Mathematical Probability Theory in a Nutshell 1

Contents

1 Kolmogorov' Synthesis

The basic notion in probability ***seems*** the notion of randomness.

(There are experiments whose results are not predictable and can be determined only after performing it and then observing the outcome. The simplest familiar examples are, the tossing of a fair coin, or the throwing of a balanced die. This intuition suggests that a mathematical formalization of probability should be build on the notion of random events or random sequences.)

A different approach has been proposed by Andrey Nikolaevich Kolmogorov in the book Foundations of the Theory of Probability.

The monograph appeared as ***Grundbegriffe der Wahrscheinlichkeitsrechnung*** in 1933 and build up probability theory in a rigorous way similar as Euclid did with geometry. The ***Grundbegriffe*** was a work of synthesis of the developments of mathematical probability for more than three centuries.

In Kolmogorov's axiomatics of probability theory the con-

cept of a random event is not primary and is constructed out of more elementary notions. His approach bears a resemblance to the approach to Statistical Mechanics put forward by Boltzmann and Gibbs.

1.1 Kolmogorov's Axioms

I $\mathfrak{F}$ ***is a field of sets***.

II $\mathfrak{F}$ ***contains the set*** Ω***.***

III ***To each set*** A ***from*** $\mathfrak{F}$ ***is assigned a nonnegative real number*** $\mathbb{P}\{A\}$***. This number*** $\mathbb{P}\{A\}$ ***is called the probability of the event*** A***.***

IV $\mathbb{P}\{\Omega\} = 1$.

V ***If*** A ***and*** B ***are disjoint, then*** $\mathbb{P}\{A \cup B\} = \mathbb{P}\{A\} + \mathbb{P}\{B\}$***.***

VI If $A_1 \supseteq A_2 \supseteq \ldots$ is a decreasing sequence of events from $\mathfrak{F}$ with $\cap_{n=1}^{\infty} A_n = \emptyset$, then $\lim_{n\to\infty} \mathbb{P}\{A_n\} = 0$.

1.2 Axioms I and II

Kolmogorov starts from a set (a space) Ω of ***elementary*** events, i.e., simple events. The elements ω of this set are immaterial for the logical development of the theory of probability.

Ω is usually called the ***sample space***, the elementary events $\omega \in \Omega$ are also called ***sample points***.

The next to be considered is a certain family $\mathfrak{F}$ of subsets of the set Ω. The elements of the family $\mathfrak{F}$ are called random events. It is assumed that the following requirements are fulfilled relative to the structure of the family $\mathfrak{F}$.

1. $\mathfrak{F}$ contains the set Ω as one of its elements.

2. If A and B, subsets of Ω, are elements of $\mathfrak{F}$, then the sets $A \cup B$, $A \cap B$, $A^c = \Omega \setminus A$ and $B^c = \Omega \setminus B$ are also elements of $\mathfrak{F}$.

In other words, $\mathfrak{F}$ ***is closed under the set boolean operations complement, union and intersection.*** Any such $\mathfrak{F}$ is called a ***field*** of events or an ***algebra*** of events.

In many very important problems we shall have to demand more:

3. If subsets $A_1, A_2, \ldots, A_n, \ldots$ of set Ω are elements of the set $\mathfrak{F}$, then their union $\bigcap_n A_n$ and intersection $\bigcup_n A_n$ are also elements of $\mathfrak{F}$.

In other words, $\mathfrak{F}$ is closed under countable-fold set boolean operations . The set $\mathfrak{F}$ formed in this fashion is called a ***Borel field*** of events; another now frequently used term is ***sigma-algebra*** of events.

Kolmogorov motivations of I and II Suppose we have a certain system of conditions $\mathfrak{S}$, capable of unlimited repetition and let us consider a fixed circle of phenomena that can arise when the conditions $\mathfrak{S}$ are realized. For example, if water at atmospheric pressure is heated above 100^o C (the set of conditions $\mathfrak{S}$), it is transformed into steam. Or, the system of conditions $\mathfrak{S}$ may consists of flipping a coin twice. In general, these phenomena can come out in different ways in different cases where the conditions are realized. Let Ω be the set of the different possible variants $\omega_1, \omega_2, \ldots$ of the outcomes of the phenomena. Some of these variants might actually not occur.

We include in the set Ω all the variants we regard a priori as possible. In the case of the coin flipped twice, there are four possible variants—***elementary events***; namely

$$\text{head—head , head—tail , tail—head , tail—tail .} \qquad (1)$$

Consider the event A that there is a repetition. This event consists of the first and fourth elementary events. Every event can similarly be regarded as a set of elementary events.

It is natural to introduce the following definitions.

If two random events A and B do not contain the same elements of the set Ω, that is they are ***disjoint***, we shall call them ***mutually exclusive***. The random event Ω will be called ***certain***, and the random event $\emptyset$ (empty set) the ***impossible event***. Events A end A^c are ***contrary*** (complementary) events.

1.3 Axioms III, IV, and V

These are the axioms that define probability. Note that for the classical definition of probability,

$$\mathbb{P}\{A\} = \frac{\text{number of possible trial outcomes favourable}}{\text{number of all possible outcomes}}, \qquad (2)$$

there was no need to postulate the properties expressed by Axioms IV and V, since these properties of probability follows from the definition eq. (2). And the assertion in Axiom III is contained in the classical definition of probability itself. Yet, these axioms can be ***justified*** on the basis of relative frequencies.

Kolmogorov motivations of III and IV and V Consider an experiment that is repeated n times and suppose that m times ($m \leq n$) the event A occurred. Clearly, the frequency $f = m/n$ is a number between 0 and 1, so the assumption that $0 \leq \mathbb{P}\{A\} \leq 1$ appears completely natural. . We always have $m = n$ for the event Ω, so we naturally set $\mathbb{P}\{\Omega\} = 1$. Finally, if A and B are mutually incompatible (in other words, the sets A and B are disjoint), then $m = m_1 + m_2$,

where m, m_1, and m_2 are the numbers of experiments in which the events $A \cup B$, A, and B happen, respectively. It follows that

$$\frac{m}{n} = \frac{m_1}{n} + \frac{m_2}{n} \quad \text{i.e.,} \quad f_{A\cup B} = f_A + f_B \,.$$

So it appears appropriate to set $\mathbb{P}\{A \cup B\} = \mathbb{P}\{A\} + \mathbb{P}\{B\}$.

1.4 Elementary corollaries

From the obvious equality $\Omega = \emptyset \cap \Omega$ and and Axiom V we conclude that $\mathbb{P}\{\Omega\} = \mathbb{P}\{\emptyset\} + \mathbb{P}\{\Omega\}$. Thus, the probability of the impossible event is zero.

Similarly, for any event A, we have $\mathbb{P}\{A^c\} = 1 - \mathbb{P}\{A\}$.

If the event A implies the event B, that is $A \subseteq B$, then $\mathbb{P}\{A\} \leq \mathbb{P}\{B\}$.

Let A and B be two arbitrary events. Insofar as the terms in the unions $A\cup B = A+(B\setminus A\cap B)$ and $B = A\cap B+(B\setminus A\cap B)$, in accordance

with Axiom V

$$\mathbb{P}\{A \cup B\} = \mathbb{P}\{A\} + \mathbb{P}\{B \setminus A \cap B\}$$

and

$$\mathbb{P}\{B\} = \mathbb{P}\{A \cap B\} + \mathbb{P}\{B \setminus A \cap B\}$$

From this follows the addition theorem for arbitrary events A and B:

$$\mathbb{P}\{A \cup B\} = \mathbb{P}\{A\} + \mathbb{P}\{B\} - \mathbb{P}\{A \cap B\} \ , \tag{3}$$

By virtue of the nonnegativity of $\mathbb{P}\{A \cap B\}$ we conclude that

$$\mathbb{P}\{A \cup B\} \leq \mathbb{P}\{A\} + \mathbb{P}\{B\} \ . \tag{4}$$

By induction we now derive that if $A_1, A_2, \ldots, A_n$ are arbitrary events, we have the inequality

$$\mathbb{P}\{A_1 \cup A_2 \cup \ldots \cup A_n\} \leq \mathbb{P}\{A_1\} + \mathbb{P}\{A_2\} + \ldots + \mathbb{P}\{A_n\} \ . \tag{5}$$

The system of axioms of Kolmogorov is ***consistent***, for there exist real objects that satisfy all these axioms. For example, the head tail

system of eq. (1) with

$$\mathbb{P}\{\text{head—head}\} = \mathbb{P}\{\text{head—tail}\} =$$

$$= \mathbb{P}\{\text{tail—head}\} = \mathbb{P}\{\text{tail—tail}\} = \frac{1}{4},$$

or the relative frequencies considered above or the volume of bodies in physical space. The system of axioms of Kolmogorov is ***incomplete***, that is ***is non categorical***: even for one and the same set Ω we can choose the probabilities in the set $\mathcal{F}$ in different ways. To take an example, in the case of the the two dice examined we can either put the above assignment of probabilities or

$$\mathbb{P}\{\text{head—head}\} = \mathbb{P}\{\text{head—tail}\} = \frac{1}{6}$$

$$\mathbb{P}\{\text{tail—head}\} = \mathbb{P}\{\text{tail—tail}\} = \frac{1}{3},$$

The incompleteness of a system of axioms in probability theory is not an indication of an inapt choice, but is due to the essence of the

matter: in various problems there may be phenomena whose study demands consideration of identical sets of random events but with different probabilities.

1.5 Axiom VI and the notion of Probability Space

In probability theory one constantly has to deal with events that decompose into an infinite number of particular cases. To cope with this, Kolmogorov added the sixth axiom, redundant for finite $\mathfrak{F}$ but independent of the first five axioms for infinite ***sigma-algebra*** $\mathfrak{F}$. This is the axiom of continuity. Given the first five axioms, it is equivalent to countable additivity a.k.a sigma-additivity:

VI' If $A_1, A_2, \ldots, A_n, \ldots$ is a sequence of pairwise esclusive events from $\mathfrak{F}$, i.e, $A_i \cap A_j = \emptyset$ for$i \neq j$, then

$$\mathbb{P}\left\{\bigcup_n A_n\right\} = \sum_n \mathbb{P}\left\{A_i\right\} . \tag{6}$$

So, from the point of view of set theory, the axiomatic definition of probability given by Kolmogorov is nothing other than the introduction into the set Ω of a normalized, countably additive, nonnegative measure defined on all elements of the set $\mathcal{F}$.

Measure In general, a measure on a set is a systematic way to assign a number to each suitable subset of that set, intuitively interpreted as its size—it is a generalization of the concepts of length, area, and volume. More precisely, a set function μ from $\mathcal{F}$ to the extended real number line is called a measure if it satisfies the following properties:

(i) Non-negativity: For all $A \in \mathcal{F}$, $\mu(A) \geq 0$.

(ii) Null empty set: $\mu(\emptyset) = 0$.

(iii) Countable additivity (or σ-additivity): For all countable collections of pairwise disjoint sets in $\mathcal{F}$:

$$\mu\left(\bigcup_i A_i\right) = \sum_i \mu(A_i)\,. \tag{7}$$

The pair $(\Omega, \mathcal{F})$ is called a ***measurable space***; the triple $(\Omega, \mathcal{F}, \mu)$ is called a ***measure space***.

Then the six axioms can be summarized by the following proposition.

Probability is represented by a nonnegative measure $\mathbb{P}$ ***on the measurable space*** $(\Omega, \mathcal{F})$ ***with*** $\mathbb{P}\{\Omega\} = 1$***.***

The triple $(\Omega, \mathcal{F}, \mathbb{P})$ is called a ***probability space***. The notion of probability space summarizes the basic mathematical structure of probability. Note the importance of specifying the sigma-algebra $\mathcal{F}$, which is to act as the domain of the probability measure. It defines which sets of outcomes are considered to be "events" and therefore to have probability values.

1.6 The relation to the world of experience

Kolmogorov's ***Grundbegriffe*** put probability's modern mathematical formalism in place. It also provided a ***philosophy of probability***—an explanation of how the formalism can be connected to

the world of experience. Unfortunately, this second contribution of Kolmogorov is almost forgotten or misunderstood.

Kolmogorov relates probability theory with the world of experience on two principles.

From Section 2 of Chapter 1 of the ***Grundbegriffe*** :

> *Under certain conditions, that we will not go into further here, we may assume that an event A that does or does not occur under conditions $\mathfrak{S}$ is assigned a real number $\mathbb{P}\{A\}$ with the following properties:*
>
> *(A) One can be practically certain that if the system of conditions $\mathfrak{S}$ is repeated a large number of times, n, and the event A occurs m times, then the ratio m/n will differ only slightly from $\mathbb{P}\{A\}$.*
>
> *(B) If $\mathbb{P}\{A\}$ is very small, then one can be practically certain that the event A will not occur on a single realization of the conditions $\mathfrak{S}$.*

Kolmogorov concluded Section 2 with two remarks:

Remark I. *If two assertions are both practically certain, then the assertion that they are simultaneously correct is practically certain, though with a little lower degree of certainty. But if the number of assertions is very large, we cannot draw any conclusion whatsoever about the correctness of their simultaneous assertion from the practical certainty of each of them individually. So it in no way follows from Principle A that* m/n *will differ only a little from* $\mathbb{P}\{A\}$ *in every one of a very large number of series of experiments, where each series consists of* n *experiments.*

Remark II. *By our axioms, the impossible event (the empty set) has the probability* $\mathbb{P}\{\emptyset\} = 0$*. But the converse inference, from* $\mathbb{P}\{A\} = 0$ *to the impossibility of* A*, does not by any means follow. By Principle B, the event* A*'s having probability zero implies only that it is practically impossible that it will happen on a particular unrepeated realization of the conditions* $\mathfrak{S}$*. This by no means implies that the event* A *will not appear in the course of a sufficiently long series of experiments. When* $\mathbb{P}\{A\} = 0$ *and* n *is very large, we can only say, by Principle A,*

that the quotient m/n will be very small—it might, for example, be equal to $1/n$.

Cournot principle Kolmogorov's Principle B is also known as ***Cournot principle***. Cournot, a mathematician of the XIX Century, now remembered as an economist and a philosopher of science, was probably the first to to say explicitly that probability theory does gain empirical meaning only by declaring events of vanishingly small probability to be impossible. More precisely, what later came to be known as Cournot principle, is the statement that ***an event of very small probability singled out in advance will not happen***. This principle was advanced as the means by which a probability model is given empirical meaning.

Paul Lèvy was the first to make the point absolutely clear: ***Cournot's principle is the only connection between probability and the empirical world.*** He first said it clearly in his 1919 course. In a 1922 lecture, Hadamard made the point that the principle of the negligible event connects the mathematics with the real world. In

his book of 1940, Émile Borel was as clear as Lévy about Cournot's principle being the ***only*** link between probability and the world. He wrote: ***The principle that an event with very small probability will not happen is the only law of chance***.

By the 1970s, only Yuri Prokhorov, carried Kolmogorov's views, expressing the principle paradoxically in the Soviet Encyclopedia of Mathematics:

Only probabilities close to zero or one are meaningful.

Note that if the empirical meaning of probability resides precisely in the non-happening of small-probability events singled out in advance, then we need no additional principles to justify rejecting a hypothesis that gives small probability to an event we single out in advance and then observe to happen. In other words, the Cournot principle provides a justification to classical statistics, as formulated by Fisher, Neyman and Pearson, which is the "working statistics" of the overwhelming majority of scientists.

2 Conditional Probabilities and Independence

Independence is a the basic concepts of probability. According to Kolmogorov, it is exactly what distinguishes probability theory from measure theory. In order to characterize independence, we introduce first the notion of conditional probability.

2.1 Conditional probability

In a number of cases it is necessary to find the probability of events, given the supplementary condition that a certain event B has occurred that has a positive-probability. We shall call such probabilities ***conditional*** and denote them by the symbol $\mathbb{P}\{A|B\}$ which signifies the probability of event A on condition that event B has occurred. Strictly speaking, unconditional probabilities $\mathbb{P}\{A\}$ are also conditional probabilities, since for the starting point of the theory the existence of a certain invariable set of conditions $\mathfrak{S}$.

To find the formula for $\mathbb{P}\{A|B\}$, consider an experiment that is repeated n times and suppose that m times ($m \leq n$) the event B

occurred and in k of the experiments ($k \leq n$) the event A occurred. The frequencies of the events are given by $f_A = k/n$ and $f_B = m/n$, respectively. If the event $A \cap B$ occurred l times ($l \leq m$) , then the frequency $f_{A|B}$ at which A occurred, provided that B occurred, is l/m. Since

$$\frac{l}{m} = \frac{l/n}{m/n} = \frac{f_{A\cap B}}{f_B}$$

we have

$$f_{A|B} = \frac{f_{A\cap B}}{f_B}$$

This relation suggests to define the conditional probability $\mathbb{P}\{A|B\}$ in the following way. ***Let*** $\mathbb{P}\{B\} > 0$***. Then the conditional probability of*** A ***given*** B ***is defined as***

$$\mathbb{P}\{A|B\} = \frac{\mathbb{P}\{A \cap B\}}{\mathbb{P}\{B\}}\,. \tag{8}$$

This may be visualized as restricting the sample space to B. In other words, if the outcomes are restricted to B, this set serves as the new sample space. Note that the notion of conditional probability is introduced as a definition and not as an axiom.

2.2 Independent events

That two events A and B are independent (alternatively called statistically independent or stochastically independent) means that the occurrence of one does not affect the probability of the other. Equivalently, they are are independent if and only if their joint probability equals the product of their probabilities:

$$\mathbb{P}\{A \cap B\} = \mathbb{P}\{A\}\,\mathbb{P}\{B\} \quad \Leftrightarrow \quad \mathbb{P}\{A|B\} = \mathbb{P}\{A\}$$

$$\Leftrightarrow \mathbb{P}\{B|A\} = \mathbb{P}\{B\}$$

More generally, the events $A_1, A_2, \ldots$ are independent if if they are pairwise independent, that is, if the intersection (= joint occurrence) of any subset of them has as its probability the product of probabilities of the individual events. For more elaboration on the notion independence, see section 5.7.

2.3 Total probability and Bayes Theorem

A partition of a set is a division of the set as a union of non-overlapping and non-empty subsets. Let

$$\Omega = \bigcup_n B_n \quad \text{with } B_n \cap B_m = \emptyset\,,\ n \neq m \tag{9}$$

be a partition of the sample space Ω into disjoint sets. Then for any set A we have

$$\mathbb{P}\{A\} = \sum_n \mathbb{P}\{B_n\}\,\mathbb{P}\{A|B_n\}\ . \tag{10}$$

This formula is easily established by writing

$$A = \Omega \cap A = (\cup_n B_n) \cap A = \cup_n (B_n \cap A)\,.$$

By countable additivity, $\mathbb{P}\{A\} = \sum_n \mathbb{P}\{B_n \cap A\}$ and by writing $\mathbb{P}\{B_n \cap A\} = \mathbb{P}\{B_n|A\}$, one arrives at the desired result. The formula (10) is usuallly referred to as that of ***total probability***. Here is a standard

interpretation. Suppose that the event A may occur under a number of mutually exclusive circumstances (or "causes"). Then the formula shows how its total probability is compounded from the probabilities of the various circumstances, and the corresponding conditional probabilities figured under the respective hypotheses.

By noting that $\mathbb{P}\{B_n|A\} = \mathbb{P}\{B_n\}\,\mathbb{P}\{A|B_n\}/\mathbb{P}\{A\}$ and using eq. (10) to express $\mathbb{P}\{A\}$, we arrive at the formula

$$\mathbb{P}\{B_n|A\} = \frac{\mathbb{P}\{B_n\}\,\mathbb{P}\{A|B_n\}}{\sum_n \mathbb{P}\{B_n\}\,\mathbb{P}\{A|B_n\}}\,. \tag{11}$$

This simple proposition with a very easy proof is very famous under the name of ***Bayes Theorem***, published in 1763. It is supposed to yield an"inverse probability," or probability of the "cause" B, on the basis of the observed "effect" A. Whereas $\mathbb{P}\{B_n\}$ is the ***a priori***, $\mathbb{P}\{B_n|A\}$ is the ***a posteriori*** probability of the cause B_n. Laplace, one of the great mathematicians of all times, who wrote a monumental treatise on probability around 1815, used the theorem to estimate the probability that "the sun will also rise tomorrow." His analysis was a lot to be desired. In modern times Bayes lent his

name to a school of statistics quite popular among philosophers. Let us merely comment that Bayes has certainly hit upon a remarkable turn-around for conditional probabilities, but the practical utility of his formula is very limited by our usual lack of knowledge on the various a priori probabilities.

3 Physical examples of probability spaces

3.1 Classical systems

The complete microscopic state of an isolated classical system of N point particles is specified at any time t by a point X in its phase space Γ, with X representing the positions and the momenta of all the particles. Given an X at some time t_0, the micro-state at any other time $t \in \mathbb{R}$, X_t, is given (as long as the system stays isolated) by the evolution generated by the Hamiltonian $\mathsf{H}(x)$ of the system. Such an evolution takes place on an energy surface, Γ_E, specified by $\mathsf{H}(x) = E$. It is useful—especially when dealing with macroscopic systems—to actually think of Γ_E as an energy shell of thickness $\Delta E \ll E$, i.e., as the the subset of space phase

$$\Gamma_E = \{x \in \Gamma : E \leq \mathsf{H}(x) \leq E + \Delta E\}\,. \tag{12}$$

Suppose that the particles are confined in a box $\Lambda \subset \mathbb{R}^3$ of finite volume $V = |\Lambda|$. Then $|\Gamma_E|$ is finite (if the particles were not confined, the volume would be infinite). Then consider the probability

measure

$$\mathbb{P}_E\{A\} = \frac{|A \cap \Gamma_E|}{|\Gamma_E|}, \tag{13}$$

on the Borel sets A of Γ_E. It is the Lebesgue measure restricted to the shell of constant energy and normalized to the volume of the shell and is called the ***microcanonical measure***. (Eventually, one may pass to the limit $\Delta E \to 0$ of both sides of eq. (13) and obtain in this way the probability measure on the surface $\mathsf{H}(x) = E$.)

Thus, the probability space of a classical system of constant energy E (possibly,with a certain tolerance $\Delta E \ll E$) is $(\Gamma_E, \mathscr{B}(\Gamma_E), \mathbb{P}_E)$: the sample space is the shell of constant energy and the probability measure is the normalized Lebesgue measure on it. We call it the ***fundamendal probability space*** of a classical system of constant energy. In view of Liouville theorem,

$$\mathbb{P}_E\left\{T_t^{-1}(A)\right\} = \mathbb{P}_E\{A\}\ , \tag{14}$$

for all $A \in \mathscr{B}(\Gamma_E)$ and $t \in \mathbb{R}$. This is the basic condition of a ***measure preserving dynamical system***, which is defined as a proba-

bility space $(\Omega, \mathscr{F}, \mathbb{P})$ endowed with a continuous family of measure-preserving transformation T_t (i.e., fulfilling eq. (14) for $\mathbb{P}_E = \mathbb{P}$ and all $A \in \mathscr{F}$). The presence of an invariant measure means that the probabilities of the events do not change with time. This is of paramount importance in the analysis of the theory.

3.2 The Ising model

The Ising model is easy to define, but its behavior is incredibly rich. To begin with we need a lattice. For example we could take $\mathbb{Z}^d$, the set of points in $\mathbb{R}^d$ all of whose coordinates are integers. In two dimensions this is usually called the square lattice, in three the cubic lattice and in one dimension it is often referred to as a chain. We shall denote a site in the lattice by $\boldsymbol{i} = (i_1, \ldots, i_d)$. For each site $\boldsymbol{i}$ we have a variable which only takes on the values $\sigma_{\boldsymbol{i}} = +1$ and $\sigma_{\boldsymbol{i}} = -1$. Physically, we may think of the variable $\sigma_{\boldsymbol{i}}$ as representing the "spin" or the magnetic moment (by suitable choice of the units of measures) at the site $\boldsymbol{i}$.

Let L some (possibly large) number and let us consider the lattice

Λ_L of side L, that is, in the case of the lattice $\mathbb{Z}^d$, consider its subset

$$\Lambda_L = \{\boldsymbol{i} = (i_1, \ldots, i_d) : |i_k| \leq L, k = 1, \ldots, d\} \ . \tag{15}$$

The subset Λ_L is finite, so the number of possible spin configurations $\{\sigma_{\boldsymbol{i}}\}_{\boldsymbol{i}\in\Lambda_L}$ is finite. We will use σ as shorthand for one of these spins configurations, i.e., an assignment of $+1$ or -1 to each site. Their totality is the sample space Σ of the Ising model. If Λ_L contains N sites, Σ contains 2^N points.

We shall now define a probability measure on the set of configurations Σ. It will depend on an "energy function" or "Hamiltonian" H. The simplest choice for H is

$$H(\sigma) = -J \sum_{\boldsymbol{i},\boldsymbol{j}:|\boldsymbol{i}-\boldsymbol{j}|=1} \sigma_{\boldsymbol{i}}\sigma_{\boldsymbol{j}} \ , \tag{16}$$

where J is a (coupling) constant. The sum is over all pairs of sites in our finite volume which are nearest neighbors in the sense that the Euclidean distance between them is 1. The probability measure is given by

$$\mathbb{P}_\beta \{\sigma\} = Z^{-1} e^{-\beta H(\sigma)} \ . \tag{17}$$

where Z is a constant chosen to make this a probability measure. Explicitly,

$$Z = \sum_{\sigma} e^{-\beta H(\sigma)} \tag{18}$$

where the sum is over all the spin configurations on our finite volume. Mathematically, $0 \leq \beta < \infty$ is a parameter; physically, it represents the ***inverse temperature*** $\beta = 1/\kappa T$.

The Ising model was invented by Wilhelm Lenz in 1920 and it is named after Ernst Ising, a student of Lenz who chose the model as the subject of his doctoral dissertation in 1925. The model was devised as a tool to study the thermodynamic properties of a ferromagnet in thermal equilibrium.

The Ising model is the main tool we have to study phase transitions. Its value transcends its original use as a model of ferromagnetism. It has also been applied to problems in chemistry, biology and other areas where "cooperative behavior" of large systems is studied. These applications are possible because the Ising model is formulated as a well defined problem in mathematical probability.

4 Random Variables (I)

A basic concept of the theory of probability is the notion of ***random variable***. A random variable is a function that expresses numerical characteristics of sample points. In physics, where the sample space is identified with the set of microscopic states of a physical system, a random variable thus represents a physical quantity (infamously called an "observable"), e.g., the density of a gas as a function of its ***microstate***.

We shall begin by assuming that the sample space is a set with a finite or a most countable number of elements. The probability space of the Ising model is of this type. In order to understand the basic notions of probability theory and their relevance to physics, it is useful not to be distracted by mathematical complications and this is what the above assumption buys us.

4.1 Discrete Random Variables

Let the sample space Ω be a set with a countable number of elements ω. Assume that a non-negative number $p(\omega)$ is assigned to each point $\omega \in \Omega$ such that

$$\sum_{\omega \in \Omega} p(\omega) = 1$$

and define

$$\mathbb{P}\{A\} = \sum_{\omega \in A} p(\omega)$$

for every subset $A \subseteq \Omega$. Then, as it can easily be seen, $\mathbb{P}$ is a probability measure on $(\Omega, \mathcal{F})$ where $\mathcal{F}$ is the family of all the subsets of Ω.

A numerically valued function Y of ω with domain Ω,

$$\omega \to \mathsf{Y}(\omega) \tag{19}$$

is called a random variable (on Ω). It is a custom (not always observed) to use a capital letter to denote a random variable. The

adjective "random" is just to remind us that we are dealing with a sample space and trying to describe certain things called "random events." What might be said to have an element of randomness in $\mathsf{Y}(\omega)$ is the sample point ω which is picked "at random," such as in a throw of dice. Once ω is picked, $\mathsf{Y}(\omega)$ is thereby determined and there is nothing vague, indeterminate or chancy about it anymore.

Given a family of random variables, $\mathsf{Y}_1, \ldots, \mathsf{Y}_n$, other random variables can be formed by forming linear combinations, by multiplying them and more generally by considering functions of them. A particularly important case is the sum of n random variables,

$$\mathsf{S}_n = \mathsf{Y}_1 + \ldots + \mathsf{Y}_n \,. \tag{20}$$

4.2 Probability Distributions

As remarked above, random variables are defined on a sample space Ω before any probability is mentioned. They acquire their probability distributions through a probability measure imposed on the space. Consider the event $\{\omega \in \Omega : a \leq \mathsf{Y}(\omega) \leq b\}$ where a and b are to

constants. Given a probability measure $\mathbb{P}$ on Ω, every subset of Ω has a probability assigned to it when Ω is countable. The set above has a probability which will be denoted by $\mathbb{P}\{a \leq \mathsf{Y} \leq b\}$, with

$$\mathbb{P}\{a \leq \mathsf{Y} \leq b\} = \mathbb{P}\{\omega \in \Omega : a \leq \mathsf{Y}(\omega) \leq b\} \ . \tag{21}$$

More generally, let Δ be a subset of real numbers, then we can write

$$\mathbb{P}\{\mathsf{Y} \in \Delta\} = \mathbb{P}\{\omega \in \Omega : \mathsf{Y}(\omega) \in \Delta\} \tag{22}$$

An important case occurs when Δ reduces to a single point y. Then the probability

$$\mathbb{P}\{\mathsf{Y} = y\} = \mathbb{P}\{\omega \in \Omega : \mathsf{Y}(\omega) = y\} \equiv p(y) \tag{23}$$

is ***the probability that*** Y ***takes the value*** y***.***

Now the hypothesis that Ω is countable will play an essential simplifying role. It is clear that the range of Y must be finite when Ω is finite, and at most countably infinite when Ω is so, and many of these numbers may be the same. In fact, the mapping $\omega \to \mathsf{Y}(\omega)$ is in general many-to-one, not necessarily one-to-one. In the extreme

case when Y is a constant random variable, the its range reduces to a single number. Be that as it may, it should be obvious that if we know all the $p(y)$'s, then we can calculate all probabilities concerning the random variable Y, alone. These ***elementary probabilities*** provide the ***probability distribution*** of the values of the random variable Y. For example, eq. (22) becomes

$$\mathbb{P}\left\{\mathsf{Y} \in \Delta\right\} = \sum_{y \in \Delta} p(y)\,. \tag{24}$$

4.3 Random vectors

Often we need to consider several random variables $\mathsf{Y}_1, \ldots, \mathsf{Y}_n$ at the same time. Their ***joint probability distribution*** is given by

$$\begin{aligned}\mathbb{P}\left\{\mathsf{Y}_1 = y_1, \ldots, \mathsf{Y}_n = y_n\right\} &= \mathbb{P}\left\{\omega \in \Omega : \mathsf{Y}_1(\omega) = y_1, \ldots, \mathsf{Y}_n(\omega) = y_n\right\} \\ &\equiv p(y_1, \ldots, y_n)\end{aligned} \tag{25}$$

The vector whose components are the random variables of the family, $\mathsf{Y} = (\mathsf{Y}_1, \ldots, \mathsf{Y}_n)$, is called a random vector and eq. (25) is its probability distribution.

The ***marginal distribution*** of one of the variables, say Y_1 is

$$p_1(y_1) = \sum_{y_2,\ldots,y_n} p(y_1, \ldots, y_n) \tag{26}$$

The marginal distributions $p_2, \ldots p_n$ of the other variables are similarly defined. Let us observe that these marginal distributions do not in general determine the joint distribution.

4.4 Partition of sample space

A random variable generate a partition of the sample space Ω, that is, a decomposition of Ω into disjoint sets $\mathsf{Y}^{-1}(y) = \{\omega \in \Omega : \mathsf{Y}(\omega) = y\}$:

$$\Omega = \bigcup_y \mathsf{Y}^{-1}(y)\,. \tag{27}$$

The sample points in $\mathsf{Y}^{-1}(y)$ have all the same numerical characteristic, namely the value y of the random variable Y. From a physical point of view, this is clear: if y is the density of a gas, $\mathsf{Y}^{-1}(y)$

is the set of phase points that have the same density y, and the set of phase points having a different density has no phase points in common with it. Similar consideration apply if instead of a random variable one considers a random vector $\mathsf{Y} = (\mathsf{Y}_1, \ldots, \mathsf{Y}_n)$. We still have the decomposition (27) with $y = (y_1, \ldots, y_n)$.

4.5 Independent random variables

Two random variables X and Y are independent if the events they generate are independent; that is to say, the events $\mathsf{X} \in \Delta_1$ and $\mathsf{Y} \in \Delta_2$, for all subsets of real numbers Δ_1 and Δ_1. A necessary and sufficient condition for this is that their joint distribution factorizes,

$$\mathbb{P}\{\mathsf{X} = x, \mathsf{Y} = y\} = \mathbb{P}\{\mathsf{X} = y\}\,\mathbb{P}\{\mathsf{Y} = y\} \quad \text{i.e.,} \quad p(x, y) = p_{\mathsf{X}}(x) p_{\mathsf{Y}}(y)\,, \tag{28}$$

where p_{X} and p_{Y} are the probability distributions of X and Y

From eq. (28) it follows and important equation for the probability distribution of the sum $\mathsf{Z} = \mathsf{X} + \mathsf{Y}$ of two independent random variables. Given $\mathsf{X} = x$, we have $\mathsf{X} + \mathsf{Y} = z$ if and only if $\mathsf{Y} = z - x$, hence

by eq. (10):

$$\mathbb{P}\left\{\mathsf{X}+\mathsf{Y}=z\right\}=\sum_{x}\mathbb{P}\left\{\mathsf{X}=x\right\}\mathbb{P}\left\{\mathsf{Y}=z-x|\mathsf{X}=x\right\}=$$

$$=\sum_{x}\mathbb{P}\left\{\mathsf{X}=x\right\}\mathbb{P}\left\{\mathsf{X}+\mathsf{Y}=z-x\right\}\ ,$$

that is, the distribution of the sum is the (discrete) convolution of the distributions,

$$p_{\mathsf{X}+\mathsf{Y}}(z)=\sum_{x}p_{\mathsf{X}}(x)p_{\mathsf{Y}}(z-x)\,. \tag{29}$$

More generally, the random variables $\mathsf{Y}_1,\ldots,\mathsf{Y}_n$ are independent if every pair of random variables is independent. This is equivalent to the factorability of their joint distribution:

$$p(y_1,y_2\ldots,y_n)=p_1(y_1)p_2(y_2)\cdots p_n(y_n) \tag{30}$$

One of main area of the research in probability has been the study of the study of the asymptotic behavior of the sums of independent

identically distributed random variables (abbreviated "i.i.d. random variables). If $\mathsf{Y}_1, \ldots, \mathsf{Y}_n$ are i.i.d. random variables with common distribution p, then

$$p(y_1, y_2 \ldots, y_n) = p(y_1)p(y_2) \cdots p(y_n) \tag{31}$$

4.6 Mean value

The values taken by a random variable Y can vary considerably. It is useful to isolate some salient characteristics of this variability, such as the average value and the way in which the values spread out about this average value. The first characteristic is called the ***mean value*** or ***expected value***. It is defined as

$$\mathbb{E}\{\mathsf{Y}\} = \sum_{\omega \in \Omega} \mathsf{Y}(\omega)\mathbb{P}\{\omega\} \tag{32}$$

provided that the series converges absolutely (in case Ω is countable and not merely finite), namely

$$\sum_{\omega \in \Omega} |\mathsf{Y}(\omega)|\mathbb{P}\{\omega\} < \infty$$

In this case we say that the mathematical expectation of Y exists or that Y is ***summable***. The process of "taking expectations"' may be described in words as follows: take the value of Y at each ω, multiply it by the probability of that point, and sum over all ω in Ω. If we think of $\mathbb{P}\{\omega\}$ as the weight attached to ω then $\mathbb{E}\{\mathsf{Y}\}$ is the weighted average of the function Y.

Let us state explicitly a general method of calculating $\mathbb{E}\{\mathsf{Y}\}$ which is often useful. Let us group together all the sample points that have the same value y. Then, by observing (27), we obtain

$$\mathbb{E}\{\mathsf{Y}\} = \sum_y y\mathbb{P}\left\{\mathsf{Y}^{-1}(y)\right\} = \sum_y yp(y)\,. \tag{33}$$

Let now highlight some basic properties of the expectation. If X and Y are summable, then $\mathsf{X}+\mathsf{Y}$ is summable too and by linearity we obtain

$$\mathbb{E}\{a\mathsf{X}+b\mathsf{Y}\} = a\mathbb{E}\{\mathsf{X}\} + b\mathbb{E}\{\mathsf{Y}\}\,, \tag{34}$$

where a and b are constants. In other words, the expectation is a linear functional. This is really a general result which is widely used in many branches of mathematics. In contrast, the following

fact requires independence and is special to probability theory: If X and Y are ***independent*** summable random variables, then

$$\mathbb{E}\{XY\} = \mathbb{E}\{X\}\,\mathbb{E}\{Y\}\,. \tag{35}$$

This is an immediate consequence of the factorability (30) of the joint distribution of X and Y. The extension to any finite number of independent summable random variables is immediate.

4.7 Moments

For positive integer r, the mathematical expectation $\mathbb{E}\{Y^r\}$ is called the r-th moment (or moment of order r) of Y. Thus if Y^r has a finite expectation, we say that Y has a finite r-th moment. For $r = 1$ of course the first moment is just the expectation or mean. The case $r = 2$ is of special importance. Since $Y^2 > 0$, we shall call $\mathbb{E}\{Y^2\}$ the second moment of Y whether it is finite or equal to $+\infty$ according as the defining series (in a countable Ω) $\sum_\omega Y(\omega)^2 \mathbb{P}\{\omega\}$ converges or diverges.

When the mean $\mathbb{E}\{\mathsf{Y}\}$ is finite, it is often convenient to consider

$$\mathsf{Y}^o = Y - \mathbb{E}\{\mathsf{Y}\} \tag{36}$$

instead of Y because its first moment is equal to zero. We shall say theat Y^o is obtained from X by ***centering***.

4.8 Variance and Standard Deviation

The second moment of Y^o is called the variance of Y and denoted by $\mathrm{Var}\,(\mathsf{Y})$; its positive square root is called the ***em standard deviation***. It is a measure of the spread about the mean of the values taken by Y. We have

$$\mathrm{Var}\,(\mathsf{Y}) = \mathbb{E}\{(\mathsf{Y} - \mathbb{E}[\mathsf{Y}])^2\} = \mathbb{E}\{\mathsf{Y}^2\} - 2\mathbb{E}[\mathsf{Y}]\mathbb{E}\{Y\} - \mathbb{E}[\mathsf{Y}]^2$$

$$= \mathbb{E}\{\mathsf{Y}^2\} - \mathbb{E}[\mathsf{Y}]^2\,,$$

The variance is consequently always nonnegative. In particular, we have $\mathrm{Var}\,(|\mathsf{Y}|) = \mathbb{E}\{\mathsf{Y}^2\} - \mathbb{E}\{|\mathsf{Y}|^2\} \geq 0$. Thus, if if the second moment

of Y is finite, Y is summable. The general method for calculating the variance is obtained from eq. (33) applied to $(\mathsf{Y} - \mathbb{E}\{Y\})^2$:

$$\mathrm{Var}\,(\mathsf{Y}) = \sum_y (y - \mathbb{E}\{Y\})^2 p(y) \tag{37}$$

Observe that Y and $\mathsf{Y} + a$ have the same variance for any constant a; in particular, this is the case for Y and Y^o. Moreover, $\mathrm{Var}\,(b\mathsf{Y}) = b^2\mathrm{Var}\,(\mathsf{Y})$ for any constant b. So, we have

$$\mathrm{Var}\,(a + b\mathsf{Y}) = b^2\mathrm{Var}\,(\mathsf{Y})\ . \tag{38}$$

for constants a an b.

The following result holds.

If X ***and*** Y ***are independent and both have finite variances, then***

$$\boldsymbol{Var}\,(\mathsf{X} + \mathsf{Y}) = \boldsymbol{Var}\,(\mathsf{X}) + \boldsymbol{Var}\,(\mathsf{Y})\ . \tag{39}$$

This is easily seen. By the preceding remark that Y and Y^o have the same variance, we may suppose that X and Y both have mean

zero. Then $X + Y$ also has mean zero and the variances in eq. (39) are the same as second moments. Now, $\mathbb{E}\{XY\} = \mathbb{E}\{X\}\mathbb{E}\{Y\} = 0$ by eq. (35), and

$$\mathbb{E}\{(X+Y)^2\} = \mathbb{E}\{X^2 + 2XY + Y^2\} = \mathbb{E}\{X^2\} + 2\mathbb{E}\{XY\} + \mathbb{E}\{Y^2\}$$

$$= \mathbb{E}\{X^2\} + \mathbb{E}\{Y^2\}$$

and this is the desired result. □

The extension of this result to any finite number of independent random variables is immediate:

$$\mathrm{Var}\,(Y_1 + \ldots + Y_n) = \mathrm{Var}\,(Y_1) + \ldots + \mathrm{Var}\,(Y_n)\ . \tag{40}$$

Notations Common notation for the mean of a random variable Y are also m_Y or $\langle Y \rangle$. The standard deviation of Y is often denoted σ_Y or ΔY.

5 General Random Variables

In the preceding sections we have given a quite rigorous discussion of random variables which take only a countable set of values. But even at an elementary level there are many important questions in which we must consider random variables not subject to such a restriction. This means that we need a sample space which is not countable. Technical questions of "measurability" then arise which cannot be treated satisfactorily without more advanced mathematics. This kind of difficulty stems from the impossibility of assigning a probability to every subset of the sample space when it is uncountable. The matter is resolved by confining the probability assignments to sample sets belonging to an adequate class called a Borel algebra.

Without going into this here we will take up a particular but very important situation which covers most applications and requires little mathematical abstraction. This is the case of random variable with a "density."

5.1 Absolutely continuous random variables

Consider a non negative (piecewise) continuos function ρ defined on $\mathbb{R}$ such that

$$\int_{-\infty}^{\infty} \rho(y)dy = 1\,. \tag{41}$$

Such a function is called a ***density function*** on $\mathbb{R}$. We can now define a class of random variables on an ***arbitrary*** sample space as follows. As in eq. (19) Y is a function on Ω, but its probabilities are prescribed by means of a density function ρ so that for any interval $[a, b]$ we have

$$\mathbb{P}\left\{a < \mathsf{Y} \leq b\right\} = \int_a^b \rho(y)dy\,. \tag{42}$$

Such a random variable is said to ***have a density***, and its density function is ρ. Equivalently, they are called ***absolutely continuous***.

More generally, for a random variable of this kind, if Δ is the union of intervals not necessarily disjoint and some of which may

be infinite, we have

$$\mathbb{P}\{\mathsf{Y} \in \Delta\} = \int_{\Delta} \rho(y)dy\,. \tag{43}$$

In particular, if $\Delta = (-\infty, y]$, then we can write

$$F(y) = \mathbb{P}\{\mathsf{Y} \leq y\} = \int_{-\infty}^{y} \rho(x)dx\,. \tag{44}$$

This formula defines the ***cumulative distribution function*** of Y. If ρ is continuos, from the fundamental theorem of calculus we have that ρ is the derivative of F.

All the notions introduced for discrete random variables translate into the corresponding ones for absolutely continuous random variables by means of the "dictionary"

$$\sum_{y} \longrightarrow \int_{-\infty}^{\infty} dy \qquad \text{and} \qquad p(y) \longrightarrow \rho(y) \tag{45}$$

For example, the condition of summable random variable now be-

come that of integrable random variable:

$$\int_{-\infty}^{\infty} |y|\rho(y)dy < \infty \tag{46}$$

the definitions of mean becomes

$$\mathbb{E}\{Y\} = \int_{-\infty}^{\infty} y\rho(y)dy \tag{47}$$

and so on. In particular, note that the density of the sum of two independent random variables is now the usual convolution

$$\rho_{\mathsf{X}+\mathsf{Y}}(z) = \int_{-\infty}^{\infty} \rho_{\mathsf{X}}(x)\rho_{\mathsf{Y}}(z-x)\,dx\,. \tag{48}$$

and the joint density of i.i.d random variables $\mathsf{Y}_1, \ldots, \mathsf{Y}_n$ is

$$\rho(y_1, y_2 \ldots, y_n) = \rho(y_1)\rho(y_2)\cdots\rho(y_n)\,, \tag{49}$$

where ρ is the common density.

5.2 General random variables

The most general random variable is a function Y defined on the sample space Ω such that for any real y, the probability $\mathbb{P}\{\mathsf{Y} < y\}$ is defined. More precisely, given the probability space $(\Omega, \mathscr{F}, \mathbb{P})$, a function $\mathsf{Y} : \Omega \to \mathbb{R}$ is a random variable if the event $\{\mathsf{Y} < y\} \in \mathscr{F}$ for every $y \in \mathbb{R}$.

Next, we define for every real y

$$F(x) = \mathbb{P}\{\mathsf{Y} \leq y\} \tag{50}$$

or equivalently for $a < b$

$$F(b) - F(a) = \mathbb{P}\{a < \mathsf{Y} \leq b\} \tag{51}$$

and call F the ***cumulative distribution function*** of Y. From its very definition, F is a real valued, monotone non-decreasing function, defined for $-\infty \leq x \leq +\infty$, continuos from the right, which has limits 0 and 1 at $-\infty$ and $+\infty$ respectively. Monotonicity holds because if $y \leq y'$ then $\{\mathsf{Y} \leq y\} \subset \{\mathsf{Y} \leq y'\}$. The existence of the limits 0 and 1 at $-\infty$ and $+\infty$ is intuitively obvious because the event

$\{\mathsf{Y} < y\}$ becomes impossible as $y \to -\infty$ and certain as $y \to +\infty$. The rigorous proofs are however more sophisticated and depend on the countable additivity of $\mathbb{P}$; similarly, the proof of right continuity.

It can be easily checked that these properties are satisfied for the the cases considered so far:

$$F(y) = \sum_{x \leq y} p(x) \quad \text{and} \quad F(y) = \int_{-\infty}^{y} \rho(x)dx\,. \tag{52}$$

As a matter of fact, the general F turns out to be a mixture of these two kinds together with a more weird kind which correspond to ***singularly continuous*** random variables. These are variables which are neither discrete nor continuous in any interval. In this group of random variables are all those functions whose distributions are continuous but which only increase on a set of Lebesgue measure zero (an example of such a random variable is the quantity having the well-known Cantor curve as its distribution function). We shall not go into this but just note that we can operate quite well with the F as defined by eq. (50) without further specification. We finally note that a ***degenerate distribution*** or deterministic distribution

is the probability distribution of a random variable which only takes a single value. A random variable so distributed is called ***degenerate*** or ***deterministic***.

5.3 Generalized densities

The distribution function $F(y)$ of a discrete random variable is constant except for a countable number of discontinuities y_n, where $F(y_n) = F(y_n - 0) + p_n$.

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:

$$\rho(y) = \frac{dF(y)}{dy} = \sum_n p_k \delta(y - y_n) \tag{53}$$

So, by allowing densities to be defined by combination of continuos functions and delta functions, we are able to express with same notation both discrete and absolutely continuos random variables as well as mixture of them.

5.4 Location-scale transformations

The linear transformation

$$\mathsf{Z} = a + b\mathsf{Y}\,, \tag{54}$$

with $a \in \mathbb{R}$ and $b \in (0, \infty)$, if often useful in the applications and is called a ***location-scale transformation***; a is called the ***location parameter*** and b the ***scale parameter***. When two variables Z and Y are related by such a transformation they are said of the ***same type***. Clearly, $\mathbb{E}\{\mathsf{Z}\} = a + b\mathbb{E}\{\mathsf{Y}\}$ and the relations among the the variances is given by eq. (38). It can be easily checked that

$$F_{\mathsf{Z}}(y) = F_{\mathsf{Y}}\left(\frac{y-a}{b}\right)\,, \quad \text{equivalently, for the density,}$$
$$\rho_{\mathsf{Z}}(y) = \frac{1}{b}\rho_{\mathsf{Y}}\left(\frac{y-a}{b}\right)\,. \tag{55}$$

To be of the same type is an equivalence relation among random variables. Particularly useful is the transformation which gives the

standardization

$$\mathsf{Y}^* = \frac{\mathsf{Y} - m_{\mathsf{Y}}}{\sigma_Y} \tag{56}$$

of a (square summable) random variable Y.

5.5 Covariance

For a pair of random variables X and Y, their covariance is defined as

$$\mathsf{Cov}(\mathsf{X},\mathsf{Y}) = \mathbb{E}\{(\mathsf{X} - m_{\mathsf{X}})(\mathsf{Y} - m_{\mathsf{Y}})\} = \mathbb{E}\{\mathsf{XY}\} - m_{\mathsf{X}} m_{\mathsf{Y}}\,. \tag{57}$$

The covariance is a measure of how much two random variables are dependent on each other (if they are independent, their covariance is zero).

The covariance of the random vector $\mathsf{Y} = (\mathsf{Y}_1, \ldots, \mathsf{Y}_n)$ is defined as the $n \times n$ matrix C_{Y} whose (i,j)-entry is $\mathsf{Cov}(\mathsf{Y}_i, \mathsf{Y}_j)$. It can be shown that this is equivalent to the following geometrical definition

$$\mathsf{C}_{\mathsf{Y}} = \mathbb{E}\{|\mathsf{Y} - m_{\mathsf{Y}}\rangle\langle \mathsf{Y} - m_{\mathsf{Y}}|\} \tag{58}$$

where $|\mathsf{Y} - m_{\mathsf{Y}}\rangle\langle \mathsf{Y} - m_{\mathsf{Y}}|$ is the projector in $\mathbb{R}^n$ onto the vector $\mathsf{Y} - m_{\mathsf{Y}}$ (bra-ket notation). Moreover, it can be proven the following proposition: ***The covariance matrix corresponding to any random vector*** Y ***is symmetric positive semidefinite***.

The covariance enters in the computation of the variance of the sum of two random variables:

$$\begin{aligned} \mathrm{Var}\,(\mathsf{X}+\mathsf{Y}) &= \langle (\mathsf{X} - m_{\mathsf{X}}) + (\mathsf{Y} - m_{\mathsf{Y}}), (\mathsf{X} - m_{\mathsf{X}}) + (\mathsf{Y} - m_{\mathsf{Y}})\rangle \\ &= \|\mathsf{X} - m_{\mathsf{X}}\|^2 + 2\langle \mathsf{X} - m_{\mathsf{X}}, \mathsf{Y} - m_{\mathsf{Y}}\rangle + \|\mathsf{Y} - m_{\mathsf{Y}}\|^2 \\ &= \mathrm{Var}\,(\mathsf{X}) + 2\mathsf{Cov}(\mathsf{X},\mathsf{Y}) + \mathrm{Var}\,(\mathsf{Y})\ . \end{aligned} \tag{59}$$

5.6 Uncorrelated random variables

Two random variables X e Y are said ***uncorrelated*** if $\mathsf{Cov}(\mathsf{X},\mathsf{Y}) = 0$. This is weaker than independence. Correlation is only a measure of linear dependence. It does not necessarily imply anything about other kinds of dependence. Consider for example $\mathsf{X} \sim N(0,\sigma)$ and $\mathsf{Y} = \mathsf{X}^2$. The two variables are clearly dependent, however,

$$\mathsf{Cov}(\mathsf{X},\mathsf{Y}) = \mathbb{E}\{\mathsf{XY}\} - \mathbb{E}\{\mathsf{X}\}\mathbb{E}\{\mathsf{Y}\} = \mathbb{E}\{\mathsf{X}^3\} - 0 = 0\,.$$

The random variables $\mathsf{Y}_1, \ldots, \mathsf{Y}_n$ are said uncorrelated if their are pairwise uncorrelated.

5.7 Geometrical meaning of variance and covariance

The random variables with $\mathbb{E}\{\mathsf{Y}^2\} < \infty$ form the Hilbert space $L^2(\Omega, \mathscr{F}, \mathbb{P})$ with scalar product

$$\langle \mathsf{X}, \mathsf{Y} \rangle = \int \overline{\mathsf{X}(\omega)} \mathsf{Y}(\omega) \mathbb{P}(d\omega) \,. \tag{60}$$

The Hilbert space structure provides a nice geometrical framework for various probabilistic notions. In particular, we have:

- The variance can be expressed in terms of this scalar product and the associated L^2 norm as follows:

$$\mathrm{Var}\,(\mathsf{Y}) = \langle \mathsf{Y} - m_{\mathsf{Y}}, \mathsf{Y} - m_{\mathsf{Y}} \rangle = \|\mathsf{Y} - m_{\mathsf{Y}}\|^2 \,. \tag{61}$$

- The covariance of two variables is the scalar product of the corresponding centered variables

$$\mathrm{Cov}(\mathsf{X}, \mathsf{Y}) = \langle \mathsf{X}^o, \mathsf{Y}^o \rangle = \langle \mathsf{X} - m_{\mathsf{X}}, \mathsf{Y} - m_{\mathsf{Y}} \rangle \,. \tag{62}$$

- Two variables are uncorrelated, i.e. their covariance is zero, if and only if the corresponding centered variables are orthogonal in the L^2 sense.

- The random variables $\mathsf{Y}_1, \ldots, \mathsf{Y}_n$ are uncorrelated if and only if the corresponding centered variables $\mathsf{Y}_k - m_{\mathsf{Y}_k}$ form a family of orthogonal vectors in $L^2(\Omega, \mathscr{F}, \mathbb{P})$. As consequence of Pitagora's theorem, the variance of their sum $\mathsf{S}_n = \mathsf{Y}_1 + \ldots + \mathsf{Y}_n$ is the sum of the variances.

6 Generating Functions

Thee notion of "generating function" is a very useful mathematical device invented by the great mathematician Euler to study the partition problem in number theory. In probability theory, can be introduced as follows. Suppose that Y is a random variable that takes only nonnegative integer values with the probability distribution given by $\mathbb{P}\{\mathsf{Y} = k\} = p_k$. The idea is to put all the information contained above in a compact capsule. For this purpose a dummy variable z (real or complex) is introduced and the following power series in z set up:

$$g(z) = \sum_{n=0}^{\infty} p_n z^n \,. \tag{63}$$

This is called the generating function associated with the random variable Y. Remembering that $\sum_n p_n = 1$, it is easy to see that the power series in eq. (63) converges for $|z| \leq 1$. If we now differentiate the series term by term to get the derivatives of g, so long as we restrict its domain of validity to $|z| \leq 1$, we obtain $p_n = g^{(n)}(0)/n!$. This

shows that we can recover all the p_ns from g. Therefore, not only does the probability distribution determine the generating function, but also vice versa.

The generating function leads to extensions. This is easily seen when it is expressed as

$$g(z) = \mathbb{E}\{z^{\mathsf{Y}}\}\,. \tag{64}$$

If Y can take arbitrary real values this expression still has a meaning. The generating function is very useful for studying the sum of independent variables. ***If*** X ***and*** Y ***are independent random variables, then***

$$g_{\mathsf{X}+\mathsf{Y}}(z) = g_{\mathsf{X}}(z)g_{\mathsf{Y}}(z)\,. \tag{65}$$

This follows immediately from eq. (35) and eq. (64): $\mathbb{E}\{z^{\mathsf{X}+\mathsf{Y}}\} = \mathbb{E}\{z^{\mathsf{X}}e^{\mathsf{Y}}\} = \mathbb{E}\{z^{\mathsf{X}}\}\mathbb{E}\{z^{\mathsf{Y}}\}$. That is, the generating function of the sum of independent variables is the product of the generating functions of the summands, and products are much easier to handle than the convolution products.

6.1 Moment generating function

Let us consider only $0 \leq z \leq 1$. Every such z can be represented as $e^{-\lambda}$ with $0 \leq \lambda < \infty$, in fact the correspondence $z = e^{-\lambda}$ one-to-one. if Y takes the values y_i with probabilities p_i, then $\mathbb{E}\{e^{-\lambda \mathsf{Y}}\} = \sum_j p_j e^{-\lambda y_j}$ provided that the series converges absolutely. This is the case if all the values $y_i \geq 0$. Moreover, if Y has the density function ρ, then

$$\mathbb{E}\{e^{-\lambda \mathsf{Y}}\} = \int e^{-\lambda y} \rho(y) dy = \tilde{\rho}(\lambda)\,, \tag{66}$$

provided that the integral converges. This is the case if $\rho(y) = 0$ for $y \leq 0$ namely when Y does not take negative values (or more generally when is bounded below). If ρ is meant in the general sense of section 5.3, eq. (66) includes both discrete and continuos random variables.

We have therefore extended the notion of a generating function through to a large class of random variables. This new tool is called the ***Laplace transform*** of Y, which in probability is written with the "wrong" sign and defined as

$$M(k) = \tilde{\rho}(-k) = \mathbb{E}\{e^{k \mathsf{Y}}\}\,. \tag{67}$$

M is called the ***moment-generating function*** (MGF) of the random variable Y. It can be shown that M is always real analytic when it exists for all $k \in \mathbb{R}$. Thus, from the Taylor expansion,

$$M(k) = 1 + k\mathbb{E}\{\mathsf{Y}\} + \frac{k^2\mathbb{E}\{\mathsf{Y}^2\}}{2!} + \ldots + \frac{k^3\mathbb{E}\{\mathsf{Y}^3\}}{3!} + \ldots$$

one reads the moments $m_n = \mathbb{E}\{\mathsf{Y}^n\}$ of the random variable:

$$m_n = \left.\frac{\partial^n M}{\partial k^n}\right|_{k=0} \tag{68}$$

6.2 Characteristic function

A problem with the moment-generating functions is that moments and the moment-generating function may not exist, as the integrals need not converge absolutely. However, if we replace the negative real parameter $-\lambda$ in eq. (66) by the purely imaginary ik, where $i = \sqrt{-1}$ and k is real, we get the ***Fourier transform***

$$\hat{\rho}(k) = \mathbb{E}\{e^{ik\mathsf{Y}}\}\,. \tag{69}$$

In probability theory it is also known as the ***characteristic function*** (CF) of Y. The function $\hat{\rho}(k)$ is always defined, in fact $\hat{\rho}(k) \leq 1$ for all k. Herein lies the superiority of this new transform over the others discussed above, which cannot be defined sometimes because the associated series or integral does not converge. Given $\hat{\rho}$, one recovers the distribution of Y, by the usual Fourier inversion formula

$$\rho(y) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{\rho}(k) e^{-iky} dy \,. \tag{70}$$

Here are some basic properties of the characteristic function that follow from its definition.

(a) $\hat{\rho}(0) = 1$, $|\hat{\rho}(k)| \leq 1$ for all k; and $\hat{\rho}(k) \to 0$ as $k \to \infty$;

(b) Under a location-scale transformation (54) the characteristic function changes as

$$\hat{\rho}_{a+b\mathsf{Y}}(k) = e^{ika} \hat{\rho}_{\mathsf{Y}}(bk) \,; \tag{71}$$

(c) $\overline{\hat{\rho}(k)} = \hat{\rho}(-k)$ (where the bar denotes complex conjugation). If Y has distribution symmetric around zero, the its characteristic function is real valued.

(d) If $\mathbb{E}\{\mathsf{Y}^n\} < \infty$, $n \geq 1$, then there exists the continuous n-th derivative of the characteristic function and

$$\hat{\rho}^{(n)}(0) = i^n \mathbb{E}\{\mathsf{Y}^n\}\,. \tag{72}$$

In particular if $\mathrm{Var}\,(\mathsf{Y}) = 1$ and $\mathbb{E}\{Y\} = 0$, the characteristic function is

$$\hat{\rho}(k) = 1 - \frac{k^2}{2} + o(k^2)\,, \quad k \to 0 \tag{73}$$

where $o(k^2)$ is some function of k that goes to zero more rapidly than k^2.

(e) If a random variable has a moment-generating function, then the domain of the characteristic function can be extended to the complex plane, and

$$\hat{\rho}(k) = M(ik)\,. \tag{74}$$

7 Important Random Variables

The probability distribution of a random variable is also called its ***probability law*** or simply its ***law***. The word ***law*** is particularly appropriate since among all positive integrable functions only a very limited class of distributions governs the random variables which describe natural phenomena. To find out such laws and to explain their origin is a problem that has accompanied most of the research in mathematical probability since its inception.

7.1 Common Probability Distributions

Bernoulli distribution It is the distribution of a random variable which takes value 1 with success probability p and value 0 with failure probability $q = 1 - p$. It can be used, for example, to represent the toss of a coin, where "1" is defined to mean "heads" and "1" is defined to mean "tails" (or vice versa). It is denoted by Bernoulli(p). Mean

and variance of a random variable Y so distributed are

$$\mathbb{E}[\mathsf{Y}] = 1 \cdot p + 0 \cdot q = p, \quad \mathbf{Var}\,(\mathsf{Y}) = p(1-p) = pq\,. \tag{75}$$

Uniform distribution In an interval $[a, b]$, the uniform distribution is given by $\rho(y) = 1/(b-a)$ for $y \in [a, b]$ and equal to 0 otherwise. Mean and variance are

$$\mathbb{E}[\mathsf{Y}] = \frac{1}{12}(a+b), \quad \mathbf{Var}\,(\mathsf{Y}) = \frac{1}{12}(b-a)^2\,. \tag{76}$$

Poisson distribution

$$\mathbb{P}\left\{\mathsf{Y} = n\right\} = p_n = \frac{\lambda^n}{n!}e^{-\lambda}\,, \quad \lambda > 0 \quad \textbf{(intensity)}\,. \tag{77}$$

For example, in radioactive decay the number of atoms decaying per unit time is such a random variable taking values $0, 1, 2, \ldots$ without any upper bound. In general, it is the discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time and/or space if these events occur

with a known average rate and independently of the time since the last event. Mean and variance of a random variable Y so distributed are

$$\mathbb{E}[\mathsf{Y}] = \sum_0^\infty n \frac{\lambda^n}{n!} e^{-\lambda} = \lambda\,, \quad \mathrm{Var}\,(\mathsf{Y}) = \sum_0^\infty n^2 \frac{\lambda^n}{n!} e^{-\lambda} - \lambda^2 = \lambda\,. \tag{78}$$

The Poisson distribution can also be used for the number of events in other specified intervals such as distance, area or volume.

Normal or Gaussian distribution It is characterized by two parameters $m \in \mathbb{R}$ and $\sigma > 0$. It is usually denoted by $N(m, \sigma)$ and it is given by

$$\rho(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-m^2)}{2\sigma^2}}\,. \tag{79}$$

Mean and variance are

$$\mathbb{E}[\mathsf{Y}] = m\,, \quad \mathrm{Var}\,(\mathsf{Y}) = \sigma^2\,. \tag{80}$$

Gaussian random variables occur so commonly in applications, for example as measurement errors in laboratory experiments, that

they are often said to be normally distributed. Their ubiquity is explained by the fundamental theorem of probability and statistics, the ***Central Limit Theorem***.

Exponential distribution

$$\rho(y) = \lambda e^{-\lambda y} \tag{81}$$

and is the law that governs the time between events in a Poisson process, i.e. a process in which events occur continuously and independently at a constant average rate. Its mean and variance are easily computed:

$$\mathbb{E}[\mathsf{Y}] = \frac{1}{\lambda}\,, \quad \mathbf{Var}\,(\mathsf{Y}) = \frac{1}{\lambda^2}\,. \tag{82}$$

Cauchy distribution

$$\rho(y) = \frac{a}{\pi(a^2 + y^2)}\,, \tag{83}$$

where a is a positive constant (usually, $a = 1$). It arises, e.g., in the following situation. At some point P, let an emitter of particles

be placed, and at the distance a away from it, let a screen be installed. The particles are emitted in a plane which is perpendicular to the screen, and the angle ϕ between the plane and the normal to the screen is a random variable which is uniformly distributed on $(-\pi/2, \pi/2)$. The random coordinate $\mathsf{Y} = a \tan \phi$ on the screen, assuming that the particles fly along straight lines, is the Cauchy distribution. The Cauchy distribution is a symmetrical"bell- shaped" function like the Gaussian distribution, it differs from that in the behavior of their tails: the tails of the Cauchy density decrease as y^{-2}: the mean does not exist and the variance is infinite.

See table 1 for a list of the CFs and MGFs of the most common distributions.

7.2 Sums of i.i.d. random variables

The most important class of random variables is that of the sums $\mathsf{S}_n = \mathsf{Y}_1 + \ldots + \mathsf{Y}_n$ of i.i.d random variables. Denote by Y be a random variable that has the same distribution ρ of the variables. Since the joint distribution is $\rho(y_1, \ldots, y_n) = \rho(y_1) \cdots \rho(y_n)$, the distribution ρ_{S_n} of

S_n is

$$\rho_{\mathsf{S}_n} = \rho^{\star n}(y) \tag{84}$$

where $\rho^{\star n}$ is the n-th power of the convolution product. This follows by iteration from eq. (48). Mean, variance and generating function follow immediately from eq. (34), eq. (39) and eq. (65), respectively:

$$\mathbb{E}\{\mathsf{S}_n\} = nm\,, \qquad \mathrm{Var}\,(\mathsf{S}_n) = n\sigma^2\,, \qquad g_{\mathsf{S}_n} = g(z)^n\,, \tag{85}$$

where m, σ^2, and $g(z)$ are mean, variance, and generating function of Y. By rescaling, the corresponding equations for the normalized sum or ***sample mean*** $\bar{\mathsf{Y}}_n = \mathsf{S}_n/n$, follow immediately:

$$\mathbb{E}\{\bar{\mathsf{Y}}_n\} = m\,, \qquad \mathrm{Var}\left(\bar{\mathsf{Y}}_n\right) = \frac{\sigma^2}{n}\,, \qquad g_{\bar{\mathsf{Y}}_n} = g(z^{-n})^n\,, \tag{86}$$

Finally, two examples:

Binomial distribution Let $\mathsf{S}_n = \mathsf{B}_1 + \ldots + \mathsf{B}_n$, where each variable of the sum is Bernoulli(p) distributed. Then by simple combinatorics

$$\mathbb{P}\left\{\mathsf{S}_n = r\right\} = \binom{n}{r} p^r (1-p)^{n-r}\,, \quad \text{with} \quad \binom{n}{r} = \frac{n!}{j!(n-r)!}\,. \tag{87}$$

This is the ***binomial distribution*** $b(n, r, p)$. From additivity of mean and variance,

$$\mathbb{E}\{\mathsf{S}_n\} = np\,, \qquad \mathrm{Var}\,(\mathsf{S}_n) = npq\,, \qquad q \equiv (1-p)\,. \tag{88}$$

Note that Bernoulli(p) = $b(1, r, p)$.

Sum of n normal distributed variables Let $\mathsf{S}_n = \mathsf{Y}_1 + \ldots + \mathsf{Y}_n$, where each variable of the sum is $N(m, \sigma)$ distributed. Then the characteristic function S_n is

$$\hat{\rho}(k)^n = e^{imnk} e^{-\frac{1}{2}\sigma^2 n k^2}\,.$$

Thus S_n is still gaussian with distribution $N(nm, \sqrt{n}\sigma)$.

RV	$\rho(y)$	$M(k)$	$\hat{\rho}(k)$
Ber(p)	$p\delta(y-1)+q\delta(0)$	$pe^{k}+q$	$pe^{ik}+q$
Pois(λ)	$\sum_{n=0}^{\infty}\frac{\lambda^n}{n!}e^{-\lambda}\delta(y-n)$	$e^{\lambda(e^k-1)}$	$e^{\lambda(e^{ik}-1)}$
$N(m,\sigma)$	$\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(y-m^2)}{2\sigma^2}}$	$e^{km}e^{\frac{1}{2}\sigma^2k^2}$	$e^{ikm}e^{-\frac{1}{2}\sigma^2k^2}$
Exp(λ)	$\lambda e^{-\lambda y}$	$\frac{\lambda}{\lambda-k}$ **for** $k<\lambda$	$\frac{\lambda}{\lambda-ik}$
Cauchy	$\frac{a}{\pi(a^2+y^2)}$	∞ **for** $k\neq 0$	$e^{-a\lvert k\rvert}$