \text{for } i = 0, 1, \ldots, k, \]
for every $n \geq N_m$. Assuming that $N_1 < N_2 < \ldots$ (which can be arranged without problems) we can for $n = N_m, N_m + 1, \ldots, N_{m+1} - 1$ construct $X'_n$ with the same distribution as $X_n$ and such that
\[ P\left(\bigcup_{n=N_m}^{N_{m+1}-1} (|X'_n - X| > 2 \cdot 2^{-m})\right) < 2 \cdot 2^{-m} \]
The Borel-Cantelli lemma shows that
\[ P\left(\bigcup_{n=N_m}^{N_{m+1}-1} (|X'_n - X| > 2 \cdot 2^{-m}) \text{ for infinitely many values of } m\right) = 0 \]
And from this it clearly follows that $X'_n \xrightarrow{a.e.} X$.

**Example 2.6** Assume that $X_n \xrightarrow{D} X$ and that $A \in \mathcal{A}$. We want to consider the relation between $P(X_n \in A)$ and $P(X \in A)$. If $P(X \in \partial A) = 0$ we of course already know that $P(X_n \in A) \rightarrow P(X \in A)$, but it is often useful to have some handle on what happens more generally.

Assume that $A \in \mathcal{A}$ is open. We will show that
\[ \liminf_{n \rightarrow \infty} P(X_n \in A) \geq P(X \in A). \]
To prove this, consider a sequence $X'_1, X'_2, \ldots$ such that $X'_n$ has the same distribution as $X_n$ and such that $X'_n \xrightarrow{\text{a.e.}} X$. If we take a specific $\omega \in \Omega$ such that $X'_n(\omega) \rightarrow X(\omega)$ and if it holds that $X(\omega) \in A$, then clearly $X'_n(\omega) \in A$ for $n$ large enough. This show that
\[ \liminf_{n \rightarrow \infty} 1_{(X'_n \in A)} \geq 1_{(X \in A)} \quad \text{a.e.} \]
Appealing to Fatou’s lemma, we have that
\[ \liminf_{n \rightarrow \infty} P(X_n \in A) = \liminf_{n \rightarrow \infty} \int 1_{(X'_n \in A)} dP \geq \int \liminf_{n \rightarrow \infty} 1_{(X'_n \in A)} dP \geq \int 1_{(X \in A)} dP = P(X \in A) \]
2.3. The representation theorem

as desired. Similarly it can be shown that $A \in A$ is closed, then it holds that

$$\limsup_{n \to \infty} P(X_n \in A) \leq P(X \in A)$$

The easy way to obtain this, is to apply the results for open sets to $A^c$. 

$\circ$