# Mathematical Probability Theory in a Nutshell 2

## Contents

# 1 Mathematical Preliminaries

## 1.1 Borel space, measurable functions, and sigma-algebras

When on the set $\Omega$ the notion of open subsets is well defined, an important sigma-algebra arises. It is the ***Borel algebra*** $\mathscr{B}(\Omega)$ is defined as the smallest sigma-algebra containing all open sets (or, equivalently, all closed sets). The elements of $\mathscr{B}(\Omega)$ are called ***Borel sets*** and the pair $(\Omega, \mathscr{B}(\Omega))$ is called ***Borel space***. The Borel sets of the real line $\mathbb{R}$, can be generated by taking all possible countable unions and intersections of the intervals $(a, b)$. Analogous construction holds for the Borel sets of $\mathbb{R}^n$, where instead of intervals one considers parallelepipeds (that is, cartesian products of intervals), and for the Borel sets of a space which locally is like $\mathbb{R}^n$ (that is, a manifold).

**Measurable functions** A function $f$ from the measurable space $(\Omega, \mathscr{F})$ to the measurable space $(\Sigma, \mathscr{S})$ is said ***measurable*** if

$$f^{-1}(\Delta) = \{\omega \in \Omega : f(\omega) \in \Delta\} \in \mathscr{F} \quad \text{for all} \quad \Delta \in \mathscr{S}\,.$$

**Sub-sigma-algebras** Given a measurable space $(\Omega, \mathscr{F})$, sometimes it is useful to consider a sub-sigma-algebra $\mathscr{A} \subset \mathscr{F}$ and the class of functions which are $\mathscr{A}$-measurable. Clearly, all the $\mathscr{A}$-measurable functions are $\mathscr{F}$-measurable, but not vice-versa.

**Simple sigma-algebras** Among the sub-sigma-algebras that are useful in the applications, there are the ***simple sigma-algebras***, namely those that are generated by a finite (or numerable) collection of subsets. More precisely, if $A_1, A_2, \ldots$ are elements of $\mathscr{F}$), the sigma-algebra that they generate is the smallest sigma-algebra that contains them. This algebra is usually denoted by $\sigma\{A_1, A_2, \ldots\}$ and the sets $A_1, A_2, \ldots$ are sometimes called the ***atoms*** of the algebra.

## 1.2 Lebesgue measure on $\mathbb{R}^n$

Let $C = I_1 \times I_2 \times \cdots \times I_n$ the cartesian product of $n$ intervals $I_i$ forming a (rectangular) parallelepiped in $\mathbb{R}^n$. Each interval is an open (or closed, or semi-open) Borel set of $\mathbb{R}$ of length $|I_i|$. It is a well known fact that he volume of $C$ is $|C| = \prod_{i=1}^n |I_i|$. Given a subset $E \subset \mathbb{R}^n$, the Lebesgue outer volume $|E|^*$ is defined as

$$|E|^* = \inf\left\{\sum_{\alpha=1}^{\infty} |C_\alpha|\right\}$$

where $C_\alpha$, $\alpha = 1, 2, \ldots$, is a sequence of open parallelepipeds with

$$E \subset \bigcup_{\alpha=1}^{\infty} C_\alpha\,.$$

The Lebesgue measure of $E$ is given by its Lebesgue outer measure $|E| = |E|^*$ if, for every $A \subset \mathbb{R}^n$,

$$|A|^* = |A \cap E|^* + |A \cap E^c|^*\ .$$

When this condition is verified, $E$ is said Lebesgue measurable and its measure (or volume) is denoted $|E|$.

The family of subsets of $\mathbb{R}^n$ that are Lebesgue measurable form a sigma-algebra contained in the Borel algebra. In other words, if $E \in \mathsf{B}(\mathbb{R}^n)$, it is automatically Lebesgue measurable.

## 1.3 Miscellanea about measures

**Lebesgue-Stiltjes measure on the real line** Lebesgue-Stiltjes measure on $\mathbb{R}$ is a generalization of the Lebesgue measure obtained in the following way: let $F$ be a real valued, non-decreasing function, defined for $-\infty \leq x \leq +\infty$, and continuos from the right,

$$F(x+0) = \lim_{\epsilon \to 0} F(x+\epsilon) = F(x)$$

For any semi-open intervals $[a,b)$, $a \leq b$, we define $\mu\{[a,b)\} = F(b) - F(a)$. The unique extension of this set functions on the real line called the Lebesgue-Stiltjes measure on the real line.

**Discrete measures** We obtain a discrete measure by letting $F(x)$ be constant except for a countable number of discontinuities $y_k$, where $F(y_k) = F(y_k - 0) + p_k$. The measure is then said to be concentrated at the points $x_k$ with the weights $p_k > 0$. Though this $F$ is not differentiable in the usual sense, is nevertheless differentiable in the distributional sense and can be expressed using Dirac's delta function:

$$\rho(y) = \frac{dF(y)}{dy} = \sum_k p_k \delta(y - y_k) \tag{1}$$

So, by allowing densities to be represented by delta functions, we are able to express with same notation both discrete measures and measures that are absolutely continuos with respect to the Lebesgue measure. Though this does not cover all cases, it is sufficient for the most applications of measure theory to physics (the measures that are left out are those that are the singular continuos measures, i.e., those that concentrated on continuos totally disconnected sets, like the Cantor set and the other fractal sets).

**Finite measures** A measure $\mu$ on $(\Omega, \mathcal{F})$ is said ***finite*** if $\mu(\Omega) < \infty$. The Lebesgue measure is not finite. The Lebesgue-Stiltjes measure for $F(+\infty) < \infty$ is finite.

**Absolute continuity** Measures can be compared. A measure $\mu_1$ is said to be ***absolutely continuous*** with respect to a measure $\mu_2$ if all sets of $\mu_2$-measure zero are also of $\mu_1$ measure zero. Two measures $\mu_1$ and $\mu_2$ are said to be equivalent if each one is absolutely continuos with respect to the other one. The two measures have then the same null sets. For this relation, usually one writes $\mu_1 \sim \mu_2$ and we note that it is an equivalence relation. Two measures $\mu_1$ and $\mu_2$ are said to be mutually ***singular*** if there exist two disjoint sets $A$ and $B$ such that $A \cup B = \Omega$ and such that, for every measurable set $X \subset \Omega$,

$$\mu_1(A \cap X) = \mu_2(B \cup X) = 0$$

Examples of mutually singular measures are easily constructed. If a measure is absolutely continuous with respect to Lebesgue measure, it is simply called absolutely continuous.

**Support of a measure** Let $(\Omega, \mathcal{B}(\Omega))$ be a Borel space and consider a measure $\mu$ on it. The support of $\mu$ (usually denoted by $\text{supp}(\mu)$) is the complement of the union of all open sets which are $\mu$-null sets, i.e. the smallest closed set $C$ such that $\mu(\Omega \setminus C) = 0$. Usually, one says that the measure is ***concentrated*** on $C$.

**Induced measure** Let $\mu$ be a measure on $(\Omega, \mathcal{F})$. Then a measurable function $f$ from the measurable space $(\Omega, \mathcal{F})$ to the measurable space $(\Sigma, \mathcal{S})$ induces on $(\Sigma, \mathcal{S})$ the measure $\hat{\mu} = \mu \circ \mathsf{Y}^{-1}$, i.e.,

$$\hat{\mu}(\Delta) = \mu\{\omega \in \Omega : \mathsf{Y}(\omega) \in \Delta\} \quad \text{for all} \quad \Delta \in \mathcal{S}\,. \tag{2}$$

## 1.4 Integrals

Let us briefly recall the definition of integrals with respect to a measure.

The so called real-valued ***simple functions*** defined by

$$f(\omega) = \sum_{i=1}^{n} \alpha_i \mathbb{1}_{A_i}(\omega)\,, \tag{3}$$

where $A_i \in \mathcal{F}$ and $\mathbb{1}_{A_i}$ is the indicator function of the set $A_i$ (i.e., the function that is equal to $1$ if $\omega \in A_i$ and equal to $0$ otherwise) are measurable. The integral of a simple function is defined by

$$\int f(\omega)\mu(d\omega) = \sum_{i=1}^{n} \alpha_i \mu(A_i)\,. \tag{4}$$

It is a finite number since the $\mu(A_i)$ are all finite (we admit only finite measures).

We now consider sequences of functions $f_n(\omega)$ $(n = 1, 2, \ldots)$ and define ***convergence in the measure*** of such a sequence to a limit function $f(\omega)$ if, for every $\epsilon > 0$,

$$\lim_{n\to\infty} \mu\left(\{\omega : |f_n(\omega) - f(\omega)| \geq \epsilon\}\right) \tag{5}$$

Since simple functions are not only measurable but also integrable, we can define the integrable functions as follows: A finite-valued function $f$ on a measure space $(\Omega, \mathcal{F}, \mu)$ is integrable if there exists a sequence of simple functions $f_n$ such that $f_n$ tends to $f$ in the measure. It follows then that the numbers $\int f_n(\omega)\mu(d\omega)$ tend to a limit which defines the integral of $f$:

$$\int f(\omega)\mu(d\omega) = \lim_{n\to\infty} \int f_n(\omega)\mu(d\omega)\,. \tag{6}$$

Various theorems then permit the usual operations with integrals. For instance, $f_1$ and $f_2$ are integrable, so are $f_1 + f_2$ and $f_1 f_2$, and the integral of the sum is the sum of the integrals. Furthermore, if $f$ is integrable so is $|f|$, and

$$\int f(\omega)\mu(d\omega) \leq \int |f_n(\omega)|\mu(d\omega)\,. \tag{7}$$

**Example** If $\mu$ is the measure concentrated on a discrete set of points $x_k \in \mathbb{R}$ defined by

$$\mu(\Delta) = \sum_{k:x_k\in A} p_k\,, \quad \text{for all} \quad A \in \mathcal{B}(\mathbb{R}), \tag{8}$$

for a given choice of the weights $p_k \geq 0$, then

$$\int f(x)\mu(dy) = \sum_{k} f(x_k)p_k\,. \tag{9}$$

### Notations

The integral of a function $f$ with respect to a measure is denoted

$$\int f(\omega)\mu(d\omega) \quad \text{or} \quad \int f d\mu\,.$$

If $\mu$ is the Lebeseque measure on $\mathbb{R}^n$, the inegral will be denoted

$$\int f(x)\,|dx| \quad \text{or simply, according the standard conventions,}$$

$$\int f(x)dx\,.$$

## 1.5 The Radon-Nikodym theorem

Let $\mu$ be the measure defined by eq. (8). Consider another discrete measure $\nu(x)$ concentrated in the same points, but with different weights $q_k$, not necessarily strictly greater than zero. Clearly the sets of $\nu$-measure zero have also $\mu$-measure zero, so $\nu$ is absolutely continuos with respect to $\mu$. Note that we can write

$$\nu(B) = \int_B \nu(dx) = \sum_{k:x_k\in B} q_k = \sum_{k:x_k\in B} \frac{q_k}{p_k} p_k = \int_B \phi(x)\mu(dx) \tag{10}$$

where $\phi(x) = q_k/w_k$ is $x = x_k$ and is equal to $0$ otherwise. The function $\phi$ is measurable and bounded (since $p_k > 0$) and is called the Radon-Nikodym derivative of $\nu$ with respect to $\mu$. The generalization of this property to any measure is the content of the following theorem.

**The Radon-Nikodym theorem** Let $(\Omega, \mathcal{B})$ be a measurable space and $\mu$ and $\nu$ be two measures on it. If $\nu$ is absolutely continuos with respect to $\mu$, the there exists a unique $\mathcal{B}$-measurable bounded function $\phi(\omega)$ such that

$$\nu(B) = \int_B \phi(\omega)\mu(d\omega) \quad \text{for all} \quad B \in \mathcal{B}\,. \tag{11}$$

If both measures are finite, then $\phi$ is integrable. The function $\phi$ is called the ***Radon-Nikodym*** derivative of $\nu$ with respect to $\mu$ and is usually denoted by $\phi = d\nu/d\mu$.

**Note** The Radon-Nikodym theorem is one of the deepest theorems in measure theory, and its proof is one of the most technical, given such an elementary statement.

## 1.6 Lebesgue integral and Lebesgue-Stiltjes integral

If $\Omega$ is the real line, $\mathcal{F}$ are the Borel sets, and $\mu$ the Lebesgue measure, then the integral is called the Lebesgue integral that we write as $\int f(x)dx$. Similarly we obtain the Lebesgue-Stieltjes integral if we define the measure corresponding to some nondecreasing function $F(x)$. We write for this integral

$$\int f(x)dF(x)\,.$$

If $F(y)$ is differentiable, then $\rho(y) = dF/dy$ is the ***density*** of the measure (that is, the Radon-Nikodym derivative with respect to the Lebesgue measure). In this case

$$\int f(y)dF(y) = \int f(y)\rho(y)dy$$

and the measure

$$\mu(dy) = dF(y) = \rho(y)dy$$

is ***absolutely continuos*** with respect to the Lebesgue measure. Clearly, the sets of measure zero of the Lebesgue measure $dy$ are also sets of measure zero of $\rho(y)dy$.

# 2 Random Variables (II)

Another basic concept of probability is the notion of ***random variable***, meant as a function on the sample space that expresses numerical characteristics of sample points.

## 2.1 The notion of random variable

***A (one-dimensional) random variable on a probability space*** $(\Omega, \mathcal{F}, \mathbb{P})$ ***is a real-valued function*** $\mathsf{Y}$ ***defined on*** $\Omega$ ***such that the set***

$$\mathsf{Y}^{-1}(\Delta) = \{\omega \in \Omega : \mathsf{Y}(\omega) \in \Delta\} \tag{12}$$

***is a random event in*** $\mathcal{F}$ ***for all Borel sets*** $\Delta \in \mathcal{B}(\mathbb{R})$***.***

Thus, from a mathematical point of view, a ***random variable*** is a measurable function from $(\Omega, \mathcal{F})$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ (see section 1.1), in agreement with the intuition that a random variable is a way to selects relevant subsets of $\mathcal{F}$ according to their numerical characteristics: the events $\mathsf{Y}^{-1}(\Delta)$ are elements of $\mathcal{F}$ (since $\mathsf{Y}$ is measurable) and form a sub-sigma-algebra of $\mathcal{F}$ (exercise: prove it). This sigma-algebra will be denoted by $\mathcal{B}(\mathsf{Y})$ and is called the sigma-algebra generated by the random variable $\mathsf{Y}$.

If $\mathbb{P}$ is a probability measure on $(\Omega, \mathcal{F})$, then the induced measure $\widetilde{\mathbb{P}}$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, defined in eq. (2), i.e.,

$$\widetilde{\mathbb{P}}(\Delta) = \mathbb{P}\{\mathsf{Y} \in \Delta\} \quad \text{for all} \quad \Delta \in \mathcal{B}(\mathbb{R})\,.$$

expresses the ***probability distribution*** of the random variable $\mathsf{Y}$. The ***cumulative probability distribution*** of the variable is defined by

$$F(y) = \mathbb{P}\{\mathsf{Y} < y\} = \mu\{(-\infty, y])\}\,. \tag{13}$$

## 2.2 Numerical characteristics of random variables

A basic tool in probability are integrals with respect the probability measure $\mathbb{P}$ of the type

$$\mathbb{E}\{f(\mathsf{Y})\} \equiv \int f(\mathsf{Y}(\omega))\mathbb{P}\{d\omega\} \tag{14}$$

for suitable functions $f : \mathbb{R} \to \mathbb{R}$. In terms of the induced measure, eq. (14) can be rewritten as

$$\mathbb{E}\{f(\mathsf{Y})\} = \int_{-\infty}^{\infty} f(y)\mu(dy) = \int_{-\infty}^{\infty} f(y)dF(y)\,. \tag{15}$$

where the last integral is a Lebesgue-Stieltjes integral on the real line (see section 1.6), this is so because, from its definition, $F$ is a real valued, non-decreasing function, defined for $-\infty \leq x \leq +\infty$, and continuos from the right).

Salient characteristics of random variables are obtained by considering $\mathbb{E}\{f(\mathsf{Y})\}$ given by eq. (14) for suitable functions $f : \mathbb{R} \to \mathbb{R}$ (possibly, depending on a parameter).

For example, the ***$r$-th moment of*** $\mathsf{Y}$, where $r$ is a natural number, corresponds to $f(y) = y^r$. For $r = 1$ we have the ***mean*** or ***expectation***

$$\mathbb{E}\{\mathsf{Y}\} = \int \mathsf{Y}(\omega)\mathbb{P}\{d\omega\} = \int_{-\infty}^{\infty} y\widetilde{\mathbb{P}}(dy) \tag{16}$$

The random variables with finite modulus of $r$-th moment belong to the functional space $L^r(\Omega, \mathscr{F}, \mathbb{P})$; $r = 1$ corresponds to the space of integrable (or summable) random variables for which $\mathbb{E}\{|\mathsf{Y}|\} < \infty$. Particularly useful is the space corresponding to $r = 2$. The random variables with $\mathbb{E}\{\mathsf{Y}^2\} < \infty$ form the Hilbert space $L^2(\Omega, \mathscr{F}, \mathbb{P})$ with scalar product

$$\langle \mathsf{X}, \mathsf{Y} \rangle = \int \overline{\mathsf{X}(\omega)}\mathsf{Y}(\omega)\mathbb{P}(d\omega)\,. \tag{17}$$

This Hilbert space provides a nice geometrical interpretation of various probabilistic notions.

## 2.3 Joint distribution of several random variable

Often we need to consider several random variables at the same time. The joint probability distribution of the random variables $\mathsf{Y} = (\mathsf{Y}_1, \ldots, \mathsf{Y}_n)$ is defined by the set function on $(\mathbb{R}^n, \mathscr{B}(\mathbb{R})^n)$

$$\widetilde{\mathbb{P}}(\Delta_1 \times \cdots \times \Delta_n) = \mathbb{P}\{\mathsf{Y}_1 \in \Delta_1, \ldots \mathsf{Y}_n \in \Delta_n\} \tag{18}$$

which can be completed and extended to a measure $\mu$ on all the Borel sets of $\mathbb{R}^n$. The probability space $(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n), \widetilde{\mathbb{P}})$ contains all the information about the statistical properties of the random variables $\mathsf{Y} = (\mathsf{Y}_1, \ldots, \mathsf{Y}_n)$.

## 2.4 Equivalence of random variables

There are several different senses in which random variables can be considered to be equivalent. Two random variables can be equal, equal almost surely, or equal in distribution. In decreasing order of strength, the precise definition of these notions of equivalence is given below.

**Equality** Two random variables $\mathsf{X}$ and $\mathsf{Y}$ are equal ($=$) if they are equal as functions on $(\Omega, \mathscr{F})$.

**Almost sure equality** Two random variables $\mathsf{X}$ and $\mathsf{Y}$ are equal almost surely if, and only if, the probability that they are different is zero, $\mathbb{P}\{\mathsf{X} \neq \mathsf{Y}\} = 0$. In this case we shall write

$$\mathsf{X} \overset{\text{a.s.}}{=} \mathsf{Y}\,. \tag{19}$$

**Equality in distribution** The random variables $\mathsf{X}$ and $\mathsf{Y}$ are equal in distribution if they have the same distribution functions. In this case we shall write

$$\mathsf{X} \overset{\mathscr{D}}{=} \mathsf{Y}\,. \tag{20}$$

**Equality up to a location-scale transformation** The weakest sense of equivalence corresponds to the case in which the random variable Z is equal in distribution to Y up to a linear transformation, i.e., $\mathsf{Z} \overset{\mathscr{D}}{=} a + b\mathsf{Y}$, with $a \in \mathbb{R}$ and $b \in (0, \infty)$; $a$ is called the ***location parameter*** and $b$ the ***scale parameter***. In this case, Z and Y are said to be of the ***same type*** and this is indicated by writing

$$\mathsf{Z} \overset{\mathscr{D}}{\approx} \mathsf{Y}\,. \tag{21}$$

To be of the same type is an equivalence relation among random variables.

# 3 Conditional Exptectations

## 3.1 Conditional expectation with respect to a random variable

The notion of conditional probability defined by **??** extends immediately to conditioning by a random variable X taking only a countable number of values $x_k$, $k = 1, 2, \ldots$: the ***conditional probability of*** $A$ ***given*** $\mathsf{X} = x_k$ is defined by

$$\mathbb{P}\{A|\mathsf{X} = x_k\} = \frac{\mathbb{P}\{A, \mathsf{X} = x_k\}}{\mathbb{P}\{\mathsf{X} = x_k\}} \tag{22}$$

if $\mathbb{P}\{\mathsf{X} = x_k\} > 0$ and conventionally defined as zero if $\mathbb{P}\{\mathsf{X} = x_k\} = 0$. The set function $A \to \mathbb{P}\{A|\mathsf{X} = x_k\}$ assigning to any $A \in \mathscr{F}$ a number between 0 and 1 is obviously a measure on $(\Omega, \mathscr{F})$ that we shall denote $\mathbb{P}\{d\omega|\mathsf{X} = x_k\}$. We can then define the ***conditional expectation of a random variable*** Y ***given that*** $\mathsf{X} = x_n$ in the following obvious way

$$\mathbb{E}\{Y|\mathsf{X} = x_n\} = \int \mathsf{Y}(\omega)\mathbb{P}\{d\omega|\mathsf{X} = x_n\} \tag{23}$$

if the integral exists.

One would like to generalize eq. (22) and eq. (23) to random variables X taking on nondenumerably many values. Suppose that $\mathbb{P}\{\mathsf{X} \in \Delta\} > 0$, with $\Delta \in \mathscr{B}(\mathbb{R})$. Then, as above, the conditional probability of $A$ given $\mathsf{X} \in \Delta$ is defined by

$$\mathbb{P}\{A|\mathsf{X} \in \Delta\} = \frac{\mathbb{P}\{A, \mathsf{X} \in \Delta\}}{\mathbb{P}\{\mathsf{X} \in \Delta\}}\,. \tag{24}$$

Suppose now that we want to give meaning to the conditional probability of A given $\mathsf{X} = x$. Of course, if $\mathbb{P}\{\mathsf{X} = x\} > 0$, everything is fine but if $\mathbb{P}\{\mathsf{X} = x\} = 0$, there is a problem. A way to circumvent the problem is to consider $\Delta$ in eq. (24) to be the interval $(x - \delta, x + \delta)$ and then to try to define $\mathbb{P}\{A|\mathsf{X} = x\}$ as the limit $\delta \to 0$ of the right hand side of eq. (24). However, in general, there is no guarantee that the limit exists, unless we put restrictive conditions on $\mathbb{P}$ and X. A different route, which avoid imposing such restrictions consists in finding a suitable extension of the of the notion of conditional probability.

## 3.2 Conditional probability and conditional expectation

The formula (**??**) for the total probability suggests a more general notion of conditional probability. Consider the random variable which takes values $\mathbb{P}\{A|B_n\}$ with probabilities $\mathbb{P}\{B_n\}$. Then the right hand side of **??** is just the mean of this random variable. Such a random variable is usually denoted $\mathbb{P}\{A|\mathscr{B}\}$, where $\mathscr{B}$ is the sigma-algebra

generated by the partition (**??**), and is called the ***conditional probability of*** $A$ ***with respect to*** $\mathscr{B}$. From its very definition, it can be represented as as

$$\mathbb{P}\{A|\mathscr{B}\} = \sum_n \mathbb{P}\{A|B_n\}\mathbb{1}_{B_n}(\omega) = \sum_n \frac{\mathbb{P}\{A \cap B_n\}}{\mathbb{P}\{B_n\}}\mathbb{1}_{B_n}(\omega) \tag{25}$$

where $\mathbb{1}_{B_n}(\omega)$ is the indicator function of the set $B_n$. Then we can then rewrite **??** as follows:

$$\mathbb{P}\{A\} = \mathbb{E}\{\mathbb{P}\{A|\mathscr{B}\}\}\,. \tag{26}$$

The conceptual point here is that we are modeling conditional probabilities not as a number, but as a random variable and on a closer inspection this makes good sense: $\mathbb{P}\{A|\mathscr{B}\}$ is a random variable because which of the events $B_n$ in the partition will occur depends on $\omega$.

The notion just defined leads to useful extensions. The set function $A \to \mathbb{P}\{A|\mathscr{B}\}$ assigning to any $A \in \mathscr{F}$ a number between 0 and 1 is obviously a measure on $(\Omega, \mathscr{F})$, and since $\mathbb{P}\{A|\mathscr{B}\}$ is random, it is a ***random*** measure. Let us denote it by $\mathbb{P}\{d\omega'|\mathscr{B}\}(\omega)$. We can then define the ***conditional expectation of a random variable*** $\mathsf{Y}$ ***with respect to the sigma-algebra*** $\mathscr{B}$ as follows:

$$\begin{aligned}\mathbb{E}\{\mathsf{Y}|\mathscr{B}\} &= \int \mathsf{Y}(\omega')\mathbb{P}\{d\omega'|\mathscr{B}\} = \int \mathsf{Y}(\omega')\sum_n \frac{\mathbb{P}\{d\omega' \cap B_n\}}{\mathbb{P}\{B_n\}}\mathbb{1}_{B_n}(\omega) \\ &= \sum_n \frac{\int_{B_n}\mathsf{Y}(\omega')\mathbb{P}\{d\omega'\}}{\mathbb{P}\{B_n\}}\mathbb{1}_{B_n}(\omega)\end{aligned} \tag{27}$$

This is clearly a random variable and it reduces to $\mathbb{P}\{A|\mathscr{B}\}$ for $\mathsf{Y} = \mathbb{1}_A$, the indicator function of the set $A$; that is, the conditional probability of $A$ is the conditional expectation of $\mathbb{1}_A$.

Also eq. (23) defines a random variable, namely

$$\mathbb{E}\{\mathsf{Y}|\mathsf{X}\} = \int \mathsf{Y}(\omega)\mathbb{P}\{d\omega|\mathsf{X}\} \tag{28}$$

expressing the conditional expectation of the random variable $\mathsf{Y}$ given the random variable $\mathsf{X}$. Surprisingly (or maybe not), we have

$$\mathbb{E}\{\mathsf{Y}|\mathsf{X}\} = \mathbb{E}\{\mathsf{Y}|\mathscr{B}(\mathsf{X})\} \tag{29}$$

where $\mathscr{B}(\mathsf{X})$ is the sigma-algebra generated by $\mathsf{X}$, namely the sigma-algebra generated by level sets of $\mathsf{X}$, $B_n = \{\omega \in \Omega : \mathsf{X}(\omega) = x_n\}$. This follows easily from noting that the quantities that multiplies the indicator function of the set $B_n$ in the right hand side of eq. (27) are indeed equal to $\mathbb{E}\{Y|\mathsf{X} = x_n\}$ and that

$$\sum_n \mathbb{E}\{\mathsf{Y}|\mathsf{X} = x_n\}\mathbb{1}_{\{\mathsf{X}=x_n\}} = \mathbb{E}\{\mathsf{Y}|\mathsf{X}\}\,. \tag{30}$$

The conditional expectation fulfills a condition analogous to (26), namely

$$\mathbb{E}\{\mathsf{Y}\} = \mathbb{E}\{\mathbb{E}\{\mathsf{Y}|\mathscr{B}\}\} \tag{31}$$

for any (summable) random variable $\mathsf{Y}$. Note that when $\mathsf{Y}$ is measurable with respect to $\mathscr{B}$, that is, $\mathsf{Y} = \sum_k y_k \mathbb{1}_{B_n}$, then

$$\mathbb{E}\{\mathsf{Y}|\mathscr{B}\} = \mathsf{Y}\,. \tag{32}$$

In particular, since $\mathbb{E}\{\mathsf{Y}|\mathscr{B}\}$ itself is measurable with respect to $\mathscr{B}$, we have

$$\mathbb{E}\{\mathbb{E}\{Y|\mathscr{B}\}|\mathscr{B}\} = \mathbb{E}\{Y|\mathscr{B}\} \ . \tag{33}$$

Conditional probability theory is one of the most difficult parts of basic probability theory. The reason is that it is hard to come up with good intuitions for it. The following physical intuition may help. Suppose that $\Omega$ is the space of the microstates of a physical system. Then a partition $\mathscr{B}$ of $\Omega$ arises by assembling together microstates $\omega$ that have the same macroscopic appearance—one says that the partition provides a ***coarse-graining*** of the set of the microstates. Suppose that $\mathsf{Y}$ is a physical quantity that varies on a much smaller scale and denote by $\mathscr{F}$ the finer partition of $\Omega$ on which $\mathsf{Y}$ can assumed to be constant. Since $\mathscr{F}$ has a finer resolution than $\mathscr{B}$, conditional expectation of $\mathsf{Y}$ given $\mathscr{B}$ must be some way of coarse-graining $\mathsf{Y}$ as to be identified only on the basis of the resolution provided by the macroscopic scale $\mathscr{B}$. The way to do that is by averaging out $\mathsf{Y}$ on the those smaller subsets of $\mathscr{F}$ and making it constant on the larger subsets of $\mathscr{B}$. This is indeed what formula (27) does.

## 3.3 Conditional expectation as orthogonal projector

The first way to extend the notion of conditional expectation is to start form the following geometrical interpretation of formula (27). Let $\mathsf{Y} \in L^2(\Omega, \mathscr{F}, \mathbb{P})$. Then $\mathbb{E}\{\mathsf{Y}|\mathsf{B}\}$ is the orthogonal projection of $\mathsf{Y}$ onto $L^2(\Omega, \mathscr{F}, \mathbb{P}) \subset L^2(\Omega, \mathscr{B}, \mathbb{P})$, where $\mathscr{B}$ is the sigma-algebra generated by $\{B_n\}$. In fact, the functions

$$\frac{\mathbb{1}_{B_n}}{\|\mathbb{1}_{B_n}\|} = \frac{\mathbb{1}_{B_n}}{\sqrt{\mathbb{P}\{\mathbb{1}_{B_n}\}}}$$

are an orthonormal basis in $L^2(\Omega, \mathscr{B}, \mathbb{P})$ and thus the orthogonal projection of $\mathsf{Y}$ onto $L^2(\Omega, \mathscr{B}, \mathbb{P})$ is (bra-ket notation)

$$\sum_n \left| \frac{\mathbb{1}_{B_n}}{\|\mathbb{1}_{B_n}\|} \right\rangle \left\langle \frac{\mathbb{1}_{B_n}}{\|\mathbb{1}_{B_n}\|} \middle| \mathsf{Y} \right\rangle = \sum_n \frac{\mathbb{1}_{B_n}}{\mathbb{P}\{B_n\}} \int \mathbb{1}_{B_n} \mathsf{Y}(\omega) \mathbb{P}\{d\omega\} = \sum_n \frac{\mathbb{1}_{B_n}}{\mathbb{P}\{B_n\}} \int_{B_n} \mathsf{Y}(\omega) \mathbb{P}\{d\omega\} \ ,$$

which is exactly the right hand side of eq. (27). Therefore, we may regard $\mathbb{E}\{\mathsf{Y}|\mathscr{B}\}$ as ***the best approximation*** of $\mathsf{Y}$ in the space $L^2(\Omega, \mathscr{B}, \mathbb{P})$ of $\mathscr{B}$-measurable random variables with finite $2nd$ moment. Going back to the phsyical example above, if $\mathsf{Y}$ is a physical quantity, then, among the macroscopic variables (that is, the $\mathscr{B}$-measurable functions), $\mathbb{E}\{\mathsf{Y}|\mathscr{B}\}$ is the closest one to $\mathsf{Y}$ in the mean quadratic sense.

This geometrical characterization of the conditional expectation extends naturally to any sigma-algebra $\mathscr{B} \subset \mathscr{F}$, whether or not it is generated by a partition. Therefore, it is very natural to define the conditional expectation $\mathbb{E}\{\mathsf{Y}|\mathscr{B}\}$ of a random variable $\mathsf{Y}$ with respect to any sigma-algebra $\mathscr{B}$ as

$$\mathbb{E}\{\mathsf{Y}|\mathscr{B}\} = \text{orthogonal projection of } \mathsf{Y} \text{ onto } L^2(\Omega, \mathscr{B}, \mathbb{P}). \tag{34}$$

Note how transparent is the geometrical meaning of eq. (32) and eq. (33): if a vector is already in the subspace, projecting it onto the subspace will do nothing to it. The conditional expectation so defined is restricted only to $L^2$ random variables. However, since $L^2 \subset L^1$ is dense in $L^1$, we may define a unique extension of $\mathbb{E}\{\cdot|\mathscr{B}\}$ to a linear operator from $L^1(\Omega, \mathscr{F}, \mathbb{P})$ (the space of random variables that are absolutely integrable) to $L^1(\Omega, \mathscr{F}, \mathbb{P})$.

The foregoing is summarized by the following definition:

***The conditional expectation*** $\mathbb{E}\{\mathsf{Y}|\mathscr{B}\}$ ***of a summable random variable*** $\mathsf{Y}$ ***with respect to a sigma-algebra*** $\mathscr{B}$ ***is defined as the unique extension of the orthogonal projection of*** $\mathsf{Y}$ ***onto*** $L^2(\Omega, \mathscr{B}, \mathbb{P})$.

## 3.4 Conditional expectation as Radom-Nicodym derivative

The second way to extend the notion of conditional expectation is to start from a measure-theoretic interpretation of formula (27). Consider the conditional expectation (27) of $\mathsf{Y} = \mathbb{1}_A$, that is the conditional probability of $A$. By integrating the right hand side of (27) on a set $B_k$ of the partition, we obtain

$$\mathbb{Q}(B_k) = \int_{B_k} \mathbb{1}_A(\omega)\mathbb{P}\{d\omega\} = \mathbb{P}\{A \cap B_k\}\,. \tag{35}$$

By additivity, we obtain a measure $\mathbb{Q}$ on the sigma-algebra $\mathscr{B}$ generated by the partition which is absolutely continuos (see section 1.3) with respect to the measure $\mathbb{P}$ restricted to $\mathscr{B}$; in fact, if $\mathbb{P}\{B_k\} = 0$, clearly $\mathbb{Q}(B_k) = 0$. Then, it is easy to see that the conditional expectation of $\mathsf{Y} = \mathbb{1}_A$ given by (27) is the Radon-Nikodym derivative (see eq. (10) in the appendix) of $\mathbb{Q}$ with respect to $\mathbb{P}$. For a general random variable $\mathsf{Y}$ one need to define

$$\mathbb{Q}(B_k) = \int_{B_k} \mathsf{Y}(\omega)\mathbb{P}\{d\omega\}\,. \tag{36}$$

and convince himself that the absolute continuity of $\mathbb{Q}$ to the restriction of $\mathbb{P}$ still holds. In this way one arrives at the following formula

$$\mathbb{E}\{\mathsf{Y}|\mathscr{B}\}(\omega) = \frac{d\mathbb{Q}}{d\mathbb{P}}(\omega)\,, \tag{37}$$

which is meaningful for a general sigma-algebra $\mathscr{B} \subset \mathscr{F}$, whether or not it is generated by a partition.

More precisely, let $\mathsf{Y}$ be a summable random variable on $(\Omega, \mathscr{F})$ and let $\mathscr{B}$ a sub-sigma-algebra of $\mathscr{F}$. Then we may define a measure $\mathbb{Q}$ on $(\Omega, \mathscr{B})$ by setting

$$\mathbb{Q}(B) = \int_B Y(\omega)\mathbb{P}\{d\omega\} \quad \text{for all} \quad B \in \mathscr{B}\,. \tag{38}$$

The measure $\mathbb{P}$ may be restricted to a measure on $(\Omega, \mathscr{B})$, which we shall also denote by $\mathbb{P}$. As measure on $(\Omega, \mathscr{B})$, $\mathbb{Q}$ is absolutely continuous with respect to $\mathbb{P}$ (see section 1.3). Indeed if $\mathbb{P}\{B\} = 0$ then it is easy to see that

$$\mathbb{Q}(B) = \int_B Y(\omega)\mathbb{P}\{d\omega\} = 0\,,$$

say by approximating $\mathsf{Y}$ by an increasing sequence of step functions. Then, by the Radon-Nikodym theorem (see section 1.5) there is a $\mathscr{B}$-measurable function $\phi$ such that

$$\int_B Y(\omega)\mathbb{P}\{d\omega\} = \mathbb{Q}(B) = \int_B \phi(\omega)\mathbb{P}\{d\omega\} \quad \text{for all} \quad B \in \mathscr{B}\,. \tag{39}$$

The function $\phi$ is unique up to sets of measure zero and the Radon-Nikodym derivative of $\mathbb{Q}$ with respect to $\mathbb{P}$ (see section 1.5), whence eq. (37) for a general sigma-algebra.

This definition is completely equivalent to definition **??**. To see this, use eq. (39), firstly, to show that if $\mathsf{Y} \in L^2(\Omega, \mathscr{F}, \mathbb{P})$, then $\phi \in L^2(\Omega, \mathscr{B}, \mathbb{P})$ and secondly, to see that $Y - \mathbb{E}\{\mathsf{Y}|\mathscr{B}\}$ is orthogonal to all functions in $L^2(\Omega, \mathscr{B}, \mathbb{P})$. Finally, check that $\mathsf{Y} \to \phi = \mathbb{E}\{\mathsf{Y}|\mathscr{B}\}$ is a bounded linear functional, and conclude that it must be the same object characterized by definition **??**.

We summarize below this second characterization of the conditional expectation.

***The conditional expectation*** $\mathbb{E}\{\mathsf{Y}|\mathscr{B}\}$ ***of a summable random variable*** $\mathsf{Y}$ ***with respect to a sigma-algebra*** $\mathscr{B}$ ***is defined as the Radon-Nikodym derivative of the measure***

$$\mathbb{Q}(B) = \int_B Y(\omega)\mathbb{P}\{d\omega\} \quad \textbf{\textit{for all}} \quad B \in \mathscr{B}$$

***with respect to the measure*** $\mathbb{P}$ ***restricted to*** $(\Omega, \mathscr{B})$***.***

## 3.5 Conditional distributions

Let us now go back to the problem of defining the conditional probability $\mathbb{P}\{A|\mathsf{X}=x\}$ when the event $\mathsf{X}=x$ has probability zero. The right had side of eq. (24) is indeed the ratio of two measures on $\mathscr{B}(\mathbb{R})$: $\hat{\mathbb{Q}}(\Delta)=\mathbb{P}\{A,\mathsf{X}\in\Delta\}$ and $\widetilde{\mathbb{P}}(\Delta)=\mathbb{P}\{\mathsf{X}\in\Delta\}$, the first being absolutely continuos with respect to the second. So $\mathbb{P}\{A|\mathsf{X}=x\}$ is naturally defined as the Radon-Nicodym derivative of $\hat{\mathbb{Q}}$ with respect to $\widetilde{\mathbb{P}}$. Thus,

$$\mathbb{P}\{A|\mathsf{X}=x\}=\frac{d\hat{\mathbb{Q}}}{d\widetilde{\mathbb{P}}}(x) \tag{40}$$

The connnection with foregoing is established as follows. Let $\mathbb{Q}$ be the measure on $\mathscr{F}$ defined by eq. (38) for $\mathscr{B}=\mathscr{B}(\mathsf{X})$ and $Y=\mathbb{1}_A$. Then

$$\mathbb{P}\{A|\mathsf{X}\}(\omega)=\frac{d\hat{\mathbb{Q}}}{d\widetilde{\mathbb{P}}}(\mathsf{X}(\omega))=\frac{d\mathbb{Q}}{d\mathbb{P}}(\omega) \tag{41}$$

is the conditional probability of $A$ with respect to the sigma-algebra generated by $\mathsf{X}$, that is, it the extension to the general case of the conditional probability (expectation) (29) with respect to a random variable $\mathsf{X}$.

The following point deserves attention. Since the Radon-Nicodym derivative is unique up to sets of measure $0$, $\mathbb{P}\{A|\mathsf{X}=x\}$ is not defined pointwise with respect to $x$. If you choose a specific function $\phi(x)$ which satisfies

$$\hat{\mathbb{Q}}(\Delta)=\int_\Delta \phi(x)\widetilde{\mathbb{P}}(dx)\quad\text{for all}\quad\Delta\in\mathscr{B}(\mathbb{R})\,,$$

then any other function differing from $\phi(x)$ on a set of $\mathbb{P}$-measure $0$ is good as well. A similar comment applies to $\mathbb{P}\{A|\mathsf{X}\}$ as a function of $\omega$.

Let us now consider the case in which the event $A$ is of the form $\mathsf{Y}\in\Gamma$, where $\mathsf{Y}$ is another random variable and $\Gamma\in\mathscr{B}(\mathbb{R})$. Then $\hat{\mathbb{Q}}(\Delta)=\widetilde{\mathbb{P}}(\Gamma\times\Delta)$, where $\widetilde{\mathbb{P}}(\Gamma\times\Delta)$ is the joint distribution of $\mathsf{Y}$ and $\mathsf{X}$. Suppose that $\mathsf{X}$ and $\mathsf{Y}$ have a joint density $\rho(y,x)$, i.e.,

$$\widetilde{\mathbb{P}}(\Gamma\times\Delta)=\int_\Gamma\int_\Delta \rho(y,x)dydx\,,$$

so that

$$\widetilde{\mathbb{P}}(\Delta)=\int_{\mathbb{R}}\int_\Delta \rho(y,x)dydx=\int_\Delta\rho(x)dx\,,\quad\text{where}\quad\rho(x)=\int_{\mathbb{R}}\rho(y,x)dy\,.$$

Since

$$\widetilde{\mathbb{P}}(\Gamma\times\Delta)=\int_\Delta\left[\int_\Gamma\frac{\rho(y,x}{\rho(x)}dy\right]\rho(x)dx\,, \tag{42}$$

we see that

$$\mathbb{P}\{\mathsf{Y}\in\Gamma|\mathsf{X}=x\}=\int_\Gamma\frac{\rho(y,x)}{\rho(x)}dx \tag{43}$$

where the equality is in the "almost everywhere" sense.

## 3.6 Conditional Independence

Two events $A$ and $B$ are conditionally independent given a third event $X$ precisely if the occurrence or non-occurrence of $A$ and the occurrence or non-occurrence of $B$ are independent events in their conditional probability distribution given $X$, that is,

$$\mathbb{P}\{A\cap B|X\}=\mathbb{P}\{A|X\}\,\mathbb{P}\{B|X\}\,, \tag{44}$$

or equivalently,

$$\mathbb{P}\{A|B\cap X\}=\mathbb{P}\{A|X\}\ . \tag{45}$$

The notion of conditional independence clarifies various issues related to the notion of independence. Let us quote a passage from "Bertlmann's socks and the nature of reality" where John Bell illustrate this notion—relevant for a proper understanding of Bell's theorem—with a simple example:

> *For example the statistics of heart attacks in Lille and Lyons show strong correlations. The probability of $M$ cases in Lyons and $N$ in Lille, on a randomly chosen day, does not separate:*
>
> $$P(M,N)\neq P_1(M)P_2(N)$$
>
> *In fact when $M$ is above average $N$ also tends to be above average. You might shrug your shoulders and say 'coincidences happen all the time', or 'that's life'. Such an attitude is indeed sometimes advocated by otherwise serious people in the context of quantum philosophy. But outside that peculiar context, such an attitude would be dismissed as unscientific. The scientific attitude is that correlations cry out for explanation. And of course in the given example explanations are soon found. The weather is much the same in the two towns, and hot days are bad for heart attacks. The day of the week is exactly the same in the two towns, and Sundays are especially bad because of family quarrels and too much to eat. And so on. It seems reasonable to expect that if sufficiently many such causal factors can be identified and held fixed, the residual fluctuations will be independent, i.e.,*
>
> $$P(M,N|a,b,\lambda)=P_1(M|a,\lambda)P_2(N|b,\lambda) \tag{10}$$
>
> *where $a$ and $b$ are temperatures in Lyons and Lille respectively, $A$ denotes any number of other variables that might be relevant, and $P(M,N|a,b,\lambda)$ is the conditional probability of $M$ cases in Lyons and $N$ in Lille for given $(a,b,\lambda)$. Note well that we already incorporate in (10) a hypothesis of 'local causality' or 'no action at a distance'. For we do not allow the first factor to depend on $b$, nor the second on $a$. That is, we do not admit the temperature in Lyons as a causal influence in Lille, and vice versa.*

## 3.7 Conditional independence with respect to a sigma-algebra

Two events $A$ and $B$ are conditionally independent given a sigma-algebra $\mathscr{A}$ if

$$\mathbb{P}\{A\cap B|\mathscr{A}\}=\mathbb{P}\{A|\mathscr{B}\}\,\mathbb{P}\{B|\mathscr{A}\}\ . \tag{46}$$

Two random variables X and Y are conditionally independent given a sigma-algebra $\mathscr{A}$ if the above equation holds for all $A$ in $\mathscr{B}(\mathsf{X})$ and $B$ in $\mathscr{B}(\mathsf{Y})$.

Two random variables X and Y are conditionally independent given a random variable Z, if they are independent given the sigma-algebra $\mathscr{B}(\mathsf{Z})$ generated by Z. This is commonly written $\mathsf{X}\perp\!\!\!\perp\mathsf{Y}|\mathsf{Z}$ and is read " X is independent of Y, given Z." If Z assumes a countable set of values $z$, this is equivalent to the conditional independence of X and Y for the events of the form $\{\mathsf{Z}=z\}$ Conditional independence of more than two events, or of more than two random variables, is defined analogously.

The following two examples show that $\mathsf{X}\perp\!\!\!\perp\mathsf{Y}$ neither implies nor is implied by $\mathsf{X}\perp\!\!\!\perp\mathsf{Y}|\mathsf{Z}$. First, suppose Z has value $0$ with probability $1/2$ and value $1$ otherwise. When $\mathsf{Z}=0$ take X and Y to be independent, each having the value $0$ with probability $0.99$

and the value $1$ otherwise. When $\mathsf{Z} = 1$, $\mathsf{X}$ and $\mathsf{Y}$ are again independent, but this time they take the value $1$ with probability $0.99$. Then $\mathsf{X} \perp\!\!\!\perp \mathsf{Y}|\mathsf{Z}$. But $\mathsf{X}$ and $\mathsf{Y}$ are dependent, because $\mathbb{P}\{\mathsf{X} = 0\} < \mathbb{P}\{\mathsf{X} = 0|\mathsf{Y} = 1\}$. This is because $\mathbb{P}\{\mathsf{X} = 0\} = 1/2$, but if $\mathsf{Y} = 0$ then it's very likely that $\mathsf{Z} = 0$ and thus that $\mathsf{X} = 0$ as well, so $\mathbb{P}\{\mathsf{X} = 0|\mathsf{Y} = 0\} > 1/2$. (Another example showing that two variables might be dependent, but independent given a third variable is the above cited example of Bell.) For the second example—that independence of two variables does not imply their independence given a third variable—suppose $\mathsf{X} \perp\!\!\!\perp \mathsf{Y}$, each taking the values 0 and 1 with probability $1/2$. Let $\mathsf{Z} = \mathsf{XY}$. Then when $\mathsf{Z} = 0$, $\mathbb{P}\{\mathsf{X} = 0\} = 2/3$, but $\mathbb{P}\{\mathsf{X} = 0|\mathsf{Y} = 0\} = 1/2$, so $\mathsf{X} \perp\!\!\!\perp \mathsf{Y}|\mathsf{Z}$ is false.