Introduction to

Bayesian Probability

and its use in

Theoretical Physics

John Hemp

Back Page

Bayesian probability, has a long history, but has developed into its most useful, contemporary form, thanks largely to the work of E. T. Jaynes (1922-1997). A brief account of Jaynes's theory of probability is given in this book, and examples of its profitable use in theoretical physics are given. These include its use in statistical thermodynamics (to clarify, in particular, the meaning of thermodynamic entropy), and its use in the theory of Brownian motion (to derive the diffusion equation and the concepts of particle drift-velocity and probability flux). Also discussed is the way in which rational Bayesian probability theory eliminates the paradoxes that block the way toward a realist interpretation of quantum mechanics, and promises a more rational formulation of quantum mechanics involving complex-valued Bayesian probabilities.

Preface

In this book I provide a short account of the rational Bayesian theory of probability developed by E. T. Jaynes. I describe how the theory amounts to an extension of logic, providing a more general way of reasoning that helps to clarify many conceptual problems in theoretical physics, such as how to define thermodynamic entropy in classical statistical thermodynamics, and how to derive the diffusion equation in the theory of Brownian motion. I note also, how the use of a form of Bayesian reasoning, generalised to take account of Heisenberg's uncertainty principle, can remove the well-known paradoxes that seem to block the way to a realist interpretation of quantum mechanics, and provide a clearer understanding of the subject. I assume the reader is already familiar with probability theory, analytical mechanics, classical thermodynamics, and (non-relativistic) quantum mechanics. Familiarity with statistical thermodynamics would also be useful, but the Bayesian formulation of it given in this book might serve as an introduction to statistical thermodynamics for someone who has not yet studied it.

Bayesian Probability

There are two Bayesian theories of probability. One is termed ‘subjective’ and the other ‘rational’. They differ only in the way we arrive at a prior probability distribution over the possible outcomes of a process. That is, the probability distribution we should hold when we start out with the very minimum of knowledge of the process in question, hoping (by means of Bayes’ rule) to sharpen up the distribution as more knowledge comes to light.

In subjective Bayesian probability, we are at liberty to form a prior distribution over the various possibilities in a purely subjective manner. We choose a probability distribution that we judge to represent our initial state of knowledge well enough.

In rational Bayesian probability we form a prior distribution using certain *rules* of probability assignment. One such rule is the rule of indifference. It applies when our initial knowledge is such that we have no reason to believe in one possibility more than another, and so assign equal probabilities to each. There are other more sophisticated rules, such as the method of maximum entropy which will be described later.

There is much debate amongst statisticians as to which approach is ‘correct’.

In this book we use rational Bayesian probability, which Jaynes has called ‘the logic of science’.[1]

[1] Jaynes E. T. ‘Probability Theory -*The logic of science*’ „„

Probability is then an extension of logic and operates independently of the subjective thoughts of any person using it.

The mathematical calculus of probability theory remains the same when adopting the rational Bayesian approach to it. We assume the reader is familiar with the calculus of probability, and concentrate on the implications of the Bayesian interpretation.

1. Bayesian probability in theoretical physics

When rational Bayesian probability theory is used in physics, we have the general problem of logically deriving the probability of a property of a physical system based on supposed knowledge of other properties of the system. The supposed knowledge is knowledge held by an observer of the system.

In relativity theory, an observer is associated with a particular coordinate frame. Different observers being associated with different coordinate frames. In probability theory, an observer is associated with a particular state of knowledge. Different observers being associated with different states of knowledge.

So, when we speak of supposed knowledge, we will mean knowledge of a system held by a particular observer. Different observers may be in different states of knowledge about properties of the same system. They may know different things about the properties of the

system. When we speak of knowledge *we* hold, we will mean knowledge we hold as an observer.

In theoretical physics, we reason about *models* of physical systems rather than the systems themselves. We need, therefore to consider how supposed knowledge of an observer should be included in any model. This is done by marking certain modelled properties as 'known'. We then apply probability theory to calculate probabilities of other properties.

Henceforth, when we speak of a 'system' or 'physical system' we will mean a *model* of the system; and when we speak of knowledge of properties of a system, we will mean *supposed* knowledge of properties in the model of the system.

Let Y be a proposition claiming the knowledge we hold about system properties. Let G be a proposition claiming our general knowledge including our knowledge of the form of the system, the physical laws governing the system, and the usual laws of reasoning. We suppose that knowledge Y is limited in the sense that we cannot deduce from it, using our general knowledge G, the truth or falsehood of all propositions relating to properties of the system.

When calculable, the probability $P(X|YG)$ of a proposition X claiming a particular property of the system, under knowledge Y and knowledge G, is a non-

negative real number in the range 0 to 1, measuring our degree of belief in the truth of X under knowledge YG.

In the rational Bayesian approach, probability theory is thus viewed as an extension of logic. (Hence Jaynes' use of the phrase 'the logic of science' to describe probability theory.) Accordingly, probabilities are not limiting frequencies in infinitely many cases or 'trials' of the process in question. They are not physically necessary frequencies under specified physical conditions. They are instead, logically deduced degrees of belief, given the (limited) knowledge we (as an observer) hold.

As different observers may be in different states of knowledge about a system, they may deduce different probabilities for one and the same property of a system. As probabilities are not physical properties of a system, there is no contradiction here.

A probability equal to 1 means we have (full) belief of the property in question, i.e. belief without qualification. We will then say simply, that we *believe* that property. A probability equal to 0 means we have no belief in the property.

Note that 'belief' does not imply *certainty* of the presence of the property. We may believe something that is actually not true, i.e. we may be mistaken in our belief. A proposition calculated to have a probability of 1 or 0 therefore leaves the actual truth or falsity of the proposition open to question. It is, after all, not really

justified to suppose we know anything for certain about any physical system. Certainty is present only in mathematics, the logically deduced propositions of which are certainly true (not just believed to be true). In rational Bayesian probability theory, the propositions representing our knowledge may have probabilities equal to 1, but are still, only *believed* to be true. Propositions that follow logically from propositions believed to be true, are themselves only *believed* to be true. Thus if $Y \Longrightarrow X$ (i.e. proposition Y implies proposition X) and we believe in Y, then $P(X|YG) = 1$ means we believe in X whether or not Y is actually true.

2. Rules of the probability calculus

When employing the basic rules of the probability calculus, we use the simple (Boolean like) notation for conjunctions and disjunctions of propositions. So, as in Jaynes (2003), the disjunction and conjunction of propositions A and B are denoted $A + B$ and AB respectively.

Since our knowledge G (defined in Section 1) is common to probabilities relating to a particular system, we drop the G in $P(X|YG)$ writing it as $P(X|Y)$, the presence of G being understood. The basic rules of the probability calculus are, the same as in ordinary probability theory. These are the sum rule and the product rule:

$$P(A+B|Y) = P(A|Y) + P(B|Y) - P(AB|Y)$$

$$P(AB|Y) = P(A|Y)P(B|AY)$$

where, for us, $P(B|AY)$ is our degree of belief in B when we believe A (as well as Y) to be true.

3. The probability of a particular value of a continuous variable

Suppose a continuous variable x, represents some physical property of a system, and that under knowledge Y of the physical state of a system, we have a valid probability density $p(x)$ over the range of allowed values of x. The probability that x lies in a given *interval* is calculable, and x is, of course, supposed to take on a definite real value even though we don't know which.

Now under knowledge Y, the probability that x has any one particular real value x' has to be zero. For the probability $p(x')\Delta x$ for x lying in the small interval $x' < x < x' + \Delta x$ tends to zero as $\Delta x \to 0$ even though $p(x') \neq 0$. Hence the probability that $x = x'$ is 0. Therefore, the probability that $x \neq x'$ is 1. According to our interpretation of probability, a probability equal to 1, this does not mean that x is actually different from x'. It means only that we *believe* x differs from x', leaving open the possibility that x might in fact equal x'.

4. Conditional probabilities

Consider a *sample space* representing mutually exclusive and exhaustive propositions relating to the state of a system. Any point in the space is an atomistic proposition claiming a possible state, and vice versa. Any *set* of points is the proposition claiming one or other member of the set is true. Let A and B be propositions of this kind, then the product rule gives us the formula

$$P(B|AY) = \frac{P(BA|Y)}{P(A|Y)}$$

This is Bayes' rule which may be used to calculate the probability of B conditional on knowledge A.

In the usual probability theory, Bayes rule plays a part, but its interpretation is different. The rule is supposed to give us the probability of B on condition certain physical properties described by A are present in addition to the physical properties described by Y. The distinction is critical when it comes, for example, to resolving paradoxes in a realist interpretation of quantum mechanics.

5. Consequences of the basic rules of probability

We have already noted that the mathematical formalism of Bayesian probability theory remains the same in the probability theory normally used in theoretical

physics[2]. This means we can draw upon the many derivable results of probability theory (e.g. the extended product rule, and the central limit theorem) without having to reconstruct new proofs of these results. The formal proofs remain mathematically the same. Only the interpretation of the results is changed.

In particular we note here (without repeating the formal proof) the extended product rule, or multiplication theorem:

$$P(A_1A_2\ldots A_n|Y) = P(A_1|Y)P(A_2|A_1Y)P(A_3|A_1A_2Y) \\ \ldots P(A_n|A_1A_2\ldots A_{n-1}Y)$$

where $A_1, A_2, \ldots A_n$ are propositions claiming properties of the system in question.

We note also, other forms of Bayes' rule, following from the product rule. Firstly

$$P(B|AY) = \frac{P(B|Y)}{P(A|Y)}P(A|BY)$$

where $P(B|Y)$ can be viewed as our 'prior' probability for B, and $P(B|AY)$ our 'posterior' probability (after

[2] in which probability of an event is taken to be the fraction of times the event occurs in a large number of cases or trials; that fraction being set by physical properties of the system.

acquiring additional knowledge A). Secondly, in relation to distribution functions, we have

$$P(x_k|AY) = \frac{P(x_k|Y)P(A|x_kY)}{\sum_{i=1}^{n} P(x_i|Y)P(A|x_iY)}$$

which follows from the first relation when B is one of n mutually exclusive and exhaustive outcomes x_k ($k = 1,\ldots n$).

6. Expected values

Suppose $P(x_j|Y)$ ($j = 1,\ldots m$) is our probability distribution over a discrete physical variable x taking one or other of the values $x_1,\ldots x_m$. Then, of course, the 'expected value' $\langle x\rangle$ of x under knowledge Y is, by definition

$$\langle x\rangle_Y = \sum_{j=1}^{m} x_j P(x_j|Y)$$

If A_i ($i = 1,\ldots n$) are any other set of propositions claiming exclusive and exhaustive properties of the system, then we may speak of the conditional expected value $\langle x\rangle_{A_iY}$ under knowledge of Y and knowledge that the i^{th} property is present. If $P(A_i|Y)$, $i = 1,\ldots n$ is our probability distribution over the A_i and $P(x_j|A_iY)$ our

conditional probability distribution over the x_i, we can write our unconditional distribution as

$$P(x_j|Y) = \sum_{i=1}^{n} P(A_i|Y)P(x_j|A_iY)$$

(This is proved by writing the LHS as $P(x_j(A_1 + A_2 + \ldots + A_n)|Y)$, which can be set equal to $P(x_jA_1 + x_jA_2 + \ldots + x_jA_n)|Y)$, and then applying the sum and product rules.)

By multiplying through by x_j and summing over j we can use this result to show that

$$\langle x \rangle_Y = \sum_{i=1}^{n} P(A_i|Y)\langle x \rangle_{A_iY}$$

Now, as well as the expected value of a variable, like x under condition Y, we have an expected value $\langle f \rangle$ of a function $f(x)$ of x defined as

$$\langle f \rangle_Y = \sum_{j=1}^{m} f(x_j)P(x_j|Y)$$

The concept of expected value extends, of course, to the case of probability distributions over two, or more

discrete variables, and to cases in which some or all of the variables are continuously distributed. For example, associated with a probability distribution $P(x_i, y_j|Y)$, where $i = 1, \ldots m$ and $j = 1, \ldots l$, we have

$$\langle x \rangle_Y = \sum_{i=1}^{m} \sum_{j=1}^{l} x_i P(x_i, y_j|Y)$$
$$\langle y \rangle_Y = \sum_{i=1}^{m} \sum_{j=1}^{l} y_j P(x_i, y_j|Y)$$

for the expected values of x and y, or, in the case both variables are continuous,

$$\langle x \rangle_Y = \int_{\mathrm{L}} \int_{\mathrm{M}} x P(x, y|Y) dx dy$$
$$\langle y \rangle_Y = \int_{\mathrm{L}} \int_{\mathrm{M}} y P(x, y|Y) dx dy$$

where M and L are the respective ranges of the variables x and y; and the expected values of $f(x)$ and $g(x, y)$ are

$$\langle f \rangle_Y = \int_{L} \int_{M} f(x) P(x, y|Y) dx dy$$

and

$$\langle g \rangle_Y = \int_{\mathrm{L}} \int_{\mathrm{M}} g(x, y) P(x, y|Y) dx dy$$

and so on.

7. Expected values as knowledge

In the application of Bayesian probability to theoretical physics, we may claim knowledge of the values of particular physical variables.

It will, however, be useful, also to claim knowledge of the *expected* values of particular physical variables. In doing this we are anticipating a calculable probability distribution over the variables whose expected values we claim to know.

This extension of possible states of knowledge, employed by E. T. Jaynes, enabled him to introduce his most profitable principle of prior probability assignment, namely the method of maximum entropy. This was the great advance in probabilistic reasoning initiated by Jaynes.

8. Rules for forming prior probability distributions

In addition to the probability calculus, we have (in rational Bayesian probability theory) rules for forming *prior* probability distributions, i.e. probability distributions under our initial knowledge.

Jaynes has employed rules for this purpose which we number 1 to 4 and give below. (Jaynes himself devised Rules 3 and 4, and was of the opinion that there may well

be further rules yet to be discovered.) The four rules useful for forming prior distributions are stated as follows[3]:

1. The principle of similarity
If one problem of finding a prior probability distribution is recognisable *similar* to another, then we should assign the same probability distribution in each case.

2. The principle of indifference
If we are quite indifferent with regard to which of a set of mutually exclusive propositions might be true, we should attach equal probabilities to each.

3. The method of transformation groups
If, after a transformation, the problem of finding a prior probability distribution is recognisably similar to that before, and if a one-to-one equivalence exists between the propositions claiming properties before and after the transformation, then we can write down an equation that must be satisfied by the probability distribution we seek.

[3] The strict adoption of these rules is the thing that distinguishes *rational* Bayesian probability theory from *subjective* Bayesian probability theory. In espousing rational Bayesian probability, prior distributions are, for us, justified *logically* from the knowledge an observer holds. This way, a scientific theory using probability remains objective; i.e. independent of opinions or subjective judgements of an observer.

This helps us find the mathematical form of the distribution.

4. The method of maximum entropy

If our knowledge only takes the form of certain mathematical constraints that our prior probability distribution must satisfy, then that distribution is the one which maximises the 'information entropy' defined below.

The information entropy (or 'entropy' for short) of a probability distribution $p_i = P(x_i|Y) \quad (i = 1, \dots n)$ is defined as

$$\mathcal{H} = -\sum_{i=1}^{n} p_i \ln p_i$$

and is taken to be the measure (under knowledge Y) of our ignorance regarding which of the propositions $x_i \ (i = 1, \dots n)$ is true.[4]

On account of the meaning of $\mathcal{H}$, it is rational for us to adopt, as our prior probability distribution, the values of the p_i that maximise $\mathcal{H}$ subject only to the constraints.

[4] The question as to whether the entropy, as defined, is the best or only rational measure of our degree of ignorance has been considered at length by Jaynes, and answered in the affirmative. See Section 11.3 of his book 'Probability Theory-The Logic of Science'.

Adopting any other values of the p_i would be to claim we are less ignorant than we actually are. This is the method of maximum entropy.[5]

The information entropy is evidently zero when all but one of the p_i are zero[6]. Also, since $p \ln p$ (as a function p) is negative for $0 < p < 1$ and zero for $p = 0$ and for $p = 1$, the information entropy $\mathcal{H}$ is clearly always positive or zero. We will show later that its greatest possible value is $\ln n$, occurring when all the p_i are equal.

$\mathcal{H} = 0$ if and only if we are in a state of *maximal* knowledge with regard to which of the propositions x_i $(i = 1, \ldots n)$ is true. This is the state of knowledge in which we believe a particular proposition x_i is true, and have no belief in the others.

$\mathcal{H} = \ln n$ if and only if we are in a state of maximal ignorance about which of the propositions x_i is true. This is the state of knowledge in which we cannot see why we should believe more in one proposition than in any of the others.

---

[5] This entropy here is not, of course, a thermodynamic entropy. That's why it is often called *information* entropy, to make a clear distinction. As we will see, in statistical mechanics, we have a calculable probability distribution over the microstates of a system, and the information entropy of that distribution turns out to be demonstratively *proportional* to the ordinary thermodynamic entropy of the system.

[6] When $p_i = 0$, we take $p_i \ln p_i$ to be zero, i.e. we take it to equal its value as $p_i \to 0$ (which is zero).

9. Examples of the application of the rules for forming prior probability distributions

Example of the use of Rule 1

Suppose a mechanical die throwing machine, and the slightly biased six-sided die it throws, are manufactured for use in gambling.

After inserting the die, the machine tosses it about so many times before throwing it out, that we can't accurately predict the outcome. By recording many observations of the outcomes of very many throws, we might however be able to come up with a *probability* of each possible die score in a single throw.

If a second die, and a second die throwing machine are manufactured *to exactly the same specifications*, we have another problem –the problem of calculating the probability of each possible outcome in this second case.

The two problems are clearly similar. So having come up with the six probabilities in the first case, we can immediately set the probabilities in the second case equal to those we came up with in the first.

Example of the use of Rule 2

Suppose, in the above gambling machine, we know the die is *unbiased*. That is, suppose we know the die is an accurately made cube of material of uniform density with a uniform surface finish, and marked in a way that does not weight its sides. Then, we are *indifferent* as

to which face will end upward, and we should therefore assign equal probabilities to the six possible scores in a single throw.

(Note that even if we know the die *is* biased but do not know toward *which face* the biased is directed, we are still indifferent with regard to which face will come up in a single throw and we should still assign a probability of $\frac{1}{6}$ to each possibility.)

Example of the use of Rule 3

The simplest example of application of this rule is its use to prove Rule 2.

As a general case, we may suppose we have a process with n possible outcomes. Let the possible outcomes be numbered $1, \ldots n$, and let the probabilities of these outcomes be p_i $(i = 1, \ldots n)$. We pose the problem of finding probability distribution p_i when our state of knowledge Y is one of total indifference as to which outcome will occur.

Making a transformation which is just a relabelling of the outcomes, we can pose another problem: that of finding the probabilities p'_j $(j = 1, \ldots n)$ of the outcomes labelled in the new way.

Under our knowledge Y of total indifference concerning which outcome occurs, the two problems are *similar*. Therefore, by Rule 1, we should set the distributions equal. That is, we should set $p'_k = p_k$ for $k =$

$1, \ldots n$. On the other hand, probabilities referring to the same outcome must be the same. This means that, for any one chosen value of k, it must be that $p'_k = p_j$ where label j, in the first way of labelling, refers to the same outcome as k in the second way of labelling. Putting together the results $p'_k = p_i$ and $p'_k = p_j$ we find $p_k = p_j$. So, two probabilities in the distribution p_i $(i = 1, \ldots n)$ are equal. Letting k run from 1 to n, we find that *all* the p_i must be equal to one another. We have thus derived Rule 2 from Rule 3 with the help of Rule 1.

Example of the use of Rule 4

Again, the simplest example of this rule is the proof of Rule 2 using it.

As a general case, we may again suppose we have a process with n possible outcomes. If we have no information that would lead us to favour some outcomes over others. Then the probabilities p_i $(i = 1, \ldots n)$, claiming each of n possible outcomes of a process, are subject only to the normalisation constraint

$$\sum_{i=1}^{n} p_i = 1$$

Under this constraint, we may employ Rule 4 to find the values of the p_i. These will be those values that maximise the information entropy

$$\mathcal{H} = -\sum_{i=1}^{n} p_i \ln p_i$$

under the normalisation constraint.

We use Lagrange's method for locating turning values under constraints. According to this method, we start by introducing a constant parameter λ and define the function

$$S = -\sum_{i=1}^{n} p_i \ln p_i + \lambda(\sum_{i=1}^{n} p_i - 1)$$

where λ multiplies a term that is zero when the constraint is applied.

For any chosen values of the p_i we calculate how S changes under a small variation of the p_i away from the chosen values.

Let the variation in the p_i be εq_i, so that the p_i change to $p_i + \varepsilon q_i$, where ε is a small real number and the q_i are arbitrary real numbers of order 1. Then, S changes to

$$S + \delta S = -\sum_{i=1}^{n} (p_i + \varepsilon q_i) \ln(p_i + \varepsilon q_i) + \lambda(\sum_{i=1}^{n} (p_i + \varepsilon q_i) - 1)$$

where, to second order in ε,

$$\ln(p_i + \varepsilon q_i) = \ln p_i (1 + \frac{q_i}{p_i}\varepsilon) = \ln p_i + \ln(1 + \frac{q_i}{p_i}\varepsilon) = \ln p_i + (\frac{q_i}{p_i}\varepsilon) - \frac{1}{2}(\frac{q_i}{p_i}\varepsilon)^2$$

assuming none of the p_i are zero. Hence

$$S + \delta S = -\sum_{i=1}^{n} (p_i + \varepsilon q_i)(\ln p_i + \varepsilon \frac{q_i}{p_i} - \frac{1}{2}\varepsilon^2 (\frac{q_i}{p_i})^2) + \lambda(\sum_{i=1}^{n} (p_i + \varepsilon q_i) - 1)$$

from which we find

$$\delta S = -\varepsilon \sum_{i=1}^{n} q_i (1 + \ln p_i - \lambda) - \varepsilon^2 \sum_{i=1}^{n} (\frac{1}{2}\frac{q_i^2}{p_i})$$

This is zero (to first order in ε) only when

$$1 + \ln p_i - \lambda = 0$$

for all i. So, the p_i that mark the one and only locally flat region or turning point of the function S must each have the value $e^{\lambda-1}$.

That $p_i = e^{\lambda-1}$ $(i = 1,2, \ldots n)$ mark the one and only *maximum* value of S is evident from the sign of the term in ε^2 in the above expression for δS. These values of p_i thus maximise

$$S = \mathcal{H} + \lambda(\sum_{i=1}^{n} p_i - 1)$$

under no constraint.

To satisfy the normalisation constraint, it is necessary that λ be $1 - \ln n$. Thus the p_i are

$$p_i = e^{\lambda-1} = 1/n \qquad i = 1,2, \ldots n$$

These normalised values of the p_i maximise S. They must also maximise $\mathcal{H}$. For, if some other normalised values p'_i made $\mathcal{H}$ larger, they would make S larger, which is impossible.

The value of the information entropy at its maximum is

$$\mathcal{H} = -\sum_{i=1}^{n} p_i \ln p_i = -\sum_{i=1}^{n} (1/n) \ln(1/n) = \ln n$$

The above derivation of the result $p_i = 1/n$ $(i = 1,..n)$ assumed none of the p_i where zero. It is clear, however, that if some *were* zero this could not make the entropy any larger than $\ln n$.

For if we started out by setting some (say m) of the p_i equal to zero. We would then have to maximise

$$-\sum_{i=1}^{n-m} p_i \ln p_i$$

subject to the constraint

$$\sum_{i=1}^{n-m} p_i = 1$$

The same method as that already explained would then result in a maximum value of $\ln(n-m)$, which is less than $\ln n$.

Therefore, $p_i = 1/n$ for $i = 1,\ldots n$ is certainly the distribution giving the maximum possible entropy under the single constraint. We have therefore shown that the

principle of indifference (Rule 2) is derivable from the method of maximum entropy (Rule 4). QED

Rule 4 has, of course, much wider application than the reproduction of Rule 2; and the method of maximum entropy, together with the method of transformation groups (Rule 3) forms the basis of rational Bayesian probability theory and its application to theoretical physics.

10. The effect (on the entropy) of knowledge other than the requirement for normalisation

Consider again, a process with n possible outcomes numbered $i = 1,2, \ldots n$. At the start, we generally hold knowledge in addition to the requirement for normalisation of the distribution p_i.

Then, when applying the method of maximum entropy to find the prior distribution, this additional knowledge (for, example knowledge of the expected value of some function of i) imposes another constraint on the p_i.

The method of maximum entropy now leads to a distribution p_i different from that resulting from the normalisation requirement alone. That is, different from that which maximised the entropy under the normalisation requirement alone. The new entropy must therefore be less than it was before.

If we claim to hold yet further knowledge expressible as a further constraint on the possible distribution p_i, then again, the method of maximum entropy must lead to a yet lower value of the entropy; and so on.

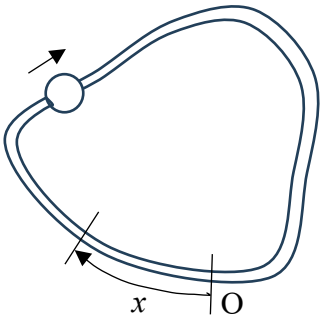
All this is to be expected, since additional knowledge diminishes the level of our ignorance with regard to which value of i actually applies.

If further knowledge expressible as a further constraint on the possible distribution p_i, leads, by the method of maximum entropy, to a probability distribution the same as before, it must be that the further constraint is already a consequence of the previous constraints. It does not then represent new knowledge after all.

11. The role of measure in the case of continuous variables

Although we have stated the above Rules 1 to 4 in relation to discrete distributions, they also apply to continuous distributions, whenever we can divide up the continuous domain into small parts of *equal measure*.

Consider, for example, the case of a bead sliding around of closed loop of wire of any shape. Suppose the loop is rigid and of infinite mass, and floats freely in space without rotating.



Let x be the arc length along the wire from a chosen origin on the wire.

Suppose our prior knowledge Y makes no reference to the position of the bead. For example, Y may represent only knowledge of its velocity. Then, we can find our prior distribution $p(x)$ as follows.

We note that the arc length x has a natural measure, as lengths or distances do in Euclidean geometry. We thus divide the length l of the wire, into a number of segments of *equal measure*, i.e. in this case of equal length, and we label these $i = 1, \ldots n$. Let x_i denote the proposition claiming the coordinate of the bead lies in the i^{th} segment. Then $P(x_i|Y)$ may stand for our probability, under our knowledge Y, that the bead is in the i^{th} segment at a particular time t.

By the principle of indifference, we should set $P(x_i|Y) = 1/n$.

Letting the number of segments tend to infinity (as described in Section 3), we are thus led to set the continuous probability density $p(x)$ equal to a constant, i.e. to set

$$p(x) = 1/l$$

Crucial here, is the *equality* of the lengths into which the wire is divided before taking the limit.

Suppose, instead of x we employed another variable say y $(= x^3)$ $(0 \leq y < l^3)$ to specify the bead's position. Then we could set up a representative $y$axis for the physical quantity y; but it would be quite wrong to divide this axis (or the part of it from $y = 0$ to $y = l^3$) into equal small segments, and to claim indifference with regard to which of these segments contained the actual value of y. This because there *is* now a difference between the segments, as they are known to relate to segments of different lengths (of different natural measure) along the $x$axis.

To find the density $\tilde{p}(y)$ we can equate the probability of being in a small segment Δy at y on the $y$axis to the probability of being in the corresponding segment Δx at x on the $x$axis. That is, we can set

$$\tilde{p}(y)\Delta y = p(x)\Delta x$$

Figure 2

where $p(x) = l^{-1}$, $\Delta y = 3x^2 \Delta x$ and $x = y^{1/3}$. We thus find, for $\tilde{p}(y)$, the formula

$$\tilde{p}(y) = (3l)^{-1} y^{-2/3}$$

This is the correct probability distribution over y.

The derivation can be put another way. If we divide the domain of y into *equal* segments, we cannot claim the probability density $\tilde{p}(y)$ takes the same value in each segment. This is because variable y does represent a physical quantity with a *natural* measure. A small segment Δy does not have a natural measure equal to Δy or proportional to Δy.

However, y does have a measure, and the measure of Δy can be deduced from its relation to the corresponding segment Δx under the relation $y = x^3$. We have (in the limit)

$$\Delta y = \frac{dy}{dx} \Delta x$$

So, we need to apply the weight dx/dy to Δy to get a measure of Δy. We should therefore take the 'measure density' $m(y)$ over y equal to dx/dy. The relative measure of Δy may thus be written

$$m(y)\Delta y$$

where $m(y)(= dx/dy)$ is the 'relative measure density function'; it is a function defined to within an unimportant constant factor. We often refer to a relative measure density as a measure density for short.

Therefore, the probability $\tilde{p}(y)\Delta y$ that y falls in the interval Δy (under indifference with regard to the value of x), should be taken proportional to the measure of Δy, i.e. to $m(y)\Delta y$. That is,

$$\tilde{p}(y)\Delta y \propto m(y)\Delta y$$

or

$$\tilde{p}(y) = km(y)$$

The constant of proportionality k, is found by integrating over y and setting the result equal to 1. Thus

$$k\int_0^{l^3} \frac{dx}{dy}dy = k[x(y)]_0^{l^3} = k[x]_0^l = kl$$

giving

$$\tilde{p}(y) = l^{-1}m(y)$$

In the present case $m(y) = dx/dy = \frac{1}{3}y^{-\frac{2}{3}}$ giving $\tilde{p}(y) = \frac{1}{3}l^{-1}y^{-\frac{2}{3}}$.

12. The entropy in the case of continuous variables

We now address the problem of defining information entropy in the case of outcomes specified by the value of a single continuous physical variable x, or by a number of such variables.

Let us use an x-axis to represent values of the variable x. In so doing, we do not imply that equal intervals of length along this x-axis represent segments of equal natural measure. The x-axis serves only to geometrically represent the variable's possible values.

The domain L of our probability density over x is supposed to consist of certain parts of the x axis. The domain may be maximal, i.e. include the whole x axis $(-\infty < x < \infty)$.

Now, we suppose there is a known measure of the physical variable x, but this is not necessarily a uniform one. We write the measure of a small element Δx of x as $m(x)\Delta x$, where $m(x)$ is a continuous smooth measure density function containing, as we have said, an unimportant constant factor.

We next divide the domain L into small parts Δx_i *of equal measure* lying adjacent to each other along the x-axis, the label i taking successive integer values. We let p_i stand for the probability that x lies in the part Δx_i.

The information entropy for the discrete distribution p_i is

$$-\sum_i p_i \ln p_i$$

This is supposed valid even when the series is infinite (on account of L being infinite); that is, we assume in such a case, that p_i tends to zero fast enough as $i \to \pm\infty$ for the series to converge.

We now express p_i as $p(x_i)\Delta x_i$ where $p(x)$ is the continuous probability density evaluated at a point x_i inside Δx_i. As the Δx_i are of equal measure, we can set

$$m(x_i)\Delta x_i = \text{const.} = \Delta c$$

so that

$$p_i = p(x_i)\Delta x_i = \frac{p(x_i)\Delta c}{m(x_i)}$$

The information entropy is thus

$$-\sum_i p_i \ln p_i = -\sum_i p(x_i)\Delta x_i \ln \frac{p(x_i)\Delta c}{m(x_i)}$$

$$= -\sum_i p(x_i)\Delta x_i(\ln\frac{p(x_i)}{m(x_i)} + \ln\Delta c)$$

or since

$$\sum_i p(x_i)\Delta x_i = 1$$

it is

$$-\sum_i p_i \ln p_i = -\sum_i \left(p(x_i)\Delta x_i \ln\frac{p(x_i)}{m(x_i)}\right) - \ln\Delta c$$

The sum on the RHS becomes an integral in the limit as all the $\Delta x_i \to 0$.

The last term on the RHS goes to infinity. However, it is clear that, from the point of view of actual applications of the method of maximum entropy, the presence of $\ln\Delta c$ is of no consequence; it is just a constant under variation of $p(x)$, and can be omitted.

In this way, we are led to start afresh and *define* the information entropy in the continuous case as

$$\mathcal{H} = -\int_L \left(p(x)\ln\frac{p(x)}{m(x)}\right)dx$$

where L is the domain of $p(x)$, and $m(x)$ is the continuous smooth natural measure density associated with the physical variable x.

Because $m(x)$ is specified only to within an arbitrary constant factor, $\mathcal{H}$ is defined only to within an arbitrary additive constant.

While there is a finite maximum value of the entropy (occurring when we hold no knowledge about the value of x), there is no finite minimum value. For, as we come more and more to believe that the actual value of x is close to a particular value, i.e. as $p(x)$ becomes more and more sharp (in delta-function fashion), $\mathcal{H}$ goes more and more toward $-\infty$. It does not go to zero.[7]

The problem of calculating or confirming the measure density $m(x)$ can be difficult. There is as yet, no general way to do it. But it can sometimes be done using the method of transformation groups.[8]

Generalisation to the case of outcomes specified by two or more continuous physical variables spanning a representative space of two or more dimensions is straight-forward. The measure density is then a function of all the variables.

13. Applications of the method of maximum entropy in the continuous case

[7] This can be seen by representing the delta function as a normal distribution in the limit as the standard deviation σ tends to zero. The value of the entropy, as we have defined it, is then a constant plus the term $\ln(\sigma\sqrt{2\pi})$, which tends to $-\infty$.

[8] We give examples in Appendix A.

First application

We first apply the method of maximum entropy to the case of a single continuous physical variable x with a known measure. We use, again, an x-axis to represent values of the variable x.

In the case of total indifference with regard to the value of x,

$$\int_L p(x)dx = 1$$

is the only constraint on $p(x)$. Maximising the entropy subject to it alone gives

$$p(x) = \frac{m(x)}{\int_L m(x)dx}$$

where $m(x)$ is the measure density. This result is proved as follows

Using Lagrange's method, we first maximise

$$S = -\int_L p \ln \frac{p}{m} dx + \lambda \left(\int_L p dx - 1 \right)$$

subject to no constraint, then apply the constraint equation to find the constant λ.

Under any small variation δp of p, i.e. variation of the function $p(x)$ by an arbitrary but small $\delta p(x)$, the

corresponding change δS in S is, to second order, calculated as follows. We use the general expansion

$$f(x+\delta) - f(x) = \delta \frac{\partial f}{\partial x} + \tfrac{1}{2}\delta^2 \frac{\partial^2 f}{\partial x^2}$$

which gives

$$\begin{aligned}(p+\delta p)\ln\frac{(p+\delta p)}{m} - p\ln\frac{p}{m} \\ &= \delta p \frac{\partial}{\partial p}\left(p\ln\frac{p}{m}\right) + \tfrac{1}{2}\delta p^2 \frac{\partial^2}{\partial p^2}\left(p\ln\frac{p}{m}\right) \\ &= \delta p\left(\ln\frac{p}{m}+1\right) + \tfrac{1}{2}\delta p^2 \frac{1}{p}\end{aligned}$$

Hence

$$\delta S = -\int_L \left(\delta p\left(\ln\frac{p}{m}+1\right) + \tfrac{1}{2}\delta p^2\frac{1}{p} - \lambda\delta p\right)dx$$

To *first* order, this is zero only when

$$\ln\frac{p}{m} + 1 - \lambda = 0$$

That is, only when

$$p(x) = m(x)e^{\lambda-1}$$

Because of the sign of the term in δp^2 in the equation for δS, this is the probability distribution that does maximise S under no constraint. Choosing λ so that

$$e^{\lambda-1} = \frac{1}{\int_L m(x)dx}$$

we satisfy the normalisation constraint on $p(x)$. So, we arrive at

$$p(x) = \frac{m(x)}{\int_L m(x)dx}$$

as the probability density maximising the entropy $\mathcal{H}$. QED

The value of the information entropy $\mathcal{H}$ associated with the probability density maximising it, is

$$-\int_L \left(p(x)\ln\frac{p(x)}{m(x)}\right)dx = -\int_L \left(\frac{m(x)}{M}\ln\frac{1}{M}\right)dx = \ln M$$

where

$$M = \int_L m(x)dx$$

The value $\ln M$ is the maximum possible value of the entropy, because total indifference represents the highest degree of ignorance possible.

So as our ignorance of the value of x increases from no ignorance to total ignorance, the information entropy increases from $-\infty$ to $\ln M$.

We note that the term in δp^2 in the expression for δS arises only from variation of the term $p \ln (p/m))$ in the entropy integral, not from variation of the term arising from the constraint. This is generally the case when one or more constraint equations are *linear* functionals of p.

Therefore, whenever the constraint equations are linear in the probability distribution (as is generally the case) Lagrange's method always leads to a distribution *maximising* the entropy.

It has been pointed out by Jaynes (2003) that if the measure density $m(x)$ is unknown, the problem of finding it is the same as that of finding the prior probability density in the case of no knowledge (other than the requirement for normalisation). The relation

$$p(x) = \frac{m(x)}{\int_L m(x)dx}$$

derived above, clearly shows this to be the case.

Second application

Suppose we know a particle lies somewhere within an (otherwise empty) enclosure of volume V. Let $p(\boldsymbol{r})$ be the probability density for the particle to be at position $\boldsymbol{r}$ in the enclosure, so that the probability it is within a small volume ΔV at $\boldsymbol{r}$ is $p(\boldsymbol{r})\Delta V$.

Then the information entropy is

$$\mathcal{H} = -\int_V p(\boldsymbol{r}) \ln \frac{p(\boldsymbol{r})}{m} dV$$

where the measure density $m(\boldsymbol{r})$ is just a constant m, as volume is a natural measure of space.

If we have no information about where the particle is situated in V, then maximising $\mathcal{H}$ subject only to the condition

$$\int_V p(\boldsymbol{r}) dV = 1$$

results in $p(\boldsymbol{r})$ being a constant, equal of course to V^{-1}. The information entropy then has its maximum value given by

$$\mathcal{H} = \ln mV$$

The proof is left to the reader.

14. Logical independence

Suppose we hold knowledge Y of a physical process. Then a proposition B claiming an event B in this process is 'logically independent' of a proposition A claiming an event A in the process, if and only if, the probability of B is unchanged when we know A as well as Y. We write this condition for independence as

$$P(B|Y) = P(B|AY)$$

So, the above relation applies when we feel sure that extra knowledge A makes no difference to the probability we should assign to B. Knowledge A is then redundant as far as our degree of belief in B is concerned.

The various conditions for independence of *any number* of propositions are algebraically the same as in the usual probability calculus, but it will be as well to note them here.

Propositions $A_1, A_2, \ldots A_n$, are 'logically independent' if and only if the $n-1$ conditions,

$$\begin{aligned} P(A_2|A_1Y) &= P(A_2|Y), \\ P(A_3|A_1A_2Y) &= P(A_3|Y), \\ &\ldots\ldots \\ P(A_n|A_1A_2\ldots A_{n-1}Y) &= P(A_n|Y) \end{aligned}$$

are satisfied, as well as the same conditions obtained after rearrangement of the propositions in any way.

So, $A_1, A_2, \ldots A_n$ are independent if and only if they are (as we say) pair-wise independent, triple-wise independent, quarto-wise independent, …etc.

If the propositions are only pair-wise independent, that is, if $P(A_2|A_1Y) = P(A_2|Y)$, etc., this is insufficient, to imply logical independence of the propositions $A_1, A_2, \ldots A_n$.[9]

Under the above conditions for logical independence, conjunctions of the propositions $A_1, A_2, \ldots A_n$ are also logically independent. So, for example

$$P(A_3A_5|A_1A_2Y) = P(A_3A_5|Y),$$
$$P(A_8A_5A_1|A_1A_2A_7A_5Y) = P(A_3A_5A_1|Y), \ldots.\text{etc.}$$

This is easily demonstrated by repeatedly applying the product rule.

The product rule

[9] A simple example is given in Mood, Graybill and Boes. Suppose two dice are thrown. Let A denote 'score of first die is odd', B denote 'score of second die is odd', and C denote 'sum of the scores is odd'. Then it is easy to show that A, B and C are pair-wise independent but not triple-wise independent.

$$P(A_1A_2\ldots A_n|Y) = P(A_1|Y)P(A_2|A_1Y)P(A_3|A_1A_2Y) \ldots P(A_n|A_1A_2\ldots A_{n-1}Y)$$

can also be employed to show that when $A_1, A_2, \ldots A_n$ are logically independent, $P(A_1A_2\ldots A_n|Y)$ factors, i.e.

$$P(A_1A_2\ldots A_n|Y) = P(A_1|Y)P(A_2|Y) \ldots P(A_n|Y)$$

But the converse is not always true. That is, factorisation of $P(A_1A_2\ldots A_n|Y)$ does not imply logical independence of $A_1, A_2, \ldots A_n$.

For factorisation to imply independence of $A_1, A_2, \ldots A_n$, it is required that all the following $n-1$ factorisations hold,

$$P(A_1A_2|Y) = P(A_1|Y)P(A_2|Y),$$
$$P(A_1A_2A_3|Y) = P(A_1|Y)P(A_2|Y)P(A_3|Y),$$
$$\ldots..$$
$$P(A_1A_2\ldots A_n|Y) = P(A_1|Y)P(A_2|Y) \ldots P(A_n|Y)$$

as well as the similar conditions obtained by rearrangement of the $A_1, A_2, \ldots A_n$. These conditions are, all together, equivalent to the original conditions involving conditional probabilities. Proof of this equivalence is left to the reader.

So far, the propositions $A_1, A_2, \ldots A_n$ are just that, single propositions. But we often encounter cases in which the propositions $A_1, A_2, \ldots A_n$ are variables.

This happens in dealing with joint probability distributions like $p(r_1, r_2, \ldots r_n)$ over variables $r_1, r_2, \ldots r_n$, where r_1 takes values $1,2, \ldots m_1$, r_2 takes values $1,2, \ldots m_2$, …etc. Letting $r_1, r_2, \ldots r_n$ stand also for propositions respectively claiming particular values of the variables, the propositions $r_1, r_2, \ldots r_n$, are, as we have seen, logically independent if and only if the $n-1$ relations

$$p(r_1 r_2) = p(r_1)p(r_2),$$
$$p(r_1 r_2 r_3) = p(r_1)p(r_2)p(r_3),$$
$$\ldots\ldots$$
$$p(r_1 r_2 \ldots r_n) = p(r_1)p(r_2) \ldots p(r_n)$$

between the normalised distributions, as well as those obtained by rearrangement of the propositions, are all met. When, and only when, these relations hold for *any* values of variables $r_1, r_2, \ldots r_n$, the propositions $r_1, r_2, \ldots r_n$ are said to be logically independent (without the need to refer to particular values of the variables). But only the final condition

$$p(r_1 r_2 \ldots r_n) = p(r_1)p(r_2) \ldots p(r_n)$$

is in fact needed for independence. This is because the other conditions are derivable from it. This is shown by

summing over the variables omitted in each case. For example, we have, by the sum rule, that

$$p(r_1 r_2) = \sum_{r_3=1}^{m_3} \sum_{r_4=1}^{m_4} \dots \sum_{r_n=1}^{m_n} p(r_1 r_2 \dots r_n)$$

which, on account of the factorisation of $p(r_1 r_2 \dots r_n)$ becomes

$$p(r_1 r_2) = p(r_1)p(r_2) \sum_{r_3=1}^{m_3} p(r_3) \sum_{r_4=1}^{m_4} p(r_4) \dots \sum_{r_n=1}^{m_n} p(r_n)$$

where each sum equals 1.

Note that, very often in our Bayesian reasoning, we will speak of 'the probability of a certain event' rather than 'the probability of the proposition claiming a certain event'. In so doing, we in no way mean to imply that events themselves have probabilities. It is just that it would make our explanations very verbose if we did not adopt this simplification of expression.

Also, for simplification, we will often say that 'one event is logically independent of another', when we mean that 'the *proposition* claiming one event is logically independent of the proposition claiming another'. Sometimes, too, when the meaning is clear, we will use the term 'independence' to mean 'logical independence',

though we should stress that logical independence of events is not the same as *physical* independence of them. Neither on its own implies the other.

Examples when physical independence and logical independence do not imply one another

(i) If two identically biased dice are separately thrown many times, the scores of one are *physically* independent of the scores of the other, on account of there being no physical interaction between the dice.

Now suppose we are initially ignorant of the manner in which the dice are biased, but know that they are biased in the same way. Then, the scores of one die are not, for us, *logically* independent of the scores of the other. For, by observing and gaining knowledge of the scores of one die, we gain information about the likely nature and degree of the bias of the other.

(ii) Suppose a machine can toss a coin so the result is always heads, or always tails, depending on which way a lever inside the machine is turned.

Now, suppose we know only that the machine can toss a coin. Then, our probability for a head in a single toss is (by indifference) one-half.

If we are then shown the position of a lever in the machine (without being told what the lever is for), our probability one-half remains the same, because for us, the knowledge of the lever's position gives us no reason to change our probability for a head. For us, the outcome of

the toss is logically independent of the position of the lever. But, of course, the outcome of the toss is not *physically* independent of the position of the lever.

15. The relation between probability and frequency

In this book, we theorise about probabilities of events occurring in physical processes.

We do not view probabilities as frequencies, even though physicists generally do. That is to say, most physicists think the 'probability' of an event in (or an outcome of) a process is the relative frequency of its occurrence in infinitely many trials of the process, or in an 'ensemble' of identical processes going on simultaneously and independently.

We understand frequencies of occurrence as they do, but we view probabilities as degrees of belief, not as frequences.

So strongly held is the view, that probability is frequency, that physicists have become hostile to Bayesian approaches to probability, at least when they are applied directly to events occurring in physical processes. They may tolerate Bayesian statistical methods in relation to the analysis of likely measurement errors etc. However, when it comes to actually theorising about the physical world, 'probabilities' are viewed as physically real. Popper's claim that the probability of an event is a measure of the

physical *propensity* of the event to occur under certain physical conditions is considered appropriate, and taken to account for and justify the frequency interpretation.

So, for the physicists, a probability of any one of the various possible outcomes of a process is the relative frequency at which that outcome occurs in many trials of the process. The probability is always conditional on the initial physical state of the system. The sequence of outcomes occurring in successive trials form what is called a random sequence. [10]

Even though we do not identify probability with frequency or propensity, we do acknowledge that, often, our degree of belief in a proposition claiming an event equals our *expected* frequency of the event in very many repetitions of the process in question. Also, the observed relative frequency of an event in very many repetitions of the process can be equal to our probability of it in a single trial. These equalities, when they are present, can, as we will show, be derived from the laws of rational Bayesian probability.

[10] Controversy has arisen regarding the meaning of a random sequence. Sometimes it is suggested that a sequence is 'random' when it cannot be generated by any finite set of rules, as if randomness was a *property of the sequence*. But in Bayesian theory it is sufficient to say that a sequence is random when the observer is *unable* to *predict* it on account of their limited knowledge of the process in question.

16. Use of the method of maximum entropy to show that expected frequencies can equal probabilities

Suppose a physical process results in one of m possible outcomes, and we have (using our knowledge of the process and the laws of probability assignment) come up with a probability distribution $p(r)$ over the possible outcomes numbered $r = 1,2, \dots m$.[11] Should we then expect the relative frequency of any chosen outcome in n trials of the process, to be equal to its single trial probability? Well, we should if we know the outcomes of the trials are logically independent.

For then, as shown in all textbooks on probability (Bayesian or not), the distribution $P(s)$ over the number, s, of occurrences of the chosen outcome in n trials is given by the binomial distribution

$$P(s) = {}_nC_s p^s q^{n-s}$$

where p is the probability the chosen outcome occurs in a single trial, and q $(= p - 1)$ is the probability it does not.

Also, as shown in textbooks, the expected value of s in the binomial distribution is np with standard

[11] For example, if the process consisted of throwing an unbiased die, we might have calculated, using the principle of indifference, that $p(r) = \frac{1}{6}$ for $r = 1,2, \dots 6$.

deviation $\sqrt{npq}$; and for large n, the binomial distribution becomes a normal (i.e. Gaussian) distribution. That is,

$$ {}_nC_r p^r q^{n-r} \approx \frac{1}{\sqrt{2\pi npq}} e^{-(r-np)^2/(2npq)} $$

provided both $np >> 1$ and $nq >> 1$.

So, provided trial outcomes are independent, we expect s to be close to the mean np, with standard deviation $\sqrt{npq}$. So, the *relative* number of occurrences of the chosen outcome, in n throws, is close to np/n, i.e. close to p, with a likely error of order $\sqrt{npq}/n$. Hence, our degree of belief that the relative number of the chosen outcome will lie between $p - \varepsilon$ and $p + \varepsilon$, where ε is a small positive number, becomes 1 in the limit $n \to \infty$. As ε may be taken as small as we like, we come to believe that the long-term relative frequency of occurrences of the chosen outcome will be equal to the probability of it occurring in a single trial. The same of course applies to the long-term frequency of any chosen outcome.

Thus, the expected relative frequency distribution $f(r)$ over the possible outcomes in a large enough number of trials would indeed be equal to the probability distribution $p(r)$ over the possible outcomes in a single trial.

That the outcomes of repeated trials of the process *are* logically independent, is provable using the method of maximum entropy. The proof is as follows.

Suppose the process is repeated n times, then the number of possible outcomes is m^n. We prove logical independence of the outcomes of different trials, by showing that the joint probability distribution $p(r_1, r_2, \ldots r_n)$ factors. Here, r_1 $(= 1,2 \ldots m)$ is the variable labelling the possible outcomes in the first trial, r_2 $(= 1,2 \ldots m)$ is the variable labelling the possible outcomes in the second trial, and so on.

In applying the method of maximum entropy to find $p(r_1, r_2, \ldots r_n)$, the constraints to be imposed on $p(r_1, r_2, \ldots r_n)$ relate to the marginal probability distributions $p_1(r_1)$, $p_2(r_2)$,… $p_n(r_n)$. The first of these is

$$p_1(r_1) = \sum_{r_2=1}^{m} \sum_{r_3=1}^{m} \ldots \sum_{r_n=1}^{m} p(r_1, r_2, \ldots r_n)$$

the sum over r_1 being omitted. We have a similar expression for $p_2(r_2)$, in which the sum over r_2 is omitted, and so on. Instead of omitting sums we can employ a Kronika delta in the summand. Then, for example, the marginal distribution function $p_1(r)$ over r can be written as

$$p_1(r) = \sum_{r_1=1}^{m} \sum_{r_2=1}^{m} \dots \sum_{r_n=1}^{m} p(r_1, r_2, \dots r_n)\, \delta_{r_1 r}$$

Now, we are supposing we have already established (and therefore have knowledge of) the probability distribution $p(r)$ over the outcomes of a single trial. So, the marginal distributions must be all the same.

The factoring of the joint distribution $p(r_1, r_2, \dots r_n)$ does not follow *automatically* when the marginal distributions are all the same.[12] None-the-less, as will show, when $p(r_1, r_2, \dots r_n)$ maximises its entropy under necessary constraints, it does factor.

In employing the method of maximum entropy, we must impose on $p(r_1, r_2, \dots r_n)$ the m^n constraints

$$\sum_{r_1=1}^{m} \sum_{r_2=1}^{m} \dots \sum_{r_n=1}^{m} p(r_1, r_2, \dots r_n)\, \delta_{r_j r} = p(r)$$

$$j = 1,2 \dots n, \quad r = 1,2 \dots m$$

[12] For example, consider the case when $n = m = 2$ and the joint distribution $p(r_1 r_2)$ has values $p(11) = 0$, $p(12) = 1/4$, $p(21) = 1/4$, and $p(22) = 1/2$. The marginal distributions $p_1(r_1)$ and $p_2(r_2)$ are here the same. But $p(r_1 r_2)$ does not factor. The product $p_1(r_1)p_2(r_2)$ differs from $p(r_1 r_2)$.

Since $p(r)$ is already normalised, we need not impose the further constraint of normalisation on $p(r_1, r_2, \ldots r_n)$. This is already implied by the above constraint equation (with any one value of j) by summing both sides from $r = 1$ to m.

Abbreviating $p(r_1, r_2, \ldots r_n)$ to p and the sums over the $r_1, r_2, \ldots r_n$ to Σ, we have then, to maximise

$$-\sum p \ln p$$

Subject to the m^n constraints

$$\sum p\, \delta_{r_j r} = p(r) \qquad r = 1,2 \ldots m, \;\; j = 1,2 \ldots n$$

To find $p(r_1, r_2, \ldots r_n)$, we maximise the expression

$$-\sum p \ln p + \sum_{j=1}^{n} \sum_{r=1}^{m} \left(\mu_{jr} \sum p \delta_{r_j r} \right)$$

where the μ_{jr} are multipliers attached to each constraint. We can rewrite this expression as

$$-\sum p \ln p + \sum \sum_{j=1}^{n} \sum_{r=1}^{m} \mu_{jr} p \delta_{r_j r}$$

Setting to zero the variation of this under a general variation of the function p, we obtain the result

$$p(r_1, r_2, \dots r_n) = \exp\left(\sum_{j=1}^{n}\sum_{r=1}^{m} \mu_{jr}\delta_{r_j r}\right)$$

The required distribution can therefore be written as

$$p(r_1, r_2, \dots r_n) = p_1(r_1)p_2(r_2) \dots p_n(r_n)$$

where the $p_j(r_j), j = 1,2 \dots n$, are

$$p_j(r_j) = \exp\left(\sum_{r=1}^{m} \mu_{jr}\delta_{r_j r}\right)$$

Therefore, use of the method of maximum entropy has proved that $p(r_1, r_2, \dots r_n)$ factors.

We have thus proved that under knowledge, alone, of the probability distribution $p(r)$ applying in any single trial, the joint probability distribution over the sample space of all possible outcomes of any number of trials, is just the product of the (identical) distributions applying in each trial. The outcomes relating to the various trials are therefore logically independent. QED

As explained at the beginning of this section, it follows that the relative frequency distribution $f(r)$ and the probability distribution $p(r)$ are the same.

(To be sure of the validity of the proof of the factorisation of $p(r_1, r_2, \ldots r_n)$, we should show that the multipliers μ_{jr} can be evaluated. To do this, we have only to impose the constraints, i.e. to equate the factors $p_j(r_j), j = 1,2 \ldots n$ to $p(r)$. This gives

$$p_j(r_j) = \exp\left(\sum_{r=1}^{m} \mu_{jr}\delta_{r_j r}\right) = p(r)$$

Writing the function on the LHS as a column:

$$p_j(r_j) = \begin{cases} e^{\mu_{j1}} \\ \vdots \\ e^{\mu_{jm}} \end{cases}$$

and writing the function $p(r)$ as a column

$$p(r) = \begin{cases} \alpha_1 \\ \vdots \\ \alpha_m \end{cases}$$

where the α_i are known, we see that the multipliers take the values

$$\mu_{jr} = \ln \alpha_r, \text{ for } j = 1,2 \ldots n, \text{ and } r = 1,2 \ldots m$$

So, they *are* calculable.)

17. Use of the method of maximum entropy to show that probabilities can equal expected frequencies

We consider the simple example of a process in which a die is thrown over and over again.

In order to determine to what extent, the die might be biased, we employ a robot (a machine) to throw the die a very large number of times, say n times, note the results, and output the relative frequencies $f(r)$ of the scores $r = 1, \ldots 6$. These are the relative frequencies we would *expect* to be present in any large number of throws of the die, conducted at any time.

Now, we can use the method of maximum entropy to find the probability distribution $p(r)$ $(r = 1, \ldots 6)$ that rationally represents our degrees of belief in each possible outcome of any *single* throw of the die. We will find

$$p(r) = f(r)$$

We first can calculate the joint probability distribution $p(r_1, r_2, \ldots r_m)$ over the scores that would occur in another large set throws, say m throws, conducted at another time.

Here, as in the last section, each variable, say r_1, takes values 1,2 ... 6.

To find $p(r_1, r_2, \ldots r_m)$, we maximise its entropy under constraints representing our knowledge of the frequency distribution $f(r)$ that arose in the original set of n throws.

As a function of the outcome variables r_k $(k = 1, \ldots n,\ r_k = 1, \ldots 6)$ the relative frequency of a score of six in the m throws, is, by definition

$$\frac{1}{m}\sum_{k=1}^{m}\delta_{r_k,6} \qquad \left(\ \delta_{r_k,6} = \begin{cases} 1 & r_k = 6 \\ 0 & r_k \neq 6 \end{cases}\ \right)$$

We therefore impose the requirement that the *expected* relative frequency of the score six in the set of m throws, be $f(6)$. That is, we impose the constraint

$$\sum_{r_1=1}^{6}\ldots\sum_{r_m=1}^{6}\left(\frac{1}{m}\sum_{k=1}^{m}\delta_{r_k,6}\right)p(r_1,\ldots r_m) = f(6)$$

(Here $\sum_{k=1}^{m}\delta_{r_k,6}$ equals $\delta_{r_1,6} + \delta_{r_2,6} + \cdots \delta_{r_m,6}$, and is, of course, a function of the $r_1, \ldots r_m$.)

We have similar expressions for the expected relative frequencies of each other score. So, the six constraints to be imposed on $p(r_1, \ldots r_m)$, are

$$\sum_{r_1=1}^{6} .. \sum_{r_m=1}^{6} p(r_1, \ldots r_m) \frac{1}{m} \sum_{k=1}^{m} \delta_{r_k,r} = f(r), \quad r = 1,2 \ldots 6$$

Since the sum of the $f(r)$ for $r = 1,2 \ldots 6$, is 1, the constraint expressing the required normalisation of the distribution $p(r_1, \ldots r_m)$ is already covered by the six constraints.

Abbreviating $p(r_1, r_2, \ldots r_m)$ to p and the sums over $r_1, r_2, \ldots r_m$ to Σ, we have then, to maximise

$$-\sum p \ln p$$

Subject to the 6 constraints

$$\sum p \frac{1}{m} \sum_{k=1}^{m} \delta_{r_k,r} = f(r), \quad r = 1,2 \ldots 6$$

To find $p(r_1, r_2, \ldots r_m)$, we first maximise (with no constraints) the expression

$$-\sum p \ln p + \sum_{r=1}^{6} \mu(r) \left(\sum p \frac{1}{n} \sum_{k=1}^{m} \delta_{r_k r} \right)$$

where the $\mu(r)$, $r = 1,2 \ldots 6$ are multipliers attached to each constraint. This can be rewritten as

$$-\sum p \ln p + \sum \sum_{r=1}^{6} \mu(r) \left(p \frac{1}{m} \sum_{k=1}^{m} \delta_{r_k r} \right)$$

Setting the variation of this under $p \to p + \delta p$, we find

$$-\ln p - 1 + \sum_{r=1}^{6} \mu(r) \left(\frac{1}{m} \sum_{k=1}^{m} \delta_{r_k r} \right) = 0$$

So,

$$p \sim \exp \left(\sum_{r=1}^{6} \mu(r) \left(\frac{1}{m} \right) \left(\delta_{r_1 r} + \delta_{r_2 r} + \cdots \delta_{r_m r} \right) \right)$$

Hence

$$p(r_1, r_2, \ldots r_m) \sim p_1(r_1)\, p_2(r_2) \ldots p_m(r_m)$$

where $\sim$ stands for 'to within a constant normalisation factor'. Clearly

$$p_1(r_1) \sim \exp \left(\mu(r_1) \frac{1}{n} \right), \qquad p_2(r_2) \sim \exp \left(\mu(r_2) \frac{1}{n} \right), \ldots \text{etc}$$

Hence $p(r_1, r_2, \ldots r_m)$ factors, making the variables $r_1, r_2, \ldots r_m$, i.e. the scores in each throw, logically independent. Furthermore, since the same function $\mu(r)$ occurs in each factor. the individual factors $p_1(r_1)$, …etc. are the same function $p(r)$ of their variable. That is

$$p_1(r) = p_2(r) = \cdots = p_n(r) = p(r)$$

Imposing the six constraints on $p(r_1, r_2, \ldots r_m)$ we have the equation

$$\sum_{r_1,..r_m} p_1(r_1)\, p_2(r_2) \ldots p_m(r_m) \frac{1}{m} \sum_{k=1}^{m} \delta_{r_k,r} = f(r)$$

With just the first term $\delta_{r_1,r}$ in the sum from $k = 1$ to m, the LHS is

$$p_1(r) \frac{1}{n} \sum_{r_2..r_m} p_2(r_2) \ldots p_m(r_m)$$

Here, the summation factors into sums all equal to 1. The other terms $\delta_{r_2,r}$, $\delta_{r_3,r}$, …etc. make contributions $p_2(r)/n$, $p_3(r)/n$,…etc. The result is

$$p_1(r) \frac{1}{n} + p_2(r) \frac{1}{n} + \cdots + p_n(r) \frac{1}{n} = f(r)$$

As all the terms on the LHS are equal to $p(r)/n$, we obtain

$$p(r) = f(r)$$

as we said we would.

So, the probability distribution over the possible scores in any single throw is equal to the relative frequency of occurrence of those scores in the original n throws, or equal to the expected relative frequency in any large number of throws.

We have thus shown how probabilities can sometimes equal expected frequencies.

18. The determination of unknown parameters in a prior probability distribution

When we use the method of transformation groups or the method of maximum entropy to derive a prior probability distribution, we often establish only the *form* of the distribution. Our derived distribution may contain constant parameters whose values we do not know.

A prior distribution *without* unknown parameters can often be tested by repetition of the process in question. Then, probabilities of system properties are converted into measurable relative frequencies and the 'correctness' of the prior distribution may then be confirmed.

If the prior distribution contains unknown parameters, the frequencies depend on the values of those parameters and vice versa. So, having observed the relative frequencies, the values of those parameters may be chosen so as to reproduce those frequencies. If only

one choice fits the bill, the ‘correct’ values of the parameters have been found.

If it turns out that no parameter values seem to reproduce the relative frequencies, then we should reconsider our derivation of the parameterised prior distribution to see if some important knowledge has been left out, or if something new can be inferred about the process considered and used to add a further constraint in the derivation of the prior distribution.

Sometimes the evaluation of the unknown parameters can be done in a more rational manner. This is possible if probability theory can provide us with a fully specified prior probability distribution over the possible values of the *parameters*.

Starting out with this prior, we sharpen it, in the Bayesian manner, by taking into account the observed frequencies of outcomes in n repetitions of the process. With n large enough, the posterior distribution should become very sharp, and we take, as the best choice of parameter values, those giving the maximum probability in the posterior distribution.

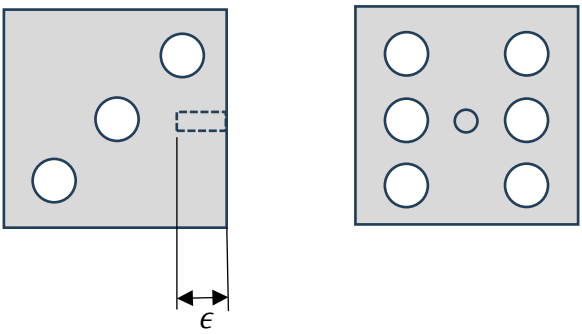
In the case of a single parameter taking one or other of possible *discrete* values, the probability of error in claiming the parameter’s value is the one maximising the posterior distribution, can be calculated by summing

the probabilities on either side of the maximum of the posterior distribution.

In the case of a single parameter taking one or other value in a *continuous* range, the probability of error in claiming its value lies in a specified narrow band around the maximum value of the posterior distribution, can be found by integrating the distribution over either side of the band.

19. Example of the rational evaluation of an unknown parameter

Suppose a manufacturer makes dice all of the same size and all of the same material, but deliberately alters some of them to give them a slight bias in favour of a score of 6. This is achieved by drilling a hole of a certain (small) diameter, into the centre of the face marked six, so displacing the centre of gravity of the die in the direction of the opposite face (i.e. the face marked one).



Some dice are thus made with different degrees of bias set by varying the small depth ϵ of the drilled hole.

Let us set ourselves the problem of calculating the probability distribution $p(r_1, \ldots r_n)$ over the scores in n throws of one of the manufactured dice.

To do the calculation, we use the method of maximum entropy, taking as constraints the requirement of normalisation,

$$\sum_{r_1=1}^{6} \ldots \sum_{r_n=1}^{6} p(r_1, \ldots r_n) = 1$$

and (what seems appropriate) the requirement that the difference between the expected relative frequency of a

score of 6 and the expected relative frequency of a score of 1, be proportional to the depth ϵ of the drilled hole.

We noted in Section 17 that, as a function of the r_k, the *actual* relative frequency of a score of six in the n throws is

$$\frac{1}{m}\sum_{k=1}^{m}\delta_{r_k,6}$$

We therefore set, as the second constraint, the requirement[13]

$$\sum_{r_1=1}^{6}\dots\sum_{r_n=1}^{6}p(r_1,\dots r_n)\left(\frac{1}{n}\sum_{k=1}^{n}\delta_{r_k,6}-\frac{1}{n}\sum_{k=1}^{n}\delta_{r_k,1}\right)=2\xi$$

where $\xi=\frac{1}{2}\alpha\epsilon$, α being a constant characteristic of any of the dice made by the manufacturer before a small hole is drilled into the centre of the side marked six.

To apply the method of maximum entropy, to calculate the form of $p(r_1,\dots r_n)$, we need to maximise

$$-\sum_{r_1=1}^{6}\dots\sum_{r_n=1}^{6}p(r_1,\dots r_n)\ln p(r_1,\dots r_n)$$

[13] We are here following the technique employed by Jaynes. See p. 258ff in 'E.T. Jaynes: Papers on Probability, Statistics and Statistical Physics' Ed. R. D. Rosenkrantz, Kluwer Ac. Pub. 1989.

subject to the constraints.

We leave the details of the calculation to the reader, as an exercise, and quote only the result, which is that

$$p(r_1,\ldots r_n) = p(r_1)p(r_2)\ldots p(r_n)$$

where the factors are all the same function $p(r)$ of their arguments, and

$$p(1) = \frac{1}{6} - \xi$$
$$p(2) = p(3) = p(4) = p(5) = \frac{1}{6}$$
$$p(6) = \frac{1}{6} + \xi$$

assuming $\xi \ll 1$.

This then, is our calculated prior probability distribution $p(r)$ over the possible scores in one throw of the die. It contains the unknown parameter ξ.

As we noted in Section 18, we can estimate ξ by conducting a large enough number of throws of the die. Then the probabilities and the relative frequencies of scores are the same, and ξ can be set equal to one-half the difference between the observed relative frequencies of a score of 6 and a score of 1 in the n throws.

Rational calculation of the parameter ξ.

The parameter ξ is equal to $\frac{1}{2}\alpha\varepsilon$, where, as we have said, α is a constant; a constant characteristic of any of the dice made by the manufacturer before small holes are drilled into them.

On the other hand, ε is variable in the sense that its value depends on how deep the die makers choose to drill the hole. Since ε is a length, it has a natural measure. As α is a simple constant, the dimensionless variable $\xi = \alpha\varepsilon$ also has a natural measure.

We have then, in ξ, a variable (with a natural measure) that can be considered to lie between 0 and a value that reflects the upper limit to the bias the die makers choose to bring about. Suppose they tell us that the dice they produce have various degrees of bias, but never so much as to make the difference in the expected relative frequencies, of a store 6 and a score of 1, greater than 0.2. Then our rational prior distribution $P(\xi)$ over ξ has the uniform density between $\xi = 0$ and $\xi = 0.1$ and is zero for $\xi > 0.1$. So

$$P(\xi) = \begin{cases} 10 & 0 < \xi < 0.1 \\ 0 & \xi > 0.1 \end{cases}$$

Now suppose we throw the die n times. Then, for a given value of ξ, the probability of r sixes in the n throws is

$$P(r|\xi) = {}_nC_r p^r q^{n-r}$$

where p is the probability of a score of 6 in a single trial, and q is the probability of a score different from 6. That is

$$p = \frac{1}{6} + \xi \text{ and } q = \frac{5}{6} - \xi$$

Using Bayes' theorem, we have, for the probability density $P(\xi|r)$ over ξ for given r, the truncated distribution

$$P(\xi|r) = \frac{P(r|\xi)P(\xi)}{P(r)} \sim \begin{cases} p^r q^{n-r} & 0 < \xi < 0.1 \\ 0 & \xi > 0.1 \end{cases}$$

where $P(r)$ is the probability of r sixes when we don't know ξ, and the symbol $\sim$ denotes 'equal to within a constant normalisation factor'. Here, of course, p and q are the simple functions of ξ noted above, so our posterior probability distribution over ξ is

$$P(\xi|r) \sim \begin{cases} \left(\frac{1}{6} + \xi\right)^r \left(\frac{5}{6} - \xi\right)^{n-r} & 0 < \xi < 0.1 \\ 0 & \xi > 0.1 \end{cases}$$

Our best guess ξ_0 of the value of ξ is the value that maximises this posterior distribution.

Instead of working with $P(\xi|r)$, it is a little easier to calculate our best guess p_0 of the value of p and the likely error in it, using the probability distribution

$$P(p|r) \sim \begin{cases} p^r(1-p)^{n-r} & \frac{1}{6} < p < \frac{1}{6} + 0.1 \\ 0 & p > \frac{1}{6} + 0.1 \end{cases}$$

over p. Then, of course, our best guess ξ_0 for the value of ξ will equal our best guess p_0 for the value of p minus one-sixth.

We will show that for large enough values of r and n, the distribution $P(p|r)$ becomes a normal distribution effectively confined to the region $\frac{1}{6} < p < \frac{1}{6} + 0.1$. So, the truncation of $p^r(1-p)^{n-r}$ will not be necessary.

Differentiating with respect to p and setting the result equal to zero, we see the maximum of $P(p|r)$ is at

$$p = p_0 = \frac{r}{n}$$

The logarithm $L(p)$ of $P(p|r)$ is

$$L(p) = r \ln p + (n-r) \ln(1-p)$$

We can expand this function about the value p_0 that maximises it. By Maclaurin's theorem

$$L(p) = L(p_0) + \frac{\partial L}{\partial p}(p - p_0) + \frac{1}{2}\frac{\partial^2 L}{\partial p^2}(p - p_0)^2 \ldots$$

where the derivatives are evaluated at $p = p_0$. The first derivative vanishes.

Leaving out powers of $(p - p_0)$ greater than the second we find

$$P(p|r) \sim \exp(L(p)) \sim \exp\left(-\frac{(p - p_0)^2}{2\sigma^2}\right)$$

where

$$\sigma^2 = \frac{p_0(1 - p_0)}{n}$$

On account of the simple relation $p = \frac{1}{6} + \xi$ between p and ξ, our posterior distribution $P(\xi|r)$ over ξ is the normal distribution

$$P(\xi|r) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(\xi - \xi_0)^2}{2\sigma^2}\right)$$

where

$$\xi_0 = p_0 - \frac{1}{6}$$

and the standard deviation σ remains the same, that is

$$\sigma = \sqrt{\frac{p_0(1-p_0)}{n}}$$

For large enough n, the normal distribution $P(\xi|r)$ is, as we have said, effectively confined to the range $0 < \xi < 0.1$. This is illustrated here.

Figure 4

We want to claim ξ lies in the range:

$$\xi_0 - \delta < \xi < \xi_0 + \delta$$

where δ is a small number chosen by us; and we want to calculate the probability that this claim is wrong. (This probability is equal to the area A in the Figure.) To do this we express δ in terms of σ, and use tabulated, and well-known, values of the area B in the above Figure. Thus, with $\delta = z_c\sigma$, we have

z_c	B	A
1	.6827	.3173
2	.9545	.0455
3	.9973	.0027

So, for example, if the die is thrown $n = 10{,}000$ times, and we find $r = 2366$, then, our best estimate of the value of p is that which maximises $P(p|r)$, i.e.

$$p_0 = \frac{r}{n} = \frac{2366}{10000} = 0.2366$$

Our best estimate of the value of ξ is

$$\xi_0 = p_0 - \tfrac{1}{6} = 0.0700$$

and

$$\sigma = \sqrt{\frac{p_0(1-p_0)}{n}} = 0.00425$$

Choosing $z_c = 1$, we can claim $\xi_0 - \sigma < \xi < \xi_0 + \sigma$, i.e.

$$0.0657 < \xi < 0.0742$$

with a probability of error A equal to 0.3173. Or, choosing $z_c = 2$, we can claim $\xi_0 - 2\sigma < \xi < \xi_0 + 2\sigma$

$$0.0615 < \xi < 0.0785$$

with a probability of error equal to 0.0455.

To increase accuracy, we need to throw the die many more times. If, for example, $n = 1{,}000{,}000$ and we get $r = 243791$, then, repeating the above calculations, we can claim (taking $z_c = 2$), that

$$0.0763 < \xi < 0.0780$$

with a probability of error equal to 0.0455.

20. The additive property of the information entropies of logically independent distributions

Let us suppose we have two processes labelled 1 and 2, and claim to hold knowledge Y_1 in relation to the first and Y_2 in relation to the second. Suppose, too, that we have calculated probability distributions $P(x_i|Y_1)$ and $P(y_j|Y_2)$ over the possible outcomes x_i $(i = 1,..n)$ of the first process, and y_j $(j = 1,\ldots m))$ of the second.

Let us suppose also, that we have no reason to think our knowledge Y_1 somehow contributes to our knowledge of process 2, or vice versa. That is, let us suppose the probability distributions are logically independent.

Now, under knowledge Y_1Y_2 of the two processes together, the probability $P(x_iy_j|Y_1Y_2)$ is the product of the probabilities $P(x_i|Y_1)$ and $P(y_j|Y_2)$.

Thus, with $p_{ij} = P(x_i y_j | Y_1 Y_2)$, $p_i = P(x_i | Y_1)$ and $p_j = P(y_j | Y_2)$, we have

$$p_{ij} = p_i p_j$$

The information entropy of the processes viewed together as one, is

$$\mathcal{H} = -\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij} \ln p_{ij}$$

Replacing p_{ij} by $p_i p_j$, the entropy becomes

$$-\sum_{i=1}^{n}\sum_{j=1}^{m} p_i p_j \ln p_i\, p_j = -\sum_{i=1}^{n}\sum_{j=1}^{m} p_i p_j \left(\ln p_i + \ln p_j\right)$$
$$= -\sum_{i=1}^{n} p_i \ln p_i - \sum_{j=1}^{m} p_j \ln p_j$$

which is the sum of the entropies of the two distributions.

This result clearly generalises to the case of any number of logically independent distributions. It means that our ignorance of independent system outcomes is additive, as we would expect. It strengthens the idea that information entropy, as defined in Section 8, serves as the legitimate measure of our ignorance as to which one of several possible outcomes of a process actually occurs.

In the case of continuous distributions where the information entropy is definable as an integral, the above proof of the additivity of information entropy under logical independence goes through just as well.

If, for example, $p_1(x)$ and $p_2(y)$ are probability distributions of continuous variables relating to logically independent outcomes of a process 1 and a process 2, the probability density for the two processes considered as one is $p(x,y) = p_1(x)p_2(y)$ in the pair of variables x and y. If $m_1(x)$ and $m_2(y)$ are the measure densities relating to x and y, then

$$m(x,y) = m_1(x)m_2(y)$$

will be the measure density in the product space.

It easily follows that the information entropy

$$\mathcal{H} = -\int p(x,y) \ln \frac{p(x,y)}{m(x,y)} dxdy$$

for the two processes considered as one, is the sum of the information entropies of the individual processes.

This concludes our account of rational Bayesian probability theory. We turn now to examples of its use in particular areas of theoretical physics.

Classical Statistical Thermodynamics

The objective of classical statistical thermodynamics is twofold. To derive the equations of state of gases, liquids and solids, and to account for the zeroth, first and second laws of ordinary thermodynamics.[14] This is to be done by classical mechanical modelling of thermodynamic systems as collections of particles (representing atoms), which may bunch together to form molecules on account of inter-particle forces.

As we employing classical mechanics, we could, in principle, label the particles and get to know their positions and velocities at all times. However, because the number of the particles is extraordinarily large, this is practically impossible, and we have instead, to resort to statistical methods of modelling.

E. T. Jaynes was the one who made clear the great simplicity in concepts and method afforded by a rational Bayesian approach to the statistical modelling.

To provide an account of Jaynes' theory, we start by defining various terms.

[14] By 'ordinary thermodynamics' we mean the classical (macroscopic) theory of thermodynamics, as it was before the formulation of the third law of thermodynamics and the study of states of very low temperature.

By a ‘closed system’ or ‘isolated thermodynamic system’ we will mean material contained in a rigid vessel of fixed dimensions and negligible thermal capacity.

We usually work with closed systems of homogeneous material of a single phase.

Unless otherwise stated, a closed system is taken to be in a state of macroscopic quiescence, when any macroscopic motion (flow or vibration) of the material, any chemical interactions in it, and any temperature gradients in it have died away.[15] We then have a closed system ‘in a state of thermodynamic equilibrium’.

The vessel containing one closed system may be brought into contact with another to allow heat transfer between them; or, by way of pistons in the walls of the vessels, one system may perform work on the other. Actions such as these are said to be brought about by an ‘agent’ who effortlessly releases latches or operates switches that control the way systems interact, so as to bring about specified thermodynamic changes. This may include controlling the way a thermodynamic system might perform work on a purely mechanical object such as a spring or weight suspended from a pulley.[16]

---

[15] When reasoning in either ordinary thermodynamics or statistical thermodynamics, we take the dying away of inhomogeneities to be a known property of systems left alone for a sufficiently long time.

[16] During such changes the thermodynamic systems are, of course, not closed.

The effects of gravitational fields on the thermodynamic states of closed systems are supposed to be negligible.

In applying Bayesian probability theory to derive probabilities over the possible states of motion of the particles of a closed system, we take as our prior knowledge, certain macroscopic properties of the system we might reasonably claim to know in practice; and certain detailed properties regarding the particles, postulated for the purpose of modelling. These might include certain macroscopic properties of the system, such as its volume and expected energy, as well as the masses and numbers of the particles making up the system, the potentials of interaction between them and between them and the vessel wall. We may claim to know these exactly, for as we have said before, Bayesian modelling of a physical system involves idealisation of both the physical form of a system and our (i.e. the observer's) knowledge of it.

As in ordinary (non-Bayesian) statistical thermodynamics, by applying the laws of classical mechanics and the laws of probability, the expected values of *all* the macroscopic (thermodynamic) properties of a modelled system can be calculated; and the probability of error in these expected values can be extremely small.

15. Liouville's theorem

We begin our account of the statistical theory by recalling Liouville's theorem. This relates to the probability distribution over the phase space in which the possible motions of the particles making up a system may be represented.

Let the q_i $(i = 1,2,\ldots s)$ be coordinates specifying the positions of all the particles making up the system. Here, s is the number of degrees of freedom of the particle system, which is three times the number of particles. The q_i could be the Cartesian coordinates of the particles, q_1, q_2, q_3 being the x, y, z coordinates of the first particle, q_4, q_5, q_6 those of the second …etc. But they could be any alternative coordinates q_i' in one-to-one relation to the q_i as specified by transformation equations

$$q_i' = f_i(q_1, q_2, \ldots, q_s) \quad i = 1,2,\ldots s$$

In any event, we refer to the q_i as 'configuration coordinates' that specify, one way or another, the momentary positions of all the particles of the system.

For any particular choice of configuration coordinates, there will be corresponding components of generalised momenta p_i $(i = 1,2,\ldots s)$ defined in terms of the Lagrangian of the system, as explained in textbooks on analytical mechanics.

It will be convenient to form a $2s$-dimensional Euclidean space, a point in which is specified by coordinates in a rectangular Cartesian coordinate frame; the coordinates being $p_1, \ldots p_s,\ q_1, \ldots q_s$. This $2s$-dimensional Euclidean space is called 'phase space'. It is unbounded in all directions associated with the p coordinates because these may take values from $-\infty$ to ∞. The phase space is, however, bounded in all directions associated with the q coordinates, on account of the system being confined to a finite region of space.

The particles of the system will generally be in motion under a system potential $V(q_1, \ldots q_s)$. This is a function of the configuration coordinates that is not explicitly dependent on the time. One part of this potential accounts for repulsive forces between the particles and the wall of the vessel in which the system is contained. This part ensures the particles are confined to the region of space inside the vessel. The remaining part of the potential accounts for the forces of interaction between the particles. These are supposed to be repulsive when the particles are far from each other or very close to each other, but attractive otherwise. As a function of the distance d of separation between two particles, their inter-particle potential V takes the form shown in the figure.

Figure

By so supposing an inter-particle potential of this form, it is possible to model chemical reactions and changes in phase. For then, pairs (or triplets, ...etc.) of particles of a gas, may, under the right conditions, group together to form molecules. Or, under the right conditions, all the particles may group tightly together to form a solid, or loosely together to form a liquid.

The shape and volume of the region of fixed space occupied by the system will normally be considered constant in time. But we will have need, on occasion, to allow variation of the volume and shape, i.e. variation of the system boundary, in the course of a controlled thermodynamic process. Under such variation, we will stick to the same configuration coordinate frame to specify the positions of the particles in the same unchanging way. Only the *range* of possible values of some or all of the configuration coordinates $q_1, \ldots q_s$ may change under variation of the system boundary.

The Cartesian coordinates of the point representing the momentary mechanical state of a system in phase space are, of course, functions of time. Their rates of change are given by Hamilton's equations

$$\dot{q}_i = \frac{\partial H}{\partial p_i} \qquad \dot{p}_i = -\frac{\partial H}{\partial q_i} \qquad (i = 1, \ldots s)$$

where the Hamiltonian $H(p_1, \ldots p_s, q_1, \ldots q_s)$ is the system energy as a function of the coordinates and generalised momenta. The potential function $V(q_1, \ldots q_s)$, and the Hamiltonian are not explicitly dependent on the time (unless the system boundary is being varied).

We can picture a swarm of points filling phase space each moving in accordance with Hamilton's equations. These points simultaneously represent all possible natural motions of the system.

Consider the points which, at time t, lie inside a vanishingly small ($2s$-dimensional) volume element $d\tau$ of the phase space. As time passes, the points in $d\tau$ (representing the mechanical state at time t) move. They stay close to each other and form, together, a volume element $d\tau'$ at a later time t'. Possible states at one time are in this way connected with possible states at another.

So if, under certain knowledge of the system's state of motion at time t, we hold a probability that the state representative point lies in $d\tau$ at time t, we must hold the *same* probability for it to lie in $d\tau'$ at time t'. In terms of probability density in phase space, this means that $\rho d\tau$ must equal $\rho' d\tau'$ where ρ and ρ' are the local probability densities at the volume elements at times t and t'.

Now the swarm of representative points flow like a fluid in phase space, and the rate of change of the volume of any small volume element moving with them is proportional to the divergence of the flow velocity distribution.

In 3-dimensional space a volume element dV moving with a flow, grows at the rate $\nabla . \boldsymbol{v}\, dV$ where $\boldsymbol{v}$ is the 3-dimensional flow velocity vector field, and

$$\nabla . \boldsymbol{v} = \frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y} + \frac{\partial v_z}{\partial z}$$

In the present case, the divergence of the flow of representative points in the $2s$-dimensional phase space is, likewise

$$\sum_{i=1}^{s} \left(\frac{\partial \dot{q}_i}{\partial q_i} + \frac{\partial \dot{p}_i}{\partial p_i} \right)$$

But by Hamilton's equations, this is zero.

We are therefore led to the conclusion that volume elements in the phase space retain their volume as they move. Accordingly, the probability density $\rho(p_1, \ldots p_s, q_1, \ldots q_s)$ at any one representative point must stay constant as this point moves through phase space. This is Liouville's theorem.

16. Properties of the probability density function

In forming a mechanical model of a closed system in thermodynamic equilibrium, we hope, that under certain limited knowledge of macroscopic system properties, we may have a calculable probability density $\rho(p_1,\dots p_s,q_1,\dots q_s)$ giving the probability

$$\rho(p_1,\dots p_s,q_1,\dots q_s)dp_1,\dots dp_s dq_1,\dots dq_s$$

for the representative point to lie in the infinitesimal cuboid $dpdq = dp_1,\dots dp_s dq_1,\dots dq_s$ of phase space.

The density function must of course satisfy

$$\int\dots\int \rho(p_1,\dots p_s,q_1,\dots q_s)dp_1,\dots dp_s dq_1,\dots dq_s = 1$$

where the integration extends over the whole of occupied phase space. For short, we may write this as

$$\int \rho(p,q)dpdq = 1$$

Now our limited knowledge of the detailed dynamics of a closed thermodynamic system is based on its macroscopic properties, and is captured by our density function $\rho(p,q)$. As the system is in thermodynamic equilibrium, the density function $\rho(p,q)$ is naturally *time-independent*.

By Liouville's theorem the value of $\rho(p,q)$ at a representative point (p,q) must also stay constant as this point moves through phase space.

These two requirements of the function $\rho(p,q)$ are certainly satisfied by the energy function $E(p,q)$ of the (closed) system. This suggests, that under thermodynamic equilibrium, $\rho(p,q)$ might be a function $f(E(p,q))$ of $E(p,q)$.

17. The microcanonical distribution

Key to forming a statistical description of a closed homogeneous system in thermodynamic equilibrium, is the derivation of a prior probability density $\rho(p,q)$ based on knowledge only of certain macroscopic mechanical properties of the system. We will claim to possess exact knowledge of the shape and volume of the system and certain knowledge concerning its energy.

Since the mechanical energy $E(p,q)$ includes a system potential function, it is (like potentials generally) specified only to within an additive constant. In claiming to know something about the value of a system's energy we have then to fix this constant. That is, we have to define the energy in a way that gives it a definite or 'absolute' value, otherwise our claim is empty.

The absolute value of the energy of a system

We define the absolute energy of a closed homogeneous system in thermodynamic equilibrium by considering states of the system in which the momenta of the particles making it up are all zero. The system is then in the solid phase. But it may still be under stress (under compression or tension). We set $V(q_1, \ldots q_s)$ equal to zero when this stress is relieved by allowing transfer of mechanical work in or out of the system.

The system potential $V(q_1, \ldots q_s)$ and the system energy $E(p, q)$ are now fully specified and are always positive or zero.

When the particle momenta are all zero, but stress in the system is *not* relieved, the system will have a potential energy $V_0(x)$ greater than zero and dependent on the geometrical form of the system boundary (the shape and volume of the vessel in which the system is contained). Here x stands for the values of a number of variables specifying the geometrical form of the boundary. As we have implied, $V_0(x) = 0$, when the variables x take the values for which there is zero stress in the system.

The microcanonical distribution

One way to establish a probability density $\rho(p, q)$ is to suppose we know that the total mechanical energy of the system lies between two close values E_1 and E_2. The representative point (p, q) then lies in the narrow region

between two $(2s-1)$ dimensional 'surfaces' in the phase space. It will of course remain between these surfaces for all time.

If we claim $\rho(p,q)$ is at one time uniform between the surfaces and zero outside them, $\rho(p,q)$ will, by Liouville's theorem, *remain* uniform between the surfaces and zero outside them, and will provide us with a workable probability density, referred to as the 'microcanonical distribution'.

This 'workable distribution' is, under a certain condition, derivable using rational Bayesian probability theory. At least, it is under a certain condition. We now state this condition and show how it can be justified. The condition is that the volume of any region in the system's phase space must be the natural measure of that region. That is, a natural measure of the set of dynamical states represented by it.

We claim this is so.

For consistency, the volume of a region of phase space must remain the same however we choose to represent the configuration of the particles. That is, it must stay the same under any one-to-one transformation of the $q_1,\ldots q_s$. Only then can we claim that the volume of a region in phase space is a natural measure of the set of dynamical states represented by it.

Now the volume of a region in phase space is, in fact, invariant under a transformation of the foresaid

kind.[17] Therefore, the measure density $m(p_1, \ldots p_s, q_1, \ldots q_s)$ is invariant also.

Making a definite choice of the unit of measure, and letting h denote Planck's constant, we further claim that the measure density takes the value:[18]

$$m = h^{-s} = (2\pi\hbar)^{-s}$$

in the phase space representing the dynamical states of any thermodynamic systems. As we are fixing the unit of measure, there is no 'unimportant arbitrary constant factor' on the RHS.

---

[17] This is shown in many textbooks on statistical thermodynamics.

[18] In adopting this value, we are following Jaynes, (see the Brandeis Lectures p.61). That is, we are anticipating a result in quantum statistical thermodynamics applying in the limit as quantum mechanics passes into classical mechanics. (Accounts of this result are given, for example, by Landau and Lifshitz QM p167, Fowler §2.2, and Tolman p.355.) Put simply, there is, in the classical limit, a constraint on possible knowledge of the values of the configuration coordinates and generalised momenta. The least possible uncertainty in the i^{th} pair of these is subject to the requirement $\Delta p_i \Delta q_i \approx 2\pi\hbar$. So, for a system with *s* degrees of freedom, we might well claim that any phase space is strictly made up of (what are classically speaking) vanishingly small cells $\Delta p \Delta q = \Delta p_1 \Delta q_1 \ldots \Delta p_s \Delta q_s$ each of volume $(2\pi\hbar)^s$. The measure density m is then naturally equal to the number of cells per unit volume of phase space.

As m is supposed to be of this form in *any* phase space, we are thus postulating not just the existence of a constant natural measure density over any one phase space, but also a definite relation between the natural measure densities in different phase spaces (that generally have different s values).

The assumption concerning the natural measure density is not a statistical assumption. It is rather a new law of mechanics not covered by the Newtonian laws of motion nor the analytical mechanics formulation of those laws. It is, however, an assumption which can evidently sit quite harmlessly alongside the classical laws of motion without causing contradiction.

The phase space probability density

$$\rho(p,q) = \begin{cases} \text{constant} & E_1 < E < E_2 \\ 0 & E < E_1 \text{ and } E > E_2 \end{cases}$$

applying under knowledge $E_1 < E < E_2$, with E_1 and E_2 nearly equal, may now be said to be derivable from rational Bayesian probability theory and the laws of mechanics. It is known as the 'microcanonical distribution'.

In most accounts of statistical thermodynamics, the microcanonical distribution is used to derive the more useful 'canonical' distribution by considering the system of interest to be no longer closed but in thermal contact

with a ‘heat reservoir at a definite temperature’. As a result, the energy of the system is no longer independent of the time, even though the probability distribution in its phase-space is still assumed to be time independent. The system and heat reservoir *together* form a closed system in which the energy *is* constant and the microcanonical distribution applies. The distribution in the phase space of the smaller system is then derived in a rather complicated way from the microcanonical distribution of the greater system.

This conventional approach to the derivation of the canonical distribution has its origin in the idea that probability is frequency of occurrence. In this way of thinking, the move from a phase-space distribution based on an effectively fixed energy (i.e. the microcanonical distribution) to a distribution based on an expected energy of many possible energies (i.e. the canonical distribution) is necessary for *physical* reasons. In the first case we have an isolated system at constant energy. In the second, we have a system whose energy is undergoing temporal *fluctuations* on account of its thermal contact with a heat reservoir.

The difficulties in deriving the canonical distribution from the microcanonical distribution, are entirely avoided when employing rational Bayesian probability theory. Then, as we show in the next section, the canonical distribution is a simple consequence of the

principle of maximum information entropy applied directly to an *isolated* thermodynamic system whose *expected* energy is known. There is no need for the system to be in contact with a heat reservoir. We might imagine, instead, that we have somehow given the system an approximately known energy. Being closed, the system, has of course, an actual energy with a definite constant value. We do not know this value exactly, but we can usefully claim to know its *expected* energy. In this Bayesian approach, energy fluctuation is replaced by energy *uncertainty*.

18. The canonical distribution

Derivation of the canonical distribution

Note first, that for economy, the energy function $E(p,q)$ and the probability density $\rho(p,q)$ will sometimes be denoted simply by E and ρ.

When we claim to know only the value of the expected value $\langle E \rangle$, we should (by the method of maximum entropy), rationally adopt the distribution $\rho(p,q)$ that maximises

$$-\int \rho \ln \frac{\rho}{m} dpdq$$

subject to the constraints

$$\int E\rho dpdq = \langle E \rangle, \qquad \int \rho dpdq = 1$$

Carrying out the maximisation by the method of Lagrange, we maximise

$$-\int \rho \ln \frac{\rho}{m} dpdq + \mu \left(\int \rho dpdq - 1 \right) - \lambda \left(\int E\rho dpdq - \langle E \rangle \right)$$

subject to no constraints, then impose the constraints to find the constants μ and λ.

For any small variation $\delta\rho$ in the function ρ, the variation in the above expression is

$$\int \left[-\delta\rho \ln \frac{\rho}{m} - \rho \frac{1}{\rho} \delta\rho + \mu\delta\rho - \lambda E \delta\rho \right] dpdq$$

Setting this to zero for all $\delta\rho$ gives

$$-\ln \frac{\rho}{m} - 1 + \mu - \lambda E = 0$$

or

$$\rho(p, q) = Ae^{-\lambda E(p,q)}$$

where A is a constant related to μ. This is the canonical distribution derived very simply.[19]

Since $\rho(p,q)$ is a function of $E(p,q)$, it satisfies the necessary condition mentioned at the end of Section 16.

The constants A and λ are given by solving the two equations of constraint:

$$\int E(p,q)Ae^{-\lambda E(p,q)}dpdq = \langle E \rangle$$

$$\int Ae^{-\lambda E(p,q)}dpdq = 1$$

Consequences of the above pair of equations can be derived by considering the small changes dA and $d\lambda$ that occur when $\langle E \rangle$ is slightly altered.

Since the RHS of the second equation is independent of $\langle E \rangle$, the variation of its LHS must vanish. This gives

$$dA\frac{1}{A} - \langle E \rangle d\lambda = 0$$

[19] This derivation of the canonical distribution was first given by Jaynes (see paragraph 3 of 'Information theory and statistical mechanics I', Physical Review 106, 1957, 620-630.). Jaynes' derivation is, in fact, of the more general 'grand canonical distribution' in which, in addition to the expected anergy, only the *expected* number of the molecules of the system is supposed to be known.

as the reader may verify. Under the same small variation of $\langle E \rangle$, the first equation gives

$$dA\frac{\langle E \rangle}{A} - \langle E^2 \rangle d\lambda = d\langle E \rangle$$

Hence

$$(\langle E \rangle^2 - \langle E^2 \rangle)d\lambda = d\langle E \rangle$$

and since $\langle E \rangle^2 - \langle E^2 \rangle$ $(= -\langle (E - \langle E \rangle)^2 \rangle)$ is negative, it follows that λ is a monotonic decreasing function of $\langle E \rangle$.

On dividing the first equation for A and λ by the second, we obtain

$$\langle E \rangle = \frac{\int E(p,q)e^{-\lambda E(p,q)}dpdq}{\int e^{-\lambda E(p,q)}dpdq}$$

expressing $\langle E \rangle$ as a function of λ. As this function must be a monotonic decreasing function of λ, there cannot be more than one value of λ, nor of A (which is a monotonic function of λ given by the second equation of constraint).

Because $E(p,q)$ is a positive definite quadratic function of the generalised momenta and therefore goes to infinity as the generalised momenta go to infinity, λ must be positive, otherwise the integrals in the equations of constraint would diverge.

19. The canonical distributions for systems after thermal contact.

Suppose the vessels of two closed quiescent systems with known expected energies, are placed in contact so that energy might pass from one to the other till the composite system is in a quiescent state. Suppose the contact is then ended and no energy has actually passed from one system to the other. Then, the energy of each system is the same as before, and therefore the thermodynamic states of each system are the same as before. In this case the systems are said to be 'in thermal equilibrium with each other'.

In general, the systems are not in thermal equilibrium with each other, and their expected energies after contact are different from what they were before.

In the statistical theory, interaction between systems 1 and 2 while in contact, can be modelled by introducing interaction potentials between particles of system 1 and particles of system 2. Strictly speaking, even if the interaction is weak, only the composite system (that of system 1 and system 2 taken together) is then closed.

Let system 1 have an energy function $E_1(p_1, q_1)$ with s_1 degrees of freedom. Let system 2 have energy function $E_2(p_2, q_2)$ with s_2 degrees of freedom.

The energy function $E(p, q)$ in the phase space of the two systems taken together is the sum of the energy

functions of the systems themselves, plus an interaction potential:

$$E(p,q) = E_1(p_1,q_1) + E_2(p_2,q_2) + v(p,q)$$

Here, p stands for the $(s_1 + s_2)$ momentum variables and q for the $(s_1 + s_2)$ coordinate variables of the composite system.

Before interaction $v(p,q) = 0$ and we suppose we know the expected energies $\langle E_1 \rangle$ and $\langle E_2 \rangle$ of each system, but do not know whether or not they are in thermal equilibrium with each other. When in thermal contact, we allow the composite system to reach a state of thermodynamic equilibrium. Then we know only the energy $\langle E \rangle$ of the composite system. But we may apply the method of maximum entropy, to derive the distribution $\rho(p,q)$ over the variables p and q (just as we did in Section 18). We can thus show that

$$\rho(p,q) = Ae^{-\lambda E(p,q)}$$

the constants A and λ being found from the conditions

$$\int E(p,q)Ae^{-\lambda E(p,q)}dpdq = \langle E \rangle$$

and

$$\int Ae^{-\lambda E(p,q)}dpdq = 1$$

As we have seen in Section 18, these conditions fix the (unique) value of λ in terms of $\langle E \rangle$. So, the distribution $\rho(p, q)$ is fully calculated.

Knowledge of $\langle E \rangle$ is sufficient to make accurate predictions of all macroscopic properties of the composite system, including the energies $\langle E_1 \rangle'$ and $\langle E_2 \rangle'$ of the component systems. These may differ from $\langle E_1 \rangle$ and $\langle E_2 \rangle$.

Under $\rho(p, q)$ the expected energies of system 1 and system 2 are

$$\langle E_1 \rangle' = \int E_1(p_1, q_1) A e^{-\lambda E(p,q)} \, dpdq$$

and

$$\langle E_2 \rangle' = \int E_2(p_2, q_2) A e^{-\lambda E(p,q)} dp_2 dq_2$$

To make the calculation of these energies tractable, we have to make the assumption that the interaction potential has only a negligible effect on our statistical predictions. This is justified by noting that interaction potential $v(p, q)$, acts only between the relatively tiny number of particles near the area of contact of the systems.

This said, we can now write

$$E(p, q) = E_1(p_1, q_1) + E_2(p_2, q_2)$$

and the distribution $\rho(p,q)$ in the phase space of the composite system, factors into normalised parts:

$$\rho(p,q) = \rho_1(p_1,q_1)\rho_2(p_2,q_2)$$

where

$$\rho_1(p_1,q_1) = A_1 e^{-\lambda E_1(p_1,q_1)}$$
$$\rho_2(p_2,q_2) = A_2 e^{-\lambda E_2(p_2,q_2)}$$

On account of this factorisation,

$$\langle E_1\rangle' = \int E_1(p_1,q_1) A_1 e^{-\lambda E_1(p_1,q_1)} dp_1 dq_1$$

and

$$\langle E_2\rangle' = \int E_2(p_2,q_2) A_2 e^{-\lambda E_2(p_2,q_2)} dp_2 dq_2$$

So, under knowledge of $\langle E\rangle$ alone, the systems turn out to be logically independent, and our knowledge is equivalent to knowledge of the (calculated) expected energies $\langle E_1\rangle'$ and $\langle E_2\rangle'$, of system 1 and system 2.

When the systems are separated and become closed again, they retain energies $\langle E_1\rangle'$ and $\langle E_2\rangle'$, so the canonical distributions $\rho_1(p_1,q_1)$ and $\rho_2(p_2,q_2)$, still apply.

20. The distribution over the energy

Using the canonical distribution, it is possible, as is well known, to reproduce the general concepts and laws of ordinary thermodynamics, and to confirm, by calculation, certain thermodynamic equations of state.

That this is possible at all, using the Bayesian approach, requires that supposed knowledge of the expected energy $\langle E \rangle$ of any system, amounts effectively to *exact* knowledge of the system's energy. It is necessary therefore, to show that the probability distribution over the energy E, derivable from the canonical distribution over coordinates and momenta, is very highly concentrated at its expected value $\langle E \rangle$.

That the distribution over the energy E does in fact satisfy this requirement, can be understood as follows.

Suppose the system takes the form of a cube of material (gas, liquid, or solid) in a quiescent state, the expected energy of the system being known.

Let the interior of the cube be partitioned geometrically into a large number n of smaller cubes of equal size and closely packed so as to fill the volume occupied by the system.[20] The number n should be very

[20] We note in passing, that imagining the partitioning of a system into many small (but not too small) parts, helps in setting up the differential equations in the thermo-dynamics of fluids or solids that are inhomogeneous and not in a state of thermodynamic equilibrium. For, on account of its small size, any part is (well enough) homogeneous and in a state of thermodynamic equilibrium

large (say 10^{12}), yet small in comparison with the number of molecules in the system (which will typically be of order 10^{23}). The number of molecules in any one small cube of material is then $10^{23}/10^{12} = 10^{11}$.

For our present purpose, we imagine the partitioning of the system to be done by instantly creating infinitely thin, rigid and thermally insulating boundaries around each small cube of material. This done, we have at once converted a single system into n identical subsystems, each being a closed homogeneous system (gas, liquid, or solid) in thermodynamic equilibrium.

The energy of the whole system remains the same as before, and the energy of each subsystem is the same as the energy that the material within it had just before the appearance of the partitions. (This is not, of course, exactly true owing to the need for new short-range potentials between the molecules of the system and the new boundaries. But, on account of the small number of particles affected (relative to the total number of particles in a subsystem), we take it that this disturbance of the energy can be neglected.)

Now, we may claim to know the expected energy $\langle E_c \rangle$ of any one subsystem. This is the known expected

---

during times short compared to the characteristic time of variation of the whole system. Momentarily then, the thermodynamics of each part is the thermodynamics of a homogeneous system in equilibrium, and the usual equations of state can be taken to apply to it.

energy $\langle E \rangle$ of the whole system divided by the number n of subsystems:

$$\langle E_c \rangle = \frac{1}{n} \langle E \rangle$$

Knowing $\langle E_c \rangle$, our knowledge of the dynamical state of any one subsystem is expressed by a canonical distribution over the coordinates and momenta of the particles making it up.

Associated with any canonical distribution is a distribution over the *energy*. With regard to the whole system before partitioning, we have a distribution $\rho(E)$, with a known mean value $\langle E \rangle$, and some standard deviation σ. Similarly, with regard to each subsystem, we have a distribution $\rho_c(E_c)$, with a known mean value $\langle E_c \rangle$, and some standard deviation σ_c.

We have seen that, under knowledge only of the total energy of non-interacting systems, their canonical distributions, are logically independent (see Section 19). So, the actual energies E_i $(i = 1,2, \ldots n)$ of the subsystems are random variables independent of one another. Since E is the sum of these random variables, the central limit theorem tells us that the distribution $\rho(E)$ is a normal distribution with mean

$$\langle E \rangle = \langle E_1 \rangle + \langle E_2 \rangle + \cdots \langle E_n \rangle = n \langle E_c \rangle$$

and variance

$$\sigma^2 = \sigma_1^2 + \sigma_2^2 + \cdots \sigma_n^2 = n\sigma_c^2$$

So, the coefficient of variation of $\sigma/\langle E\rangle$ of $\rho(E)$, is

$$\frac{\sigma}{\langle E\rangle} = \frac{\sqrt{n}\sigma_c}{n\langle E_c\rangle} \sim \frac{1}{\sqrt{n}}$$

assuming that, at worst, σ_c is of the same order as $\langle E_c\rangle$.

Since n is 10^{12}, the coefficient of variation is at most 10^{-6}, so the system energy is indeed close enough to the expected energy $\langle E\rangle$ we claim to know.

In the phase space of any individual system the energy function $E(p,q)$ is positive at any point. At one point P it takes its minimum possible value $E_{\min}$. Far from the origin as is possible to go, $E(p,q)$ becomes infinite, either because the generalised momenta go to infinity infinitely far from the origin, or because a boundary in phase space is reached (with respect to the configuration coordinates) at which the system potential energy goes to infinity. Hence throughout the phase space $E_{\min} \leq E(p,q) < \infty$; and the locus of points with energy equal to the expected energy $\langle E\rangle$ forms a closed 'surface' S around the point P of minimum energy.

Now, it might seem that the canonical distribution $\rho(p,q)$ (in contrast to the microcanonical distribution) is spread too widely in the phase space. It is not strongly concentrated near the closed 'surface' S as we would expect.

The canonical distribution $\rho(p,q)$ *does*, it is true, fall exponentially as we move *outward* from the surface S, in the direction of *increasing* energy, but it does not fall as we move *inward*. It remains significant everywhere in the region τ of phase space enclosed by S. So, it might seem that the distribution $\rho(p,q)$ gives significant weight to all points (and hence to all energy values) between E_{min} and $\langle E \rangle$.

However, almost all the 'volume' of region τ is in a layer between energy values $E = \langle E \rangle$ and $E = \langle E \rangle - \delta$ where δ is extremely small compared to $\langle E \rangle$. This is because of the very high dimension of the phase space.[21] So, though the probability *density* may be sizable everywhere in τ, the *probability* of being anywhere in τ,

[21] For example, the volume (or 'content') of a hypersphere of radius r centred at the origin of a Euclidean space of m dimensions is $C = 2r^m\pi^{m/2}/(m\Gamma(\frac{1}{2}m))$ (See, for example Kendall p.35). Hence, as a faction of C, the content between two such hyperspheres of radii r and $r+\delta$, is $R = (1+\delta/r)^m = e^{m\ln(1+x)}$ where $x = \delta/r$. Using power series, $R = e^{mx}e^{-mx^2/2}e^{mx^3/3}e^{-mx^4/4}$ ... etc. With $x = 10^{-11}$ and $m = 10^{23}$, we get $R = e^{10^{12}}e^{-10/2}e^{10^{-10}/3}$ ... which rapidly converges to an extremely large number.

other than in an extremely thin layer on the underside of S, is exceedingly small.

Indeed, in every specific case in which an explicit probability distribution over the energy can be calculated, the likely deviation of the energy from its declared expected value $\langle E \rangle$ turns out to be exceedingly small.

Part of the skill in doing theoretical physics lies in the choice of the idealised models set up. This choice must be made in a way that renders tractable the calculation of the physical quantities of interest. When rational Bayesian probability features in our modelling, it is necessary to make idealisations also with regard to our supposed knowledge; to choose this knowledge so as to render tractable the calculation of the probability distributions of interest.

In classical statistical mechanics viewed in the rational Bayesian way, there is no 'correct' or 'actual' probability distribution over system coordinates and generalised momenta, as the interpretation of probability as frequency would lead us to believe. The canonical distribution is but one of many that could be used for the same purpose. The reason for preferring it is one of practicality.

It might seem surprising that, having claimed to know only the expected value $\langle E \rangle$ of a system's energy, the canonical distribution comes back to us to say we may

be confident that this value is extremely close to the actual value of the system's energy, far closer than we would ever have reason to believe! But this is due only to the idealisation we have made in supposing that the expected value $\langle E \rangle$ of the energy is exactly known.

21. Definitions of the internal energy, entropy, and temperature

In order to reproduce ordinary thermodynamics from the statistical theory, we need to define internal energy, entropy, and temperature of a thermodynamic system in terms of concepts in the statistical mechanical theory alone.

Definition of the internal energy

Under the canonical distribution, with its very sharp form, it is natural to claim that the internal energy U of a system is the expected energy $\langle E \rangle$.

Definition of temperature and thermodynamic entropy

The information entropy associated with the canonical distribution of a thermodynamic system is

$$\mathcal{H} = -\int \rho(p,q) \ln((2\pi\hbar)^s \rho(p,q))\, dpdq$$

where $(2\pi\hbar)^{-s}$ is the constant measure density m as postulated in Section 17. Substituting

$$\rho(p,q) = Ae^{-\lambda E(p,q)}$$

we find

$$\mathcal{H} = -\ln((2\pi\hbar)^s A) + \lambda\langle E\rangle$$

The information entropy is therefore a function of the energy $\langle E\rangle$ of the system, parameters A and λ being themselves functions of $\langle E\rangle$.

The change $d\mathcal{H}$ in $\mathcal{H}$ accompanying a change $d\langle E\rangle$ in $\langle E\rangle$ is

$$d\mathcal{H} = -\frac{1}{A}dA + \langle E\rangle d\lambda + \lambda d\langle E\rangle$$

The relation

$$dA\frac{1}{A} - \langle E\rangle d\lambda = 0$$

established in Section 18, results in the cancelation of the first two terms on the right of the equation for $d\mathcal{H}$. Also $d\langle E\rangle$ is the increase dU in internal energy. Hence, we have

$$d\mathcal{H} = \lambda dU$$

Now, while $d\mathcal{H}$ is dimensionless, λ has the dimensions of energy to the power -1 and we write it as

$$\lambda = \frac{1}{kT}$$

where k is a chosen positive quantity of energy and equal to 1.380649×10^{-23} joules (the same for all systems). T is thus a dimensionless quantity, and can replace the parameter λ. The reason for the choice of the particular value of k (which is Boltzmann's constant) will become apparent later.

The dimensionless quantity T is not a *physical property* of the mechanical system. Like λ, it is a property relating to our knowledge of the system. It is a parameter in our canonical distribution based on the known value of $\langle E \rangle$ and the known shape and volume of the system.

We now define the *thermodynamic entropy* S as k times the information entropy:

$$S = k\mathcal{H}$$

But like T, the thermodynamic entropy is not, for us, a *physical property* of the system. It is a *measure of our ignorance of the detailed mechanical state of the system* expressed as so many units of energy.

When defined as $S = k\mathcal{H}$, the thermodynamic entropy S of a system has a definite absolute value. This is in contrast with the situation in ordinary thermodynamics, where the entropy of a system is defined only to within an additive arbitrary constant.

The absolute value of the entropy (as well as the internal energy) of a system in statistical thermodynamics, has important consequences. The additive properties of entropy are demonstrable as a result, as we will see shortly.[22]

A consequence of the last three equations is the relation

$$\frac{dU}{dS} = T$$

between differentials. It applies when the shape and volume of the system are maintained constant.

On account of the choice made of the numerical value of the constant k in the statistical definition of thermodynamic entropy, T coincides with the absolute temperature in ordinary thermodynamics. This is proved as follows.

We start with the fact that, in ordinary thermodynamics, it is found, empirically, that the ratio PV/n, where P, V and n are the pressure, volume and number of moles of a sufficiently rarefied gas, is independent of its chemical composition. It only depends on the temperature of the gas, as measured, for example, by a mercury thermometer.

[22] We note also, that the absolute value of the entropy allows the calculation of absolute values of the equilibrium constant K in the law of mass action; something that is not possible in ordinary thermodynamics.

It increases monotonically with temperature in exactly the same way for all gases. This is the truer, the more rarefied the gas. Any gas rarefied enough for this property to hold, is called an 'ideal gas', or 'perfect gas'.

These properties of an ideal gas are used to define the 'absolute temperature' of a body. The absolute temperature of a body is taken to be proportional to the value PV/n of an ideal gas in thermal contact with it. Thus

$$\theta = \frac{PV}{nR}$$

The constant R is fixed by setting to 100, the difference between the measured values of PV/n at the temperatures of boiling water and iced water, both at atmospheric pressure. This fixes R at about 8.29 joules.[23] The above equation connecting absolute temperature with pressure and volume constitutes the equation of state of an ideal gas in ordinary thermodynamics.

Now, statistical thermodynamics comes up with the equation of state

$$\frac{PV}{Nk} = T$$

[23] Today R is assigned the value 8.294462618 joules.

for an ideal gas modelled as N identical non-interacting particles moving inside a vessel. (The proof of this is given in Section 25.)

The two equations of state coincide, if, with Avogadro, we claim that one mole of any gas contains the same number N_A, of molecules; or that n moles of any gas contain nN_A molecules.[24] Then the statistical mechanics equation of state can be written as

$$\frac{PV}{nN_Ak} = T$$

It is now clear why we set k, equal to Boltzmann's constant, i.e. to R/N_A. For that way, statistical thermodynamics comes up with exactly the same equation of state as ordinary thermodynamics does, T being identified with the absolute temperature θ.

22. Derivation of the zeroth law of thermodynamics

The zeroth law of thermodynamics can be stated as follows:

[24] Avogadro's number N_A is today assigned the value 6.02214076×10^{23}.

If two closed systems are in thermal equilibrium with a third, they are in thermal equilibrium with each other. [25]

The first step toward establishing this law statistically, is to show that any two closed systems in thermal equilibrium with each other have equal temperatures.

This follows immediately from the theory in Section 19. For, there it was found that if the expected energies of closed systems 1 and 2 are supposed known, and it is also known that the systems are in thermal equilibrium with each other, then before and after thermal contact (wherein energy might have passed from one to the other but didn't) the probability distributions in their phase spaces, are canonical, and take the form:

$$\rho_1(p_1, q_1) = A_1 e^{-\lambda E_1(p_1,q_1)}$$
$$\rho_2(p_2, q_2) = A_2 e^{-\lambda E_2(p_2,q_2)}$$

Since the constant, λ, is the same in each, it follows that the systems have the same temperature. QED

The next step toward establishing the zeroth law is to show the converse of the above. That is, to show that if two thermodynamic systems with known expected

[25] The meaning of systems being 'in thermal equilibrium with each other' has been explained in Section 19.

energies have the same temperature, they are in thermal equilibrium with each other. We prove this as follows.

We are supposing that the systems 1 and 2 have the same temperature. This means the canonical probability densities in their phase spaces have the same value of λ.

Let the systems be put carefully into thermal contact with one another for as long as it takes to ensure thermodynamic equilibrium of the composite system. Let the systems be separated again, and let them stand till they are each in thermodynamic equilibrium. We complete the proof by showing that the energies of the systems are then the same as before thermal contact.

During contact, the composite system is closed. Its expected energy $\langle E \rangle$ stays equal to the sum of the expected energies we knew the component systems had originally, i.e. $\langle E \rangle = \langle E_1 \rangle + \langle E_2 \rangle$.

The value of $\langle E \rangle$ remains the same after the component systems are separated and are in quiescent states. Our knowledge of $\langle E \rangle$ is then the only knowledge we hold about the dynamics of the composite system. But (as explained in Section 19) on the basis of this knowledge, we can calculate the canonical distribution for the composite system after separation of its parts. We write this as

$$\rho(p, q) = A' e^{-\lambda' E(p,q)}$$

where, as explained in Section 18, λ' is related to $\langle E \rangle$ by

$$\frac{\int E(p,q)e^{-\lambda' E(p,q)}dpdq}{\int e^{-\lambda' E(p,q)}dpdq} = \langle E \rangle$$

where

$$E(p,q) = E_1(p_1,q_1) + E_2(p_2,q_2)$$

and λ' is a single valued function of $\langle E \rangle$.

Before the systems are put in thermal contact their distributions are canonical, and in the first of these, the equation relating λ to $\langle E_1 \rangle$ is

$$\frac{\int E_1(p_1,q_1)e^{-\lambda E_1(p_1,q_1)}dp_1dq_1}{\int e^{-\lambda E_1(p_1,q_1)}dp_1dq_1} = \langle E_1 \rangle$$

In the second, the equation relating λ to $\langle E_2 \rangle$ is

$$\frac{\int E_2(p_2,q_2)e^{-\lambda E_2(p_2,q_2)}dp_2dq_2}{\int e^{-\lambda E_2(p_2,q_2)}dp_2dq_2} = \langle E_2 \rangle$$

Adding these two equations, we obtain an equation relating λ to $\langle E_1 \rangle + \langle E_2 \rangle$, which after some rearrangement is

$$\frac{\int (E_1(p_1,q_1) + E_2(p_2,q_2)e^{-\lambda(E_2(p_2,q_2)+E_1(p_1,q_1))}dp_1dq_1dp_2dq_2}{\int e^{-\lambda(E_2(p_2,q_2)+E_1(p_1,q_1))}dp_1dq_1dp_2dq_2}$$
$$= \langle E_1 \rangle + \langle E_2 \rangle$$

This is the same as the equation for λ'; and as the solutions are unique, we have $\lambda' = \lambda$. The temperature of the composite system is therefore the same as the temperature of systems 1 and 2, before they were put in thermal contact.

Since

$$E(p,q) = E_1(p_1,q_1) + E_2(p_2,q_2)$$

the canonical distribution $\rho(p,q)$, calculated above, factors into normalised canonical distributions for each component system.

So, after contact, systems 1 and 2 are again closed and independent. Since $\lambda' = \lambda$ these distributions are the same as the distributions that applied before the systems were brought into thermal contact. The expected energies $\langle E_1 \rangle'$ and $\langle E_2 \rangle'$ after the systems are separated are therefore the same as before.

Hence no energy transfer is expected to have taken place during the time the systems were in contact. That is, when two systems have equal temperatures, they are necessarily in thermal equilibrium with each other. QED

The truth of the zeroth law of thermodynamics follows immediately now. For, being in thermal equilibrium with a third, the temperatures of the two systems must be equal

to that of the third. They are therefore equal to one another. Finally, being of equal temperature, the two systems are necessarily in thermal equilibrium with each other. QED

23. Derivation of the differential form of the first and second laws of thermodynamics

Under knowledge of the expected energy of a system, let us allow gradual (quasi-static) changes to it.

Let x_i $(i = 1,2 \ldots n)$ denote the independent real variables which together specify the geometrical shape and volume of the system. The gradual variation of these variables may change the system's internal energy.

With the x_i held constant, so no work is performed on the system, the system entropy S is a function only of $\langle E \rangle$, i.e. a function only of the internal energy U which could be changed by the slow passage of heat in or out of the system.

In general, S is a function both of U and of the geometric parameters x_i. That is,

$$S = f(U, x_1, \ldots x_n)$$

It follows that, in the neighbourhood of any particular values of S and the x_i, the internal energy U is a function of S and the geometrical parameters. That is,

$$U = U(S, x_1, \ldots x_n)$$

Small quasi-static changes in S and the x_i therefore result in a change

$$dU = \left.\frac{\partial U}{\partial S}\right|_x dS + \sum_{i=1}^{n} \left.\frac{\partial U}{\partial x_i}\right|_S dx_i$$

(x standing for all the terms x_i). Here, the term multiplying dS is just the temperature of the system as established in Section 21, i.e.

$$\left.\frac{\partial U}{\partial S}\right|_x = T$$

The sum must represent the change in U due to external work done on the system, for it vanishes when the dx_i all vanish, and is linear in the dx_i. Accordingly, the $\left.\frac{\partial U}{\partial x_i}\right|_S$ ($i = 1,2 \ldots n$) must be the generalised (external) forces X_i under which the displacements dx_i occur. We thus arrive at

$$dU = TdS + dW$$

where dW is the work done by external forces on the system, given by

$$dW = \sum_{i=1}^{n} X_i dx_i$$

The term TdS in the equation for dU (being the difference between dU and dW) is the heat dQ flowing into the system. Hence the general relation

$$dS = \frac{dQ}{T}$$

which holds whether or not the x_i are constant, i.e. whether or not the volume or shape of the system changes as heat dQ flows into it.[26]

We have thus reproduced the differential form of the first and second laws of thermodynamics.[27]

It is true that we have failed to demonstrate that temperature and entropy are *mechanical properties* of a system. The temperature and entropy are, from the statistical point of view, properties of the canonical probability distribution representing our degrees of belief over the possible microscopic states of the system.

However, this 'failure' is of no consequence. For the first and second laws will lead to the same predictions

[26] It follows from this relation, that if heat is quasi-statically removed from any system its entropy necessarily falls.
[27] Confirmation of the second law in its full form is provided later.

as ordinary thermodynamics does. That is, they will provide the correct mathematical relations between directly measurable mechanical system properties such as volume, pressure and energy. They will also provide the correct mathematical relations between these properties and the temperature and entropy of the system. The temperature and entropy may be converted into directly measurable mechanical forms by means of suitable measuring apparatus. For example, temperature may be converted into the length of a mercury column in a tube. That length is predictable by the statistical theory even though temperature itself is not thought of as a mechanical property.

In the modelling of liquid and gas systems, it can be shown, statistically, that the pressure P arising from molecular motion, is uniform throughout the system; any spatial variations of it are expected to be vanishingly small. So, for liquid and gas systems the work done by external forces is simply $dW = -PdV$ where dV is the increase in volume of the system. The differential form of the first and second laws then takes the well-known form

$$dU = TdS - PdV$$

24. Corrections needed to the definition of entropy to ensure it is an extensive property

The mechanically defined internal energy of a closed system of fixed density must increase in proportion to its volume. The same must be true of its statistically defined entropy. These conditions are required in order that the internal energy and entropy are 'extensive' properties as they are in ordinary thermodynamics.

The internal energy

On account of the extremely large number of particles in the bulk of a system compared to those near the wall, the contribution of particle-wall potential energy is dwarfed by the particle kinetic energy plus the potential energy of particle-particle interaction throughout the volume of the system. So, the internal energy, as we have defined it, is essentially that of the particles in the bulk of the system, and, it is therefore already an extensive property on account of the homogeneity of the system. So, no correction is needed here.

The entropy

In the development of (non-Bayesian) statistical thermodynamics, 'correct Boltzmann counting' was found necessary in order for the entropy of a gas to be an extensive property. In the Bayesian theory, where the entropy is a measure of an observer's ignorance of the dynamical state of the N particles making up the system,

the equivalent correction is achieved by adjusting exactly what the observer is supposed to be ignorant about.

It is not necessary to invoke the quantum mechanics idea that identical particles are indistinguishable. Instead, we suppose they are naturally distinguishable, even though their masses might be the same.

When in a state of thermodynamic equilibrium, the M molecules making up a gas, or the M ions making up a liquid are free to move around and replace one another in the roles they are playing. By 'role' we mean the occupation of a particular position with a particular velocity. (This freedom is not present in a solid thermodynamic system where the particles are bound into a permanent structure.)

To ensure it is an extensive property, we must redefine the information entropy for a gas or liquid at one time, to be the measure of our ignorance of the system's detailed mechanical properties at that time *leaving aside our ignorance of which molecule is playing which role*.

Then, at that time, the probabilities of each of the possible arrangements of the particles among the roles they are playing, are all equal. The information entropy associated with these probabilities is the logarithm of the *number* of possible arrangements (see *Example of the use of Rule 4* in Section 9 of Part 1). The number of possible arrangements is $M!$, so we should subtract $\ln M!$ from the

information entropy of a gas or liquid system, and subtract $k \ln M!$ from its thermodynamic entropy. The information entropy thus becomes

$$\mathcal{H} = -\int \rho(p,q) \ln\left(\left(2\pi\hbar\right)^{s} \rho(p,q)\right) dpdq - \ln M!$$

This can be rewritten as

$$\mathcal{H} = -\int \rho(p,q) \ln\left(\left(2\pi\hbar\right)^{s} M!\, \rho(p,q)\right) dpdq$$

So instead of subtracting $\ln M!$ From the information entropy we could, if we wished, change the measure density from $m = (2\pi\hbar)^{-s}$ to $m' = m/M!$.

Since $\ln M!$ is just a constant for a given system, the change in the information entropy of the system does not affect the derivation of the canonical distribution or any of the differential relations derived from it in Section 23, including the differential form of the first and second laws of thermodynamics. Neither does it invalidate the additive property of information entropy, noted in Section 20 of Part 1), applied here to two, or more, independent thermodynamic systems. However, it does change the value of the thermodynamic entropy, $k\mathcal{H}$, of systems in a gas or liquid phase; giving those thermodynamic entropies the extensive property they need.

There is a second correction to the entropy of fluid systems, that must be applied. This is due to the fact that the molecules are free to rotate, and if a molecule is symmetrical in some way, there are two or more similar orientations of it. Let $\sigma\ (= 1, 2, 3 \ldots)$ be the number of these similar orientations.[28] To ensure the extensive property of the entropy, we need to leave aside our ignorance regarding which of the similar orientations is occupied by any one molecule. At any one time, there are σ^M equivalent ways the M molecules can be orientated. By indifference, the probabilities of these are all equal. Accordingly, we should subtract from the entropy the term $\ln \sigma^M$.

Altogether, we need then, to subtract the term $\ln(M!\,\sigma^M)$, from the entropy to get

$$H = -\int \rho(p,q) \ln\left(\left(2\pi\hbar\right)^s \rho(p,q)\right) dpdq - \ln(M!\,\sigma^M)$$

or to change the measure density from $m = (2\pi\hbar)^{-s}$ to $m' = m/(M!\,\sigma^M)$.

[28] Values of σ for various molecules are as follows: for water H_2O (like an isosceles triangle) $\sigma = 2$, for ammonia NH_3 (regular triangular pyramid) $\sigma = 3$, for methane CH_4 (tetrahedron) $\sigma = 12$, for benzene C_6H_6 (regular hexagon) $\sigma = 12$. For diatomic molecules with different atoms (such as HCl) $\sigma = 1$, but with diatomic molecules with identical atoms (like H_2), $\sigma = 2$.

25. Equations of state.

We will not spend much space on describing how classical statistical thermodynamics comes up with equations of state. We will only summarise the way this is done in general, and in the particular case of the monotonic perfect gas.

We have seen how statistical mechanics can account for the internal energy U of a system in mechanical terms, and account for the entropy S and absolute temperature T of the system in statistical terms; and this demonstratively results in relations between them the same as those in ordinary thermodynamics.

All other thermodynamic properties of a system are functions of U, S and T, and this is how they are defined both in ordinary thermodynamics and in the statistical theory.

Take, for example, the free energy of a system. This is defined as

$$F = U - TS$$

The free energy turns out to be an important quantity in the statistical theory, for it can be found directly in terms of the energy function $E(p, q)$ of the mechanical model of the system. The canonical distribution $\rho(p, q) = Ae^{-\lambda E(p,q)}$ and its information entropy $\mathcal{H} =$

$-\ln((2\pi\hbar)^s A) + \lambda\langle E\rangle$, give, for solid systems and fluid systems, the entropies,

$$S = \begin{cases} -k\ln((2\pi\hbar)^s A) + k\lambda & \text{solid} \\ -k\ln((2\pi\hbar)^s A) + k\lambda U - k\ln(M!\,\sigma^M) & \text{fluid} \end{cases}$$

where $\langle E\rangle$ is replaced by U, and the corrections (in Section 24) have been made to the entropy in the case of a fluid. By definition, we have therefore

$$F = \begin{cases} kT\ln[(2\pi\hbar)^s A] \\ kT\ln[(2\pi\hbar)^s A/(M!\,\sigma^M)] \end{cases}$$

Or using the identity $A = \left(\int e^{-\lambda E(p,q)} dpdq\right)^{-1}$,

$$F = \begin{cases} -kT\ln\left[(2\pi\hbar)^{-s}\int e^{-\lambda E(p,q)} dpdq\right] \\ -kT\ln\left[(2\pi\hbar)^{-s}\int e^{-\lambda E(p,q)} dpdq/(M!\,\sigma^M)\right] \end{cases}$$

giving F directly in terms the temperature and the energy function $E(p,q)$ of the mechanical model of the system.

It is normal, first to calculate what is known as the 'partition function' Z given by

$$Z = \begin{cases} (2\pi\hbar)^{-s} \int e^{-\lambda E(p,q)} dpdq \\ \dfrac{(2\pi\hbar)^{-s}}{M!\,\sigma^M} \int e^{-\lambda E(p,q)} dpdq \end{cases}$$

and then put

$$F = -kT \ln Z$$

Once F is calculated, other thermodynamic quantities are easily found. For example, (for a liquid)

$$P = -\left(\frac{\partial F}{\partial V}\right)_T$$

and (for a liquid or solid) the entropy is

$$S = -\left(\frac{\partial F}{\partial T}\right)_V$$

and the internal energy is

$$U = -T^2 \left(\frac{\partial}{\partial T} \frac{F}{T}\right)_V$$

These equalities are well known in ordinary thermodynamics where they are derived from the differential form of the first and second laws. These laws hold just as well for us, because we have established the truth of them on the basis of mechanical and statistical

theory. We can therefore derive the above results in the same way as is done in ordinary thermodynamics.

Case of the perfect monatomic gas

As a well-known example of the derivation of equations of state, we take the case of a monatomic perfect gas inside a rigid vessel. We model the N atoms as particles, so the energy function takes the form

$$E(p,q) = \frac{1}{2m}\sum_{j=1}^{N} \boldsymbol{p}_j^2 + V(\boldsymbol{r}_1, \ldots \boldsymbol{r}_N)$$

Where the vector $\boldsymbol{r}_i$ stands for the Cartesian coordinates of i^{th} particle and the vector $\boldsymbol{p}_j$ stands for the associated Cartesian components of particle momentum. The function $V(\boldsymbol{r}_1, \ldots \boldsymbol{r}_N)$ represents the system potential. As the gas is 'perfect', there are no particle-particle interaction potentials. So, the potential function $V(\boldsymbol{r}_1, \ldots \boldsymbol{r}_N)$ is the sum of identical particle-wall potential functions:

$$V(\boldsymbol{r}_1, \ldots \boldsymbol{r}_N) = \sum_{i=1}^{N} V(\boldsymbol{r}_i)$$

We take $V(\boldsymbol{r}_i)$ to be zero except at the boundary of the system where we take it be positively infinite. Then any

particle hitting the vessel wall is perfectly reflected off it without loss of energy or change in its momentum component parallel to the wall.

As the first step in calculating the partition function, we note that

$$e^{-\lambda E(p,q)} = \exp\left(-\frac{\lambda}{2m}\sum_{j=1}^{N}\boldsymbol{p}_j^2\right)\exp\left(-\lambda\sum_{i=1}^{N}V(\boldsymbol{r}_i)\right)$$
$$= \prod_{j=1}^{3N} e^{-\lambda\frac{1}{2m}p_j^2}\prod_{i=1}^{N} e^{-\lambda V(\boldsymbol{r}_i)}$$

where p_j here stands for any one of the $3N$ Cartesian components of particle momenta. Here the factors in the first product are the same function of each component of momentum. Similarly, the factors in the second product are the same function of coordinate vector. So,

$$\int e^{-\lambda E(p,q)}dpdq = \left(\int_{-\infty}^{\infty} e^{-\frac{\lambda}{2m}p^2}\,dp\right)^{3N}\left(\int e^{-\lambda V(\boldsymbol{r})}\,d\boldsymbol{r}^3\right)^{N}$$

In the second integral $V(\boldsymbol{r}) = 0$ (except at the boundary which doesn't count) so the integrand equals 1 and the integral itself is just the volume V of the vessel.

On account of the pure mathematical result

$$\int_{-\infty}^{\infty} e^{-\mu x^2}\, dx = \sqrt{\frac{\pi}{\mu}} \qquad \text{for } \mu > 0$$

the first integral is just $\sqrt{2m\pi/\lambda}$.

The partition function is therefore

$$Z = \frac{(2\pi\hbar)^{-s}}{N!} \int e^{-\lambda E(p,q)} dpdq = \frac{(2\pi\hbar)^{-s}}{N!} V^N (2m\pi)^{\frac{3}{2}N} \lambda^{-\frac{3}{2}N}$$

where we have set $\sigma = 1$, since a single particle has no rotational symmetry.

Putting $\lambda = 1/kT$, we obtain, for the free energy,

$$F = -kT \left[\ln((2\pi\hbar)^{-s}(2m\pi)^{3N/2}) - \ln N! + N \ln V + \frac{3N}{2} \ln kT \right]$$

Employing Stirling's approximation $N! = (N \ln N - N)$, and putting $s = 3N$ this reduces to

$$F = -NkT \left[\ln \frac{eV}{N} + \tfrac{3}{2} \ln kT + \zeta \right]$$

where

$$\zeta = \ln \left(\frac{m}{2\pi\hbar^2} \right)^{\frac{3}{2}}$$

is known as the chemical constant of the gas.

Applying the above general formulae for the pressure, entropy and energy, we find

$$P = -\left(\frac{\partial F}{\partial V}\right)_T = \frac{NkT}{V}$$

$$S = -\left(\frac{\partial F}{\partial T}\right)_V = Nk\left[\ln\frac{eV}{N} + \tfrac{3}{2}\ln kT + \tfrac{3}{2} + \zeta\right]$$

$$U = -T^2\left(\frac{\partial}{\partial T}\frac{F}{T}\right)_V = \tfrac{3}{2}NkT$$

By the last of these, the fixed volume heat capacity of the gas is

$$C_v = \left(\frac{\partial U}{\partial T}\right)_V = \tfrac{3}{2}Nk$$

With the substitution $k = R/N_A$ and $n = N/N_A$ all these results agree with the corresponding equations in the ordinary thermodynamics theory for n moles of a perfect gas.

26. The conservation of uncorrected entropy during adiabatic changes made to a system

Consider an ideal monatomic gas enclosed in a vessel fitted with a piston. Suppose the gas starts out in a quiescent state with a known expected energy.

In the phase space of the system, let the variables q and p be the particle coordinates and generalised momenta referred to a fixed coordinate frame.

As in Section 25, the system potential function $V(q)$, becomes positively infinite when a particle touches the vessel wall or the face of the piston. So, the function $V(q)$, and therefore the Hamiltonian $H(p, q)$, depends on the position of the piston.

Now suppose the piston is linked to a purely mechanical external body such as a spring or weight suspended from a pulley, and an agent, acting on our instructions, is employing the mechanical device to control the motion of the piston. Then we should write the Hamiltonian, as $H(p, q; t)$; i.e. as a known explicit function of the time,

The probability density in the system's phase space is now also a function $\rho(p, q; t)$ of the time. It obeys a differential equation which is found as follows.

We note that Hamilton's equations and Liouville's theorem still hold when the Hamiltonian is time dependent. So, as in Section 15, a small element of representative points in phase space will move without changing its volume. As the number of representative

points inside it stays the same, the probability density inside the element must stay the same. As a result, the substantial derivative $d\rho/dt$ is zero. Expressing this derivative as

$$\frac{d\rho}{dt} = \frac{\partial \rho}{\partial t} + \sum_{i=1}^{s} \left(\dot{q}_i \frac{\partial \rho}{\partial q_i} + \dot{p}_i \frac{\partial \rho}{\partial p_i} \right)$$

as we may, we see $\rho(p, q; t)$ obeys the differential equation

$$\frac{\partial \rho}{\partial t} + \sum_{i=1}^{s} \left(\dot{q}_i \frac{\partial \rho}{\partial q_i} + \dot{p}_i \frac{\partial \rho}{\partial p_i} \right) = 0$$

or, using Hamilton's equations

$$\frac{\partial \rho}{\partial t} + \sum_{i=1}^{s} \left(\frac{\partial H}{\partial p_i} \frac{\partial \rho}{\partial q_i} - \frac{\partial H}{\partial q_i} \frac{\partial \rho}{\partial p_i} \right) = 0$$

where $H(p, q; t)$ is known.

This is the differential equation $\rho(p, q; t)$ satisfies. In principle it can be solved to find the probability density at any time, when we know it at an earlier time.

The vanishing of the substantial derivative of $\rho(p, q; t)$, has an interesting consequence with regard to the information entropy of the system. At any time t, the

information entropy (short of any corrections mentioned in Section 25) is

$$\mathcal{H} = -\int \rho(p,q;t)\ln((2\pi\hbar)^s\rho(p,q;t))\,dpdq$$

We have here, in the integrand, the product of a volume element $dpdq$ of phase space and a function of the local probability density. The integral is the sum of these products over all volume elements making up the phase space. When a volume element moves to another place, its volume, and the probability density in its vicinity, stay the same. The element therefore makes the same contribution to the information at any time in its motion. Therefore, the sum of contributions of all elements stays the same. That is, the uncorrected information entropy $\mathcal{H}$ stays the same, however the agent controls the movement of the piston.

Although this result has been proved in the case of an ideal monatomic gas in a vessel containing a single piston, it is clear from the generality of the proof, that it must be true for any mechanically modelled system and for any adiabatic changes made to the position, shape and size of the vessel containing it. It is evidently true whether or not the phase or chemical composition of the system remains constant during the changes, and true whether or not the changes are reversible.

The result is not really surprising, because of our belief in the truth of the laws of motion. During the adiabatic changes, these laws unambiguously relate the values of the p and q at one time to those at another, so that our degree of ignorance of the dynamical state of the system stays the same.

We thus reach the following conclusion:

> *The uncorrected information entropy of a system remains constant during any adiabatic changes brought about by an agent.*

Suppose that, having brought about changes, the agent allows time for the system to return to a (homogeneous) state of thermodynamic equilibrium, before declaring its work complete. If the phase and chemical composition of the material of the system remain, or have returned to being, the same as before, the corrections needed to the entropy (established in Section 24) are the same as before. Hence the corrected entropy is necessarily the same at the end of the process as it was at the beginning, whether or not the change in state of the system is reversible.

But, in ordinary thermodynamics, the entropy of a system increases after an irreversible adiabatic change in state. So, we seem to have contradiction. This can,

however, be satisfactorily resolved by making a further correction to the entropy.

27. Further correction to the entropy

The initial probability density in the phase space of the system in Section 26, may be denoted $\rho(p, q; 0)$. This is the probability density of a canonical distribution, because only the energy of the system is supposed known initially, i.e. at time zero.

After action by the agent, i.e. at time $t = t_1$ when the system is again in thermodynamic equilibrium, the evolved density $\rho(p, q; t_1)$ has (as shown in Section 26) the same uncorrected entropy as $\rho(p, q; 0)$. Knowledge Y that leads to $\rho(p, q; t_1)$ is the knowledge of the initial expected energy of the system, together with the knowledge of the action carried out by the agent, or knowledge of the function $H(p, q; t)$.

In terms of $\rho(p, q; t_1)$ the expected energy of the system at time t_1 is, in principle, calculable. It is given by

$$\langle E_1 \rangle = \int E(p, q; t_1)\rho(p, q; t_1)\, dpdq$$

However, the actual calculation of $\rho(p, q; t_1)$ and thence of $\langle E_1 \rangle$, is generally quite impossible owing to the very many variables involved. But we can get to know the expected energy $\langle E_1 \rangle$ in another way. We can measure the

work done on the system under action of the agent and add this to the known expected energy of the system at time zero.

Now we can argue that, since calculation of $\rho(p, q; t_1)$ is generally impossible, we should from time t_1 onward, dump our knowledge Y and keep only our knowledge Y' of the expected energy $\langle E_1 \rangle$.

Then knowledge Y' alone, gives us, the (time independent) canonical distribution $\rho'(p, q; t_1)$ going with the known expected energy $\langle E_1 \rangle$ of the system at time t_1.

The dumping of knowledge Y is not necessary if all changes brought about by the agent are slow and reversible. Then, the distributions $\rho(p, q; t_1)$ and $\rho'(p, q; t_1)$ are the same.

Whether or not $\rho(p, q; t_1)$ and $\rho'(p, q; t_1)$ are the same, knowledge Y' alone, is sufficient (as any canonical distribution is) for the correct prediction of all (macroscopic) thermodynamic properties of the system at and after time t_1. So, holding on only to the part Y' of our knowledge is quite in keeping with the statistical approach to thermodynamics.

Under knowledge Y' alone, $\rho'(p, q; t_1)$ maximises the entropy at time t_1, while $\rho(p, q; t_1)$ may not, because it is generally a different distribution. As a result, the

uncorrected entropy under knowledge Y' is greater than or equal to the uncorrected entropy under knowledge Y.

This evidently remains true of the corrected entropies when the phase and chemical composition of the system are the same at time t_1, as they were at time zero. So we have avoided contradiction with the principle of entropy increase that arose in Section 25.

Even when the phase or chemical composition of the system are different at the end of the change as they were at its beginning, the statistical theory still does not lead to result that contradicts the principle of entropy increase. This is because the statistical theory provides a proof of the second law of thermodynamics, from which the principle of entropy increase can be established quite generally.

28. Mathematical support for dropping unusable knowledge after an irreversible change in state.

We have claimed, in Section 27, that the phase space distribution $\rho(p, q; t_1)$, is overcomplicated, and may justifiably be replaced by the distribution $\rho'(p, q; t_1)$ applying when we know only the expected energy of the system at time t_1. In support of this claim, we compare $\rho(p, q; t_1)$ and $\rho'(p, q; t_1)$ in a case in which both are actually calculable.

We consider a perfect gas, of known expected energy. The gas is contained in a vessel taking the form of a rectangular prism fitted with pistons at each end. The pistons can be moved suddenly and symmetrically outward to a new position, allowing the gas to expand into the vacuum that then appears. The particles of the gas do not interact with each other but they undergo simple reflection every time they hit the vessel wall or a (stationary) piston end.

Figure

We take Cartesian coordinates with origin at the centre of the vessel and axes parallel to the sides of the vessel, the $x$axis lying on the central axis of the prism.

The x, y, and z components of position of any one particle of the gas change in time independently of one another.

Because the pistons are moved out instantly, there is no change in the kinetic energy of any particle.

We set up the $6N$dimensional (Euclidean) phase space based on the Cartesian coordinates of the particles and their momenta. Then of course, at any one time, there will be a point P in the phase space representing the momentary dynamical state of the gas.

We consider the independent motion of the projection P′ of P onto the 2-dimensional plane containing the axes of the $x$coordinate of position and the $x$component of momentum p_x of a particular particle.

We suppose, for simplicity, that the mass of each particle is equal to 1, so that we may sometimes replace p_x by the component v of velocity of the particle in the $x$direction.

Figure

Before expansion of the gas, P′ will lie in a region of the xvplane bounded by the lines $x = \pm a$, where $2a$ is the initial separation of the pistons. Now, suppose P′ is, at some time, moving in the positive $x$direction, with uniform velocity equal to its v coordinate. It hits the boundary at the point (a, v). It then jumps down to the point $(a, -v)$ and moves in the other direction till it hits the other boundary at $(-a, -v)$ when it jumps up to $(-a, v)$ and proceeds along the path it was on before, and so on.

We are supposing the gas is, initially in a quiescent state with known expected energy. The probability

distribution in whole phase space is then canonical, and from the work in *Case of the perfect gas* in Section 25, it evidently factors into canonical distributions in each of the three cartesian position coordinates and associated momenta for each particle. So, the probability distribution $\rho(p_x, x)$ applying to the projection of the representative point P onto the xp_x plane for any one particle will be of the form

$$\rho(p_x, x) = \begin{cases} \frac{1}{2a}\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{{p_x}^2}{2\sigma^2}} & |x| < a \\ 0 & |x| > a \end{cases}$$

where $\sigma = mkT$ is the same constant for all 3 cartesian components and all N particles The distribution $\rho(p_x, x)$ is the product of a Gaussian distribution in the momentum p_x, and a uniform distribution in x for $|x| < a$.

At time $t = 0$, let dimension a jump up to b. Then, the distribution $\rho(p_x, x)$ will start to depend on the time. We write it as $\rho(v, x; t)$, and to help deduce its variation, we draw the following diagram.

Figure

We picturing the distribution as a cloud of points in the xvplane with a density proportional to $\rho(v,x)$, at time $t=0$. As time passes, we have a shearing flow of these points towards the right for $v>0$, $|x|<a$, and towards the left for $v<0$, $|x|<a$.

Here, the boundaries of the actual region in which P′ is now confined, are marked by lines $x=\pm b$ in the diagram. But we include, to the right and to the left, repetitions of this region side by side.

Of those repetitions, those numbered $\pm 2, \pm 4$, …etc. are simple copies of the actual region, and those numbered $\pm 1, \pm 3$, …etc. are copies of the actual region after inversion of it through the origin of coordinates.

Consider the points that have velocities in the range $v=-v_m$ to $v=v_m$. These points lie within the rectangle ABCD drawn in solid lines. The motion of the points can be pictured by allowing the rectangle containing them to deform into a moving parallelogram as if no impacts with boundaries were occurring. At a time t, after the sudden expansion of the vessel, this parallelogram will appear as shown. It contains in it, of course, all the representative points that were inside the rectangle at time $t=0$.

Drawing the dotted triangle with horizontal and vertical sides $v_m t$ and v_m respectively, we see the slope of the long sides of the parallelogram is t^{-1}, making its

vertical thickness $2a/t$. As time passes, the parallelogram gets forever thinner and longer.

Within the parallelogram, at time t, the density of points is constant in the $x$direction for any one value of v. For any one value of x, it is the same function of v as it was at time $t = 0$.

At time t, neighbouring vertical lines, for example $x = 2b$ and $x = 3b$, cut off segments of the parallelogram. We get the actual spatial distribution of points in the solid-line rectangle, by sliding the segments back into it. In doing this we should invert each odd numbered segment through the centre of the vertical lines bordering it before moving it back. However, by the symmetry of the diagram, this is not necessary. We get the same result by simply horizontally translating each and every segment back. When this is done, we get the pattern shown in the next diagram. Here the particle number density remains the same function of v *inside* the shaded areas. It is zero *between* the shaded areas. This establishes the distribution $\rho(v, x; t)$ (or $\rho(p_x, x; t)$) for values of velocity in the range $v = -v_m$ to $v = v_m$.

For a larger chosen value of value v_m, the pattern of shaded parallelograms remains the same, but occupies more of the region $|x| < b$. It fills the whole of that region when $v_m \to \infty$. The distribution $\rho(p_x, x; t)$ is then fully established.

The shaded parallelogram segments are periodically spaced apart by a vertical distance $2b/t$. Each shaded segment has a vertical height of $2a/t$.

As t becomes larger and larger, the shaded parallelograms get thinner and thinner, become more and more parallel to the $x$axis, and pack closer and closer together.

Partitioning the region $|x| < b$ into equal small cells of area $\Delta x \Delta v$, and supposing t is so large that there are very many shaded rectangular parallelograms crossing each cell, we see that on taking the average of $\rho(p_x, x; t)$ over each cell we get the smooth distribution

$$\overline{\rho(p_x, x; t)} = \begin{cases} \dfrac{1}{2b}\dfrac{1}{\sigma\sqrt{2\pi}} e^{-\frac{p_x^2}{2\sigma^2}} & |x| < b \\ 0 & |x| > b \end{cases}$$

which is itself independent of the time.

The distributions $\rho(p_y, y; t)$ and $\rho(p_z, z; t)$ do not change under the expansion of the gas. They are also independent of the time, and are the same as they were before the expansion of the gas. The variable t can therefore be dropped in all the component distributions when enough time has passed for the gas to be again in a quiescent state.

The probability distribution $\rho'(p,q)$ after expansion then becomes the product of the distributions $\overline{\rho(p_x,x)}\rho(p_y,y)\rho(p_z,z)$ for each particle of the gas.

By averaging the distribution over p_x and x for each particle, we are doing the equivalent of dropping unnecessary knowledge represented by the time dependent, and complicated distribution $\rho(p_x,x;t)$, and keeping hold only of our knowledge of the expected final energy of the system, which has stayed the same.

The post-distribution $\rho'(p,q)$ is clearly canonical, and, apart from the constant factor a/b , is the same function as the $\rho(p,q)$. The constant factor arises only because of the need to renormalise the distribution to take account of the increase in volume of the gas.

The above analysis gives mathematical support to the practice (see Section 27) of dropping unusable knowledge after an adiabatic irreversible change in state of *any* system. It suggests that, under irreversible action by an agent, cells made up of representative points in the ($2s$ dimensional) phase space of a system are drawn out into thin $(2s-1)$ dimensional sheets becoming thinner and thinner as time passes during and after action by the agent; and that finally, a new state of macroscopic quiescence sets in, and the locally averaged (or smoothed over)

probability density, ceases to change with time[29]. That averaged probability distribution is a canonical distribution applying under knowledge alone of the final expected energy of the system, deducible from its known initial energy and the work done on account of action by the agent.

29. Derivation of the second law of thermodynamics

In ordinary thermodynamics, the Kelvin-Planck statement of the second law (one of many equivalent ways of stating it) is as follows:

> *It is impossible to construct an engine, which working in a cycle, produces no effect other than the extraction of a quantity of heat from a heat reservoir and the performance of an equivalent quantity of work.*

We will show how this statement is a logical consequence of the statistical theory of thermodynamics.

To avoid the difficulty in modelling a 'heat reservoir', it will be sufficient to show how the statistical theory confirms the truth of the following statement:

[29] We can liken this to the (reversible) mixing together of highly viscous liquids of different colours. After much mixing, a liquid of apparently uniform colour results even though, when examined through a microscope, the colours remain clearly separated.

It is impossible to construct an engine which does nothing more than extract heat from a system, perform an equivalent quantity of work, and return to its original state.

Here 'system', means any thermodynamic system contained in a rigid vessel and modelled as a collection of particles. The truth of this statement implies the truth of the Kelvin-Planck statement of the second law. For, the 'system' may become a 'heat reservoir' by taking the limit as its size and thermal conductivity are allowed to tend to infinity.

We will work with the system before the limit is taken, but still refer to it as the 'heat reservoir', or 'reservoir' for short.

We take the engine to be a number of statistically modelled thermodynamic systems that can interact with each other, and work together in a way decided by an agent.

The reservoir and the systems in the engine together constitute a 'composite system'; and all the actions of the agent from beginning to end will be referred to as 'the process'.

At the start of the process the component systems are each closed and in thermodynamic equilibrium; and we take it that we know their expected energies.

During the process, none of the systems making up the composite need be in thermal, mechanical or chemical equilibrium, and the systems in the engine, need not always be closed, and mixing of their molecules might occur. None-the-less, the composite system can be modelled in the manner of classical statistical mechanics. That is, as a collection of particles with particle-particle interaction potentials and (generally time-dependent) particle-wall interaction potentials.

Possible actions of the agent

We wish to prove the second law for the most general kind of engine possible. So, we must allow any number of systems of any kind in the engine, and any sort of action that could be performed on the systems making up the composite. We stipulate only that there be no change in the size shape and content of the system representing the reservoir.

We now give a few examples of possible actions, and how they can be modelled in the statistical theory. We employ a fixed coordinate system.

(i) The agent might, at some time, have the vessel containing a single system moved to another position or orientation in space. Such action can be modelled by making time dependent, parameters in the molecule-wall interaction potential functions that specify the position

and orientation of the vessel. That way, the potential function and Hamiltonian of the system become explicitly dependent on the time.

(ii) The agent might employ mechanical work from outside, to drive a paddle wheel in a thermally isolated vessel containing a liquid. This can be modelled by making time dependent, parameters in the molecule-wall interaction potential functions that specify the orientation of the paddle wheel. When the paddle-wheel stops in the same orientation as it had originally, the system Hamiltonian will be the same as before, but the energy of the system will have increased as a result of the stirring.

(iii) The agent might employ pistons fitted in containing vessels to allow one system to do mechanical work on another, or to do mechanical work on bodies external to the engine. Such actions by the agent can be modelled by making time-dependent, parameters in the potential functions of interaction between molecules and the faces of the pistons.

(iv) The agent might, remove the partition in a vessel separating gases of different chemical composition. This can be modelled by taking away, from the (two-gas) system Hamiltonian, the potential of interaction between the partition and molecules of either gas, and adding a potential of interaction between molecules of one gas and

those of the other –a potential that, without the possibility of mixing, was inactive.

(v) The agent might open and close valves for specified times, allowing quantities of gas to mix in a vessel, in order for the gases to chemically react with each other. Such chemical reactions might later be reversed by raising the temperature of the vessel. Then, the two gases might be separated again by the insertion of a suitable membrane. These actions of the agent can be modelled by introducing appropriate inter-molecular potentials and potentials between molecules of each gas and the membrane. Actions like these might lead, at the end of a process, to the molecules of a component system being made up of particles different from those making them up at the beginning of the process. However, from the macroscopic point of view (which, in the end, is all that matters) the system's chemical composition and the quantity of matter in it may still be the same at the end of the process as at the beginning.

(vi) The agent might bring into contact a part of the surface areas of two vessels in order to allow heat to flow from one system to another, or to remove heat from the heat reservoir. This can be modelled by adding to the Hamiltonian a potential of interaction between the molecules of one system and those of the other –a potential active only between molecules close to the area of contact of the vessels.

(vii) The agent might produce a change in phase of a liquid system by extracting heat from it by putting the vessel containing it in thermal contact with a monatomic gas in a long vessel fitted with a piston. By sliding the piston outward, the gas can be cooled till the liquid freezes. This action can be modelled by introducing suitable potentials of interaction between particles of the gas and molecules of the liquid across the area of contact of the vessels. A time-dependent potential of interaction between gas particles and the piston face would also be required.

The systems comprising the engine are supposed to be in the same quiescent states at the end of the process as they were at the beginning. Only the heat reservoir ends up in a different quiescent state because we are supposing the agent has had a quantity Q of heat extracted from it. We take as known the expected value of Q and the (equivalent) expected value of the work W the engine performs on outside bodies. We may therefore claim to know the expected energy of the reservoir and each system in the engine, both at the beginning and at the end of the process.

We will now show that if the engine succeeds in carrying out what is required of it, a contradiction arises. Its action results in the entropy of the system comprising

the engine and reservoir together, (i) increasing or staying the same *and* (ii) decreasing.

Proof that the entropy increases or stays the same.

This follows from the generalisation of the conservation of the uncorrected entropy of a single system in Section 26, to the case of a composite system; in particular, the composite system of the engine and reservoir taken together.

Let the coordinates and momenta p, q refer to all the particles making up the composite system. The Hamiltonian $H(p, q; t)$ is now that of the composite system, and changes in time as the agent acts. The differential equation for the probability density $\rho(p, q; t)$ is the same as in Section 26, and the proof (using Liouville's theorem) that the uncorrected information entropy of the composite system stays constant during action by the agent applies just as well.

The required correction (established in Section 27) to the information entropy of a single system after action by the agent, must now apply to the composite system. Evidently, we should drop the knowledge reflected in the probability distribution $\rho(p, q; t_1)$ at the end of the process (at time t_1), and retain only our knowledge of the expected energies of *each* component system at the end of the process. The distribution $\rho(p, q; t_1)$ under the original state of knowledge, will (except when actions of the agent

are all reversible) be different from the new distribution $\rho'(p, q; t_1)$ under the new state of knowledge. So, the former distribution will generally no longer maximise the entropy at time t_1. As a result, the corrected entropy of the composite system at the end of the process will be larger than or equal to the uncorrected entropy at time t_1; and therefore larger than or equal to the entropy of the composite system at the beginning of the process.

Our new probability distribution $\rho'(p, q; t_1)$, is the product of the canonical distributions of each of the, now closed, systems of the engine and reservoir; and the entropy of the composite system is the sum of the entropies of its component systems.

At the start of the process (at $t = 0$), our probability distribution $\rho(p, q; 0)$ was the product of the canonical distributions of the reservoir and of each of the closed systems of the engine. The entropy of the composite system was then the sum of the entropies of its component systems at the start of the process.

Since the chemical composition, the phases and the number of molecules in each of the component systems are the same at the end of the process as they were at the start of the process, the needed corrections to the entropy, established in Section 24, are the same at the end of a process as at the start of the process.[30] Therefore, after

[30] When the process involves mixing of the molecules of systems in the engine, in cannot be ensured that the number of molecules in

making these corrections, the entropy of the composite system at the end of the process (being the sum of the entropies of its component systems) must still be equal to or greater than the entropy of the composite system at the beginning of the process. QED.

Proof that the entropy decreases

During the process, the reservoir loses a quantity of heat Q. No matter how the heat is removed (reversibly or not), the reservoir (like the other component systems) is in a state of thermodynamic equilibrium both at the beginning and at end of the process. Therefore, the change in entropy of the reservoir can be calculated using the equation of state of the reservoir giving its entropy as a function of its internal energy and its volume. Differentiating this equation of state with respect to the energy we have, by the differential form of the second law with dW set equal to zero, that $(\partial S/\partial U)_V = 1/T$. So, we see that the entropy of the reservoir must be a monotonic increasing function of its internal energy. Therefore, having lost energy Q, the entropy of the reservoir must be smaller at the end of the process than it was at the beginning.

---

each system is exactly the same at start and finish. But we may reasonably assume that this discrepancy is of no consequence for our purpose.

Since the thermodynamic entropies of all the other component systems are the same at the end as at the beginning of the process, the thermodynamic entropy of the composite system must have decreased. QED.

Hence the contradiction, which is only avoided by acknowledging that the second law of thermodynamics is a logical consequence of the statistical theory of thermodynamics.

Brownian motion

In this part of the book, we reconsider (from the rational Bayesian perspective) the statistical arguments Einstein employed[31] in his theory of the random motion of tiny spherical particles suspended in a liquid (Brownian motion).

Using the same assumptions as Einstein, we employ rational Bayesian methods to derive the probability distribution for the displacement of a particle from its initial position after a specified time.

From this, follows the diffusion equation and general expressions for the particle drift-velocity and the probability flux density.

26. The process considered

We have in mind the motion of a dilute suspension of very many small particles (e.g. pollen particles) in a stationary liquid. As is now known, the slow irregular drifting motion of any one of these particles is due to numerous tiny impacts between the particle and the ions or molecules of the liquid.

Since the suspension is *dilute*, the motion of any particular particle may be supposed to proceed in a

[31] Ann. d. Phys. 17 (1905) 549-560

manner physically independent of the motion of the others. The same irregular motion of a particle is observed whether or not it is in the company of others. It is sufficient, therefore, to study the motion of a *single* particle in what we take to be an unbounded liquid.

Concerning knowledge of the particle's motion

We suppose we know from experience, that the square of the displacement of the particle's position from its starting point, tends to increase in proportion to the time of travel. The constant of proportionality is a characteristic of the motion. It has dimension $[L^2T^{-1}]$. We will assume, in our Bayesian reasoning, that this is the only physical quantity known to characterising the motion. On this basis, we will derive the form of the probability distribution over the position of the particle after a specified time; its initial position being known. This will lead us to the diffusion equation and to definitions of particle drift-velocity and probability flux density.

The propagator

We first seek an expression for the probability $P(dV_2t_2|dV_1t_1Y)$ for the particle to be in volume element dV_2 at $\boldsymbol{r}_2$ at the time t_2, when we know it was in volume element dV_1 at $\boldsymbol{r}_1$ at the earlier time t_1.

We write the probability as:

$$P(dV_2 t_2 | dV_1 t_1) = f_{\boldsymbol{r}_1 t_1}(\boldsymbol{r}_2 t_2) d^3\boldsymbol{r}_2$$

where the probability density[32] $f_{\boldsymbol{r}_1 t_1}(\boldsymbol{r}_2 t_2)$ is called the 'propagator'. Clearly

$$\int f_{\boldsymbol{r}_1 t_1}(\boldsymbol{r}_2 t_2)\, d^3\boldsymbol{r}_2 = 1$$

where the integration is over all of space.

We find the form of the propagator by applying the principles of prior probability assignment under knowledge that

(i) The liquid medium is homogeneous and isotropic.
(ii) The particle never moves infinitely fast.
(iii) The three propositions claiming the values of the Cartesian coordinates of the particle at time t_2 are logically independent of one another.
(iv) There is just one parameter characterising the motion of the particle in the liquid. This is a real positive 'diffusion coefficient' of dimension $[L^2 T^{-1}]$.

[32] For simplicity we sometimes omit commas between independent variables in functions, like between $\boldsymbol{r}$ and t here.

All these assumptions are effectively made by Einstein in his derivation of the diffusion equation.

27. Derivation of the form of the propagator

We derive the form of the propagator $f_{\boldsymbol{r}_1 t_1}(\boldsymbol{r}_2 t_2)$ in two ways.

Method 1. Use of similarity, the transformation groups of displacement and rotation, and dimensional analysis.

Taking a system of coordinates with origin O, we can form another with origin O', by performing a general displacement $\boldsymbol{\Delta}$ of the origin of coordinates and a general displacement δ of the origin of the time. An event specified by t and $\boldsymbol{r}$ in O, is represented by $t' = t - \delta$, and $\boldsymbol{r}' = \boldsymbol{r} - \boldsymbol{\Delta}$ in O'.

We have, by similarity and the homogeneity of space and time, that under this transformation of coordinates, the probability density remains the same *function* of the coordinates, So

$$f'_{\boldsymbol{r}_1 t_1}(\boldsymbol{r}_2 t_2) = f_{\boldsymbol{r}_1 t_1}(\boldsymbol{r}_2 t_2)$$

Also, according to the equations of transformation,

$$f'_{\boldsymbol{r}'_1 t'_1}(\boldsymbol{r}'_2 t'_2) = f_{\boldsymbol{r}_1 t_1}(\boldsymbol{r}_2 t_2)$$

when $t_1^{'} = t_1 - \delta$, $t_2^{'} = t_2 - \delta$, $\boldsymbol{r}_1^{'} = \boldsymbol{r}_1 - \boldsymbol{\Delta}$, $\boldsymbol{r}_2^{'} = \boldsymbol{r}_2 - \boldsymbol{\Delta}$. Combining these two gives

$$f_{\boldsymbol{r}_1 - \boldsymbol{\Delta},\, t_1 - \delta}(\boldsymbol{r}_2 - \boldsymbol{\Delta}, t_2 - \delta) = f_{\boldsymbol{r}_1 t_1}(\boldsymbol{r}_2 t_2)$$

implying

$$f_{\boldsymbol{r}_1 t_1}(\boldsymbol{r}_2 t_2) = F(\mathbf{R}, \tau)$$

where $\mathbf{R} = \boldsymbol{r}_2 - \boldsymbol{r}_1, \quad \tau = t_2 - t_1$.

Also, if we rotate vectors $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$ through the same angle about any axis passing through the origin to form new vectors $\boldsymbol{r}_1^{'}$ and $\boldsymbol{r}_2^{'}$, then, on account of isotropy of space, we must evidently have

$$f_{\boldsymbol{r}_1^{'} t_1}(\boldsymbol{r}_2^{'} t_2) = f_{\boldsymbol{r}_1 t_1}(\boldsymbol{r}_2 t_2)$$

This means $F(\mathbf{R}, \tau)$ must be a function of $\mathbf{R}$ invariant under rotation of $\mathbf{R}$, i.e.

$$F(\mathbf{R}, \tau) = f(R, \tau), \quad R = |\mathbf{R}|$$

Hence the result

$$f_{\boldsymbol{r}_1 t_1}(\boldsymbol{r}_2 t_2) = f(R, \tau)$$

This might have been obvious from the start, but we have spelt out the proof of it to illustrate the logical application of the principles of probability assignment.

Now $f(R,\tau)$ must have dimension $[L^{-3}]$. But the only quantities at our disposal are R, τ and the diffusion coefficient D, which has dimension $[L^2/T]$. So dimensional analysis gives us $f(R,\tau) = (D\tau)^{-3/2}h(R^2/D\tau)$. Or since $f(R,\tau)$ is positive, we will write

$$f_{\boldsymbol{r}_1 t_1}(\boldsymbol{r}_2 t_2) = (D\tau)^{-3/2} e^{h(R^2/D\tau)}$$

where $h(R^2/D\tau)$ is, as yet, an unknown dimensionless function of $R^2/D\tau$.

The whole argument leading to this result can be repeated for any one Cartesian component of $\boldsymbol{r}$. The only difference is that invariance under rotations must be replaced by invariance under inversion through the origin. Taking the case of the $x$component, this results in the probability density

$$f_{x_1 t_1}(x_2, t_2) = (D\tau)^{-1/2} e^{g(X^2/D\tau)}$$

where $X = x_2 - x_1$ and g is some real-valued function.

By similarity, exactly the same probability densities relating to particle movement in the y and in the z directions apply. That is,

$$f_{y_1 t_1}(y_2, t_2) = (D\tau)^{-1/2} e^{g(Y^2/D\tau)}$$

and

$$f_{z_1 t_1}(z_2, t_2) = (D\tau)^{-1/2} e^{g(Z^2/D\tau)}$$

By assumption (iii), these three probabilities densities are logically independent, therefore

$$f_{\boldsymbol{r}_1 t_1}(\boldsymbol{r}_2 t_2) = f_{x_1 t_1}(x_2, t_2) f_{y_1 t_1}(y_2, t_2) f_{z_1 t_1}(z_2, t_2)$$

It follows that

$$h(R^2/D\tau) = g(X^2/D\tau) + g(Y^2/D\tau) + g(Z^2/D\tau)$$

and since $R^2 = X^2 + Y^2 + Z^2$, we conclude that g must be a linear function. Hence

$$f_{\boldsymbol{r}_1 t_1}(\boldsymbol{r}_2 t_2) = (D\tau)^{-3/2} \exp\left(\frac{A}{D\tau}(X^2 + Y^2 + Z^2) + B\right)$$

where A and B are real constants.

The constant A must be negative, otherwise, in an arbitrary short time we would expect the particle to move to infinity, contrary to assumption (ii).

Applying the requirement for normalisation, using the mathematical result

$$\int_{-\infty}^{\infty} e^{-\lambda x^2} dx = \sqrt{\frac{\pi}{\lambda}} \quad \text{for} \quad \lambda > 0$$

we find

$$f_{\boldsymbol{r}_1 t_1}(\boldsymbol{r}_2 t_2) = (\frac{\pi D\tau}{\alpha})^{-3/2} \exp(-\frac{\alpha}{D\tau}(X^2 + Y^2 + Z^2))$$

where $\alpha\ (= -A)$ is a positive constant.

Since D is only defined to within a numerical constant, α can be set equal to any positive real number. We choose to put $\alpha = 1/4$, and finally obtain

$$f_{\boldsymbol{r}_1 t_1}(\boldsymbol{r}_2 t_2) = (4\pi D\tau)^{-3/2} \exp(-\frac{(\boldsymbol{r}_2 - \boldsymbol{r}_1)^2}{4D\tau})$$

for the propagator. This is a normal distribution where the mean square value of displacement is

$$\overline{(\boldsymbol{r}_2 - \boldsymbol{r}_1)^2} = 2D\tau$$

Method 2. Use of the method of maximum entropy

We repeat *Method 1* to the point where we have established that $f_{\boldsymbol{r}_1 t_1}(\boldsymbol{r}_2 t_2) = f(R, \tau)$.

Then, for any one value of τ, and any one position $\boldsymbol{r}_1$, we write the propagator as

$$f_{\boldsymbol{r}_1 t_1}(\boldsymbol{r}_2 t_2) = p(|\boldsymbol{r}_2 - \boldsymbol{r}_1|)$$

The RHS is a probability density p at a general point $\boldsymbol{r}_2$ in space. Applying the method of maximum information entropy, we find p by maximising

$$-\int p \ln \frac{p}{m} dV$$

where the measure density m of volume is just equal to 1, and the integral is conducted over the whole of space, dV being a volume element of space.

The maximisation of the entropy must be done under two constraints. One is just the normalisation requirement:

$$\int p \, dV = 1$$

The other relates to the mean square value of $\overline{(\boldsymbol{r}_2 - \boldsymbol{r}_1)^2}$, that is, to the mean square distance moved by the particle from $\boldsymbol{r}_1$ to $\boldsymbol{r}_2$ in time τ. Since the only characteristic of the motion is a diffusion coefficient D of dimension $[L^2/T]$, it can only be that

$$\overline{(\boldsymbol{r}_2 - \boldsymbol{r}_1)^2} = \alpha D \tau$$

where α is a dimensionless constant. Our second constraint is therefore

$$\int (\boldsymbol{r}_2 - \boldsymbol{r}_1)^2 p \, dV = \alpha D\tau$$

Following Lagrange, we first maximise

$$-\int p \ln p \, dV + \lambda\left(\int p \, dV - 1\right) - \mu\left(\int r^2 p \, dV - \alpha D\tau\right)$$

where we write $|\boldsymbol{r}_2 - \boldsymbol{r}_1|$ as r. The first order change under a small variation δp of p, is

$$\int \left(-\delta p \ln p - p\frac{1}{p}\delta p + \lambda\delta p - \mu r^2 \delta p\right) dV$$

and, as δp is arbitrary, we find $-\ln p - 1 + \lambda - \mu r^2 = 0$ or

$$p = e^{-1+\lambda} e^{-\mu r^2}$$

The constants λ and μ are found by inserting this expression into the equations of constraint. Employing the mathematical result

$$\int_0^\infty r^{2n} e^{-\mu r^2} dr = \frac{1.3.5 \ldots (2n-1)}{2^{n+1}\mu^n}\sqrt{\frac{\pi}{\mu}} \qquad n = 1,2,3 \ldots$$

we thus obtain

$$p = (4\pi D\tau)^{-3/2} e^{-r^2/4D\tau}$$

where we have chosen to put $\alpha = 2$. This reproduces the result

$$f_{\boldsymbol{r}_1 t_1}(\boldsymbol{r}_2 t_2) = (4\pi D\tau)^{-3/2} \exp(-\frac{(\boldsymbol{r}_2 - \boldsymbol{r}_1)^2}{4D\tau})$$

obtained using *Method 1*.

When we derived this equation by *Method 1*, we had to make use of assumption (iii). This assumption is not needed in *Method 2*. So, we see that assumption (iii) is not really necessary. This is a notable advantage of the method of maximum entropy, because it often avoids having to make an assumption (like assumption (iii)) which may seem rather questionable at the start of the analysis.

28. Experimental determination of the diffusion coefficient

The formula for the propagator expresses the probability for the particle to be in any volume element of space at a specified time, given we know its position at an earlier time. If at time $t = 0$ we observe it at the origin,

then at a time t later on, our probability density over the $x$component of its position $\boldsymbol{r}$ is evidently

$$f(x,t) = (4\pi Dt)^{-1/2} \exp(-\frac{x^2}{4Dt})$$

Using a microscope graticule, we can make many measurements of the particle's x coordinate during its motion, and thus confirm the *form* of the distribution; and the necessary value of the diffusion coefficient D for the liquid and particle in question.

The mathematical result

$$\int_0^\infty x^2 e^{-\mu x^2} dx = \frac{\sqrt{\pi}}{4\mu^{3/2}}, \qquad \mu > 0$$

gives the simple relation

$$\overline{x^2} = \int_{-\infty}^{\infty} x^2 f(x,t) dx = 2Dt$$

That can be used to calculate D from many measurements of x and t.

29. The diffusion equation

The propagator

$$f_{\boldsymbol{r}_1 t_1}(\boldsymbol{r}_2 t_2) = (4\pi D\tau)^{-3/2} \exp(-\frac{(\boldsymbol{r}_2 - \boldsymbol{r}_1)^2}{4D\tau})$$

derived in Section 27, can be written as

$$f_{\boldsymbol{r}_1}(\boldsymbol{r}, t) = (4\pi Dt)^{-3/2} \exp(-\frac{(\boldsymbol{r} - \boldsymbol{r}_1)^2}{4Dt})$$

where t_1 has been set equal to zero, t_2 to t, and $\boldsymbol{r}_2$ to $\boldsymbol{r}$. It is thus expressed as a function of position $\boldsymbol{r}$ after a time t has passed since the particle was at $\boldsymbol{r}_1$. In this form it is a solution of the diffusion equation

$$\frac{\partial p}{\partial t} - D\nabla^2 p = 0$$

as can be proved by direct substitution of $f_{\boldsymbol{r}_1}(\boldsymbol{r}, t)$ for p.

A *general* state of knowledge of the position of the particle at time zero may be expressed as a probability density $p(\boldsymbol{r}_1, 0)$. Then, at a later time t the probability density will be

$$p(\boldsymbol{r}, t) = \int p(\boldsymbol{r}_1, 0) f_{\boldsymbol{r}_1}(\boldsymbol{r}, t) dV_1$$

This can be seen by multiplying through by volume element dV and noting that the integrand (with dV and

dV_1 included) is, by the product rule, the probability that the particle is in dV_1 at time zero *and* in dV at time t. Then, by summing over elements dV_1, we see, as required, that the RHS equals the probability that the particle is in dV at time t.

Being a linear combination of the functions $f_{\boldsymbol{r}_1}(\boldsymbol{r}, t)$ for various values of $\boldsymbol{r}_1$, it is clear that $p(\boldsymbol{r}, t)$ satisfies the diffusion equation, and is the general solution of it.

30. Particle drift-velocity

Suppose our prior knowledge Y of the particle motion is represented by a probability density $p(\boldsymbol{r}, t)$. This could, for example be the density arising from simple knowledge that the particle was at a certain position at some early time.

Anyway, on the basis of prior knowledge Y, represented by $p(\boldsymbol{r}, t)$, we now calculate, using the method given by Jaynes[33], the expected drift-velocity of the particle when at a certain position at a certain time.

Before explaining Jaynes' method of calculation, we need to explain what we mean by the drift-velocity.

In the first place we note that the actual (thermal) particle velocity (the velocity of its centre of mass) varies

[33] 'Clearing up Mysteries -The Original Goal' in 'Maximum Entropy and Bayesian Methods', Ed. J Skilling, Cambridge, England, 1988, Kluwer Academic Publishers.

in a very irregular manner on account of the rapid fluctuating forces impressed on it by the very many ions or molecules of the liquid near its surface. So, at the larger time scale, the velocity is not a continuous, differentiable function of the time. It is, however, *integrable* over time. So, we can define the drift-velocity at time t_1 as the distance moved between times $t_1 - \tau$ and $t_1 + \tau$ divided by the short time interval 2τ. The drift velocity is thus the average of the actual (thermal) velocity from time $t_1 - \tau$ to time $t_1 + \tau$.

Defined in this way, the drift-velocity is a real property of the particle motion at any time, independently of any knowledge we may hold about the particle motion. Whatever prior knowledge Y we do hold, we will see that the calculated expected value of the drift velocity of the particle (when at a known position at a given time) has a definite value in the limit as $\tau \to 0$.

Jaynes' calculation of particle drift-velocity

To calculate the drift-velocity we consider three volume elements dV_1, dV_2 and dV_3 of the liquid as shown in the Figure. Let the positions of these be $\boldsymbol{r}_1$, $\boldsymbol{r}_2$ and $\boldsymbol{r}_3$, drawn from origin 0.

Figure

Having somehow learnt the particle is in dV_1 at time t_1, we can derive, in terms of $p(\boldsymbol{r}, t)$, an expression for the expected value of the drift-velocity $\boldsymbol{v}$ at time t_1. To do this, we work out the expected positions $\langle \boldsymbol{r}_2 \rangle$ and $\langle \boldsymbol{r}_3 \rangle$ of the particle at times $t_2 = t_1 + \tau$ and $t_3 = t_1 - \tau$ respectively, where τ is a short time; i.e. a time short compared to the characteristic time of variation of $p(\boldsymbol{r}, t)$, yet large compared to the characteristic time between impacts of ions or molecules with the particle.

The expected velocities just after and just before time t_1 are

$$\boldsymbol{v}_+ = \frac{\langle \boldsymbol{r}_2 \rangle - \boldsymbol{r}_1}{\tau} \qquad \boldsymbol{v}_- = \frac{\boldsymbol{r}_1 - \langle \boldsymbol{r}_3 \rangle}{\tau}$$

and the expected drift-velocity is

$$\boldsymbol{v} = \frac{\boldsymbol{v}_+ + \boldsymbol{v}_-}{2}$$

The probability the particle is in dV_2 at time t_2 is denoted by $P(dV_2 t_2 | dV_1 t_1 Y)$. Here knowledge Y, acquired at a time before t_3, is redundant. So, the expected position $\langle \boldsymbol{r}_2 \rangle$ is found from the propagator applying from time t_1 onward. As the propagator is a normal distribution centred at $\boldsymbol{r}_1$, $\langle \boldsymbol{r}_2 \rangle$ is clearly equal to $\boldsymbol{r}_1$. So $\boldsymbol{v}_+ = 0$.

To calculate $\langle \boldsymbol{r}_3 \rangle$, we need to employ Bayes' rule, to find the distribution $P(dV_3 t_3 | dV_1 t_1 Y)$ applying to the particle position at time $t_1 - \tau$. We have

$$P(dV_3 t_3 | dV_1 t_1 Y) = P(dV_1 t_1 | dV_3 t_3 Y) \frac{P(dV_3 t_3 | Y)}{P(dV_1 t_1 | Y)}$$

Now, on the RHS, the first term has the probability density of the propagator, i.e.

$$(4\pi D\tau)^{-3/2} \exp\left(-\frac{(\boldsymbol{r}_1 - \boldsymbol{r}_3)^2}{4D\tau}\right)$$

and clearly, $P(dV_3 t_3 | Y)$ and $P(dV_1 t_1 | Y)$ have the probability densities $p(\boldsymbol{r}_3, t_3)$ and $p(\boldsymbol{r}_1, t_1)$. So, the probability density of $P(dV_3 t_3 | dV_1 t_1 Y)$ is

$$(4\pi D\tau)^{-3/2} \exp\left(-\frac{(\boldsymbol{r}_1 - \boldsymbol{r}_3)^2}{4D\tau}\right) \frac{p(\boldsymbol{r}_3, t_3)}{p(\boldsymbol{r}_1, t_1)}$$

Since, $\boldsymbol{r}_3$ is close to $\boldsymbol{r}_1$, and t_3 close to t_1, we may write

$$p(\boldsymbol{r}_3, t_3) = p(\boldsymbol{r}_1, t_1) + \nabla p . (\boldsymbol{r}_3 - \boldsymbol{r}_1) + \frac{\partial p}{\partial t}(t_3 - t_1)$$

where ∇p and $\partial p / \partial t$ are the gradient and time derivative of $p(\boldsymbol{r}, t)$ evaluated at $\boldsymbol{r}_1$ at time t_1. So, with $\boldsymbol{r} = \boldsymbol{r}_1 - \boldsymbol{r}_3$

$$\frac{p(\boldsymbol{r}_3, t_3)}{p(\boldsymbol{r}_1, t_1)} = 1 - \frac{1}{p}\nabla p.\boldsymbol{r} - \frac{1}{p}\frac{\partial p}{\partial t}\tau$$

To calculate $\langle \boldsymbol{r}_3 \rangle$ we have, then, to find the expected value of $\boldsymbol{r}$ under the distribution

$$(4\pi D\tau)^{-3/2} \exp(-\frac{(\boldsymbol{r})^2}{4D\tau}) \left(1 - \frac{1}{p}\nabla p.\boldsymbol{r} - \frac{1}{p}\frac{\partial p}{\partial t}\tau\right)$$

Since the second and third terms in the bracket are small compared to 1, we can rewrite this distribution as

$$(4\pi D\tau)^{-3/2} \exp(-\frac{(\boldsymbol{r})^2}{4D\tau} - \frac{1}{p}\nabla p.\boldsymbol{r} - \frac{1}{p}\frac{\partial p}{\partial t}\tau)$$

Then, completing the square in $\boldsymbol{r}$**,** we obtain

$$(4\pi D\tau)^{-3/2} \exp(-\frac{(\boldsymbol{r} + 2D\tau(\nabla p/p))^2}{4D\tau} - \frac{1}{p}\frac{\partial p}{\partial t}\tau)$$

This is a normal distribution with the expected value of $\boldsymbol{r}$ equal to $-2D\tau(\nabla p/p)$. This makes

$$\boldsymbol{v}_{-} = \frac{\boldsymbol{r}_1 - \langle \boldsymbol{r}_3 \rangle}{\tau} = -2D\frac{1}{p}\nabla p$$

We arrive therefore, at the formula

$$\boldsymbol{v} = -D\frac{1}{p}\nabla p$$

for the expected drift-velocity in terms of the value and gradient of the prior probability density $p(\boldsymbol{r}, t)$ at the position $\boldsymbol{r}$ of the particle at time t.

31. Probability flux

We end this Part 3 of the book by noting why the result

$$\boldsymbol{v} = -D\frac{1}{p}\nabla p$$

for the expected particle drift velocity, enables us to view the diffusion equation

$$\frac{\partial p}{\partial t} - D\nabla^2 p = 0$$

as an equation expressing the conservation of probability.

This is because we can rewrite the diffusion equation as

$$\frac{\partial p}{\partial t} - \nabla.(p\boldsymbol{v}) = 0$$

which is the same equation as we get for conservation of mass in a compressible flow, in which $p(\boldsymbol{r}, t)$ would stand for the matter density and $p\boldsymbol{v}$ for the mass flow density.

Using Guass's theorem, the above equation for probability conservation can be written in integral form:

$$\frac{\partial}{\partial t}\int_V pdV = -\oint_S p\boldsymbol{v}.d\boldsymbol{S}$$

where V is the volume of space within and closed surface S; $d\boldsymbol{S}$ being an outward pointing element of surface area.

Assuming the probability density falls of sufficiently fast for the surface integral to tend to zero as S expands out to infinity, we have

$$\frac{\partial}{\partial t}\int_V pdV = 0$$

On account of normalisation of the probability density.

Quantum Mechanics

While the mathematical formalism of non-relativistic quantum mechanics (QM) is well established, its interpretation is not. Many have questioned the modern-day interpretation and tried to come up with alternatives. This is not the place to give an account of the many interpretations that have been offered. Sufficient to say that, in the author's opinion at least, none of them have been really convincing.

In this last part of the book, we show how, by adopting a rational Bayesian approach to probability (in which a wave function represents an observer's *knowledge* of the state of a system rather than the state itself)[34], we no longer have to face paradoxes which otherwise arise when we try to maintain realism.

The 'double-slit experiment', 'Schrödinger's cat', 'Wigner's friend' and the 'Kochen-Specker theorem' pose paradoxes of this kind, and it is easy to see why we do not have to face them.

The paradox posed by Bell's inequality, is not so easily eliminated. But we explain how it can be done when

[34] This the interpretation of the wave function was advocated by Rudolf Peierls, see his paper 'In defence of "measurement" ', Physics World, 19-20, January (1991)

we interpretate Heisenberg's uncertainty principle in a Bayesian/realist manner.

Finally, we explain how we might get closer to solving the fundamental problem of separating out from the quantum mechanical formalism, its physical and probabilistic aspects[35], by rethinking Feynman's probability amplitude theory from a Bayesian perspective.

We assume the reader is already acquainted with the formalism of (non-relativistic) quantum mechanics and its modern-day interpretation, and is familiar too, with the aforementioned paradoxes that arise when a realist interpretation is attempted.[36]

We begin by showing how the paradox posed by the double-slit experiment vanishes when we use Bayesian probability theory.

1. The double slit experiment

[35] As Jaynes has aptly remarked: '…it is the job of the laws of physics to describe physical causation at the level of ontology, and the job of probability theory to describe human inferences at the level of epistemology. The Copenhagen theory scrambles these very different functions into a nasty omelette in which the distinction between reality and our knowledge of reality is lost.'

[36] Accounts of these paradoxes can be found is many books, e.g. Hughes R.I.G., 'The structure and interpretation of quantum mechanics', Harvard Univ. Press (1989); d'Espagnat B., 'Conceptual foundations of quantum mechanics', Benjamin Inc.(1971); Redhead, M., 'Incompleteness, nonlocality and realism', OUP (1987)

The double slit experiment is often said to prove that a particle cannot be moving along a definite path. If it did, the argument goes, there would be no interference of the single-slit probability distributions over the screen.

Figure of double slit exp. Showing packets, one before the screen one emerging from each (numbered) slit and the latter two overlapping in front of the screen. Slits numbered 1 and 2

The argument employs ordinary probability theory. It makes use of ordinary conditional probabilities, that is, probabilities which apply when a certain physical condition is met; in this case the physical condition that the particle passes through a particular slit.

Of course, conditional probabilities do not have to be employed in the derivation of the probability distribution over the screen. The rules of the usual quantum mechanical formalism can be used, without mention of conditional probabilities. The result is the same for the Bayesian reasoner as it is for the non-Bayesian reasoner. Both predict the same pattern of probability distribution over the screen; one that resembles the pattern arising when waves from each slit interfere. The prediction is, in each case, arrived at by applying the time-dependent Schrödinger equation to

calculate the wave function representing the particle as it approaches and passes through the slits. The pattern of probability over the screen is then the squared modulus of interfering components of the wave function over the screen; and this agrees with experimental measurements.

But when conditional probability is employed by the non-Bayesian reasoner to predict the probability distribution over the screen, a contradiction arises.

The non-Bayesian reasoner first uses the Schrödinger equation to calculate the probability distribution over the screen on condition the particle passes through slit 1. Then on condition it passes through slit 2. The net probability distribution over the screen is then predicted to be the mean of the conditional distributions calculated for each slit.

So, for the non-Bayesian reasoner, there is now a contradiction. They are led to a probability distribution exhibiting no interference, contrary to experimental results. They therefore conclude that the particle cannot be moving on a path passing through one or other slit.

For the Bayesian reasoner, however, there is no contradiction. This is because, for the Bayesian reasoner, conditional probabilities are based not on postulated physical conditions, but on actual *knowledge* of those conditions. So, the component probability distributions are now conditional, not on which slit the particle is supposed to pass through, but on the observer *knowing*

which slit the particle passes through. This requires the observer to place particle detectors near each slit so they can find out. Then, of course, application of the ordinary quantum mechanical formalism, leads to a probability distribution over the screen different from before, but still in agreement with experimental measurements. So there is no contradiction at all.

The paradox of the double slit experiment vanishes when using Bayesian reasoning, and the particle *can* be moving on a path passing through one slit or the other.[37]

We may consider, in more detail, the double slit theory from the point of view of the Bayesian observer. This observer must base their probabilities on the knowledge they hold. In the first place, this is the initial knowledge

[37] An objection might be made regarding this Bayesian resolution of the paradox. Suppose, every time the experiment is run, the observer makes no attempt to detect which slit the particle passes through. But that, unknown to them, someone else does. The Bayesian observer will then predict interference, contrary to experimental results. Is this not a contradiction in the Bayesian approach? Well no. As we have said before, observers with different knowledge will naturally come to different conclusions. Based on their knowledge, an observer may believe certain things will happen. But that does not mean those things actually will; and when they don't, the observer looks for any information they have missed, hoping to understand why they were mistaken.

represented by the wave packet of the particle as it approaches the slits.

If they make no attempt to detect the particle, they hold no further knowledge and applying the time-dependent Schrödinger equation they correctly predict interference to occur.

If they arrange to detect the particle at one or other slit in as non-intrusive manner as possible, then they get to know which slit the particle passes through every time the experiment is run. and so must derive their expected probability distribution on the screen using the theory of mixed states. This means setting up a mixed state array with two equal probabilities associated with two normalised wave functions representing the wave packets emerging out of one or other slit.[38] This mixed state analysis leads them to correctly predict no interference of probabilities.

Therefore, when using Bayesian theory, the assumption that a particle has a definite position at any one time and moves along a definite path may be retained. This assumption is consistent both with the fact that a particle

[38] These wave packets will not be identical to the packets that emerge out of the slits when no attempt is made to detect which slit the particle passes through. Each will be slightly different on account of the disturbance caused by detection, even when this is conducted in an unintrusive manner as possible. This difference is not, however, important for the purpose at hand.

is always found at a single point in space when its position is measured, and with the fact that a particle is found to move along a path when its position is continuously measured only to classical accuracy, even if the path the particle follows might, in detail, be somewhat different from the path it would have followed without the imposition of the position measurements.

2. Schrodinger's cat

The paradox involving Schrödinger's cat, starts by claiming (as one might in principle) that we have acquired full quantum mechanical knowledge of the initial state of both the cat and the apparatus that may or may not deliver the poison. So, we have somehow determined the initial wave function for everything inside the box.

The whole system, cat and apparatus, may be viewed as three systems initially isolated from each other. The wave function of the whole system is thus the product of three wavefunctions. The first is a quasi-classical wave function representing the cat. The second is a quasi-classical wave function representing our knowledge of the apparatus. The third is a wave function of the initially isolated system which may or may not emit a particle that interacts with and triggers the apparatus to deliver the poison.

Using the Schrödinger equation, we can in principle, calculate the time evolution of the whole wave

function. This would necessarily lead to a supposition of two wave functions. In one the poison is not delivered and the cat lives, and in the other the poison is delivered and the cat is dead. The two wave functions are clearly well separated from each other in the Hilbert space of the whole system.

We gather further information when the box is opened, and use this to select which of the two aforesaid wave functions to retain. The wave function therefore 'collapses' when the box is opened. So, it appears that a physical property of the cat (being dead or alive) only manifests itself after observation.

But, for the Bayesian reasoner, there is no mystery here. The cat is either alive or dead before the box is opened. Its destiny is not determined by opening the box. All that happens is that, on seeing the contents of the box, we learn something we did not know before, and this compels us to select one of the two wave functions of the superposition that formed the wave function before opening the box.

3.Wigner's friend

The paradox involving Wigner's friend, is similar to that of Schrödinger's cat, but differs from it in that there are two observers. Here is a version of it.

It is supposed that Wigner has somehow got to know the wave function of the content of the closed-off laboratory where Wigner's friend and the apparatus she has before her are situated. That is, it is supposed that the initial wave function of all within the laboratory, including his friend, has somehow been obtained by Wigner, who sits outside unable to watch his friend's experiment.

We suppose, similarly, that Wigner's friend has somehow obtained the wave function of the apparatus alone. The apparatus is the same apparatus as that used in the case of Schrödinger's cat.

Wigner's friend waits till an agreed upon time t_1 before removing the lid of the box to observe the cat. Time t_1 is, of course, the time when her wave function collapses.

Wigner, on the other hand, can only employ Schrödinger's equation to calculate the wave function of all in the laboratory. At a time t_2, greater than t_1, this wave function is a superposition of two wave functions, one in which the cat is dead and his friend is aware of it, and another in which the cat is alive and his friend is aware of it. When Wigner opens the door of the laboratory and looks at the open box, he learns, we suppose, that the cat is alive. It is only then that his wave function collapses.

To the followers of mainstream QM, a paradoxical question now arises. At what time was the condition of the

cat physically determined? Was it when Wigner's friend opened the box to see the cat, or when Wigner opened the laboratory door to see the cat?

In QM as normally interpreted, the question itself is paradoxical and can't be answered when a wave function represents the physical state of a system, and wave function collapse is thought of as a physical process bring about a new state.

But for the Bayesian reasoner, a wave function represents an observer's knowledge, and collapse of a wave function represents only a change in that knowledge. The time of the collapse is different for the two observers because they learn something new at different times.

As the collapse of the wave function is (according to the Bayesian interpretation) not a physical event, but represents only a change in the observer's knowledge, there is no reason why the time of its occurrence cannot be different for the two observers.

The actual condition of the cat is determined physically by the apparatus at some time before anyone learns about it.

4.The claim of Kochen and Specker

Kochen and Specker claimed to be able to prove that, at any particular time, properties of a system cannot all be possessed by the system. Their proof is explained in

many text books discussing the interpretation of QM[39] and at first glance, it does seem to block the way to a realist interpretation.

The proof starts by assuming all properties *are* possessed by a system. Then, applying the quantum mechanical formalism and the normal interpretation of it, it derives a contradiction. There are several equivalent ways of arriving at a contradiction. We describe one of them[40], and that will be sufficient for the purpose at hand.

Let one property of the system be represented, in the Hilbert space, by a basis $|x_i\rangle$ $(i = 1,2,..N)$. Now, regardless of how the system has been prepared, we can imagine labelling each ket $|x_i\rangle$ of the basis with a number, 1 or 0, according as the eigenvalue x_i is the one possessed by the system at that time or not. So 1 means it is, and 0 means it isn't, and only one of the eigenkets is labelled 1. Similarly, let each ket $|y_j\rangle$ $(j = 1,2,..N)$ of a basis associated with another property be labelled with a

[39] See, for example, Redhead, M., Incompleteness, Nonlocality and Realism. OUP (1990), Belinfante, F. J., A survey of hidden-variables theories, Pergamon Press (1973), or Hughes, R. I. G., The structure and interpretation of quantum mechanics. Harvard University Press (1989).

[40] This is the one described in section 3.5 of Belinfante, F. J., A survey of hidden-variables theories, Pergamon Press (1973)

number, 1 or 0, according as the eigenvalue y_j is the one possessed by the system or not.

Now if the bases share an eigenvector, so that $|y_k\rangle = e^{i\alpha}|x_l\rangle$, for some k and some l, then, by the usual interpretation, $|y_k\rangle$ and $|x_l\rangle$ represent the same physical state of the system. This means that when property x_l is possessed by the system then property y_k must also be possessed by it. Or, looked at another way, when $|y_k\rangle = e^{i\alpha}|x_l\rangle$, the quantum mechanical formalism tells us that $|\langle x_l|y_k\rangle|^2 = |\langle y_k|x_l\rangle|^2 = 1$. So, using the ordinary (non-Bayesian) interpretation of probability, this confirms that when property x_l is supposed to be possessed by the system, then property y_k must also be possessed by the system. Note also, that when property x_l is supposed *not* to be possessed by the system, then property y_k will not be possessed by it either.

It follows, that the eigenvectors $|x_l\rangle$ and $|y_k\rangle$ must be labelled similarly. Both might be labelled 1, or both might be labelled 0.

In the many bases representing properties, cases of shared eigenkets are plenty. As we are supposing that all properties are possessed by the system, one eigenket must be labelled 1 in every basis of the Hilbert space, and eigenkets shared by any two bases must be labelled similarly. The question therefore arises as to whether such labelling is always possible. In fact, it is not always possible. A mathematical theorem proved by Kochen and

Specker, shows it is not possible in any Hilbert space of dimension $N \geq 3$.

This naturally led Kochen and Specker to claim they had shown that properties of a quantum mechanical system cannot be actually possessed by it.

The above proof does not, however, go through when we interpret the quantum mechanical formalism in the Bayesian/realist manner. The difference occurs with regard to what we think a ket vector represents, or how we interpret the conditional probability $|\langle x_l|y_k\rangle|^2$. In the proof, the ket vectors represent physical states of the system, and $|\langle x_l|y_k\rangle|^2$ is interpreted as the frequency of occurrence of property x_l on condition property y_k is present. One or the other of these supposed facts are crucial to the proof, and either would have seemed natural to Kochen and Specker.

But in the Bayesian/realist theory, kets $|x_l\rangle$ and $|y_k\rangle$ are not representing possible states of the system, but only possible states of our *knowledge* of the system; and the probability $|\langle x_l|y_k\rangle|^2$ is for us, our degree of belief that property x_l is present when we know property y_k is present. So, we can believe x_l is present only after measurement of property y gives y_k. Similarly, we can believe y_k is present only after measurement of property x gives x_l. Without measurements, we have no reason to think that properties x_l and y_k must occur together.

It is therefore not necessary to label shared eigenkets similarly; and the argument of Kochen and Specker cannot be carried through.

5. Heisenberg's uncertainty principle.

Before tackling the paradox posed by Bell's inequality, it is necessary first to work out what Heisenberg's uncertainty principle should mean to the Bayesian/realist, and what this might imply.

Heisenberg's uncertainty principle states that it is impossible in any way to measure with infinite precision, 'incompatible observables' simultaneously. For example, we cannot measure precisely, both the position and the momentum of a particle at the same time. There are, instead, specified limits to the accuracy to which this may be done.

In the usual interpretation of QM, this is taken to imply that properties can't be actually possessed by a system. Instead, a property is thought to come about and take a definite value only when a measurement of it is performed.

In a realist interpretation of QM, it is natural to think properties take on definite values independently of measurement. It is natural, also, to interpret Heisenberg's uncertainty principle as setting limits to the *knowledge* we can hold of physical properties; and to say, for example, that we cannot simultaneously hold knowledge of both the

precise position and the precise momentum of a particle on account of us not being able, in principle, to measure both at once. It is only in the classical limit of QM that we can get to know the values of incompatible properties to high accuracy (i.e. to accuracy considered good in the context of classical physics).

On the basis of the Bayesian/realist interpretation of Heisenberg's uncertainty principle, we might seek to model measurement processes, and to demonstrate *why* there are limits to the accuracy to which the values of incompatible properties can be known.

Now in adopting the realist interpretation of Heisenberg's uncertainty principle, a problem arises in connection with the product rule of probability.

For suppose A, B and Y are incompatible propositions i.e. propositions claiming precise values of incompatible physical properties such as the spin components of a particle in three different directions. Then, the product rule

$$P(AB|Y) = P(A|Y)P(B|AY) = P(B|Y)P(A|BY)$$

is meaningless in Bayesian theory, because knowledge AY, or knowledge BY on the RHS, is knowledge we can never hold. Probabilities $P(B|AY)$ and $P(A|BY)$ are therefore nonexistent. Also, the value of $P(AB|Y)$ can't

be calculated using the product rule. Neither can it be evaluated by counting the frequency at which event AB occurs in many trials, because we can never know when A and B have occurred together. For these reasons, we are led to introduce a new rule of Bayesian reasoning:

> *Whenever the truth of the conjunction of two or more propositions cannot be tested experimentally on account of Heisenberg's uncertainty principle, the probability of their conjunction under any knowledge is nonexistent, and even if we believe in the propositions separately, it does not follow that we should believe in their conjunction.*

We have no reason to think, however, that $P(A|Y)$ and $P(B|Y)$ are nonexistent. Probabilities of this form feature, after all, in the quantum mechanical transformation equations from one representation to another.

For example, consider a case in which wave functions of a system are functions of a single variable with n possible values. Under pure knowledge Y the wave function $\psi(x_i|Y)$ $(i = 1,2 \dots n)$ in the $x$representation and the equivalent wave function $\phi(y_j|Y)$ $(j = 1,2 \dots n)$ in the $y$representation are related to one another through the transformation equations

$$\phi(y_j|Y) = \sum_{i=1}^{n} f(y_j|x_i)\psi(x_i|Y)$$

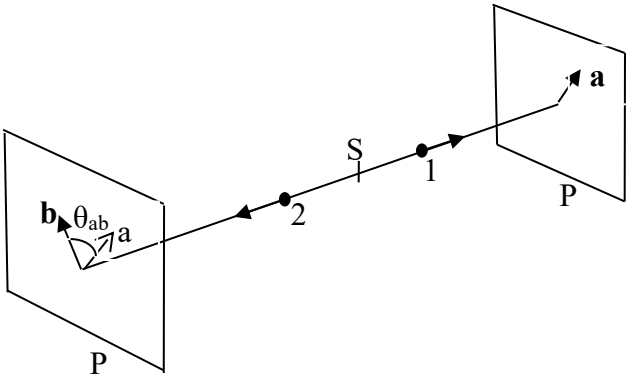
$$\psi(x_i|Y) = \sum_{j=1}^{n} f^*(x_i|y_j)\varphi(y_j|Y)$$

where $f(y_j|x_i)$ is a known transformation function. Born's rule then gives the probability $P(y_j|x_i)$ as

$$P(y_j|x_i) = |f(y_j|x_i)|^2$$

6. Bell's inequality

This inequality relates to a system composed of a pair of spin one-half particles produced in a singlet spin state and sent flying out from a source S in opposite directions towards apparatus designed to measure particle spin components in any directions ($\boldsymbol{a}$ and $\boldsymbol{b}$) in planes P perpendicular to their line of motion.



The translational motions of the particles are independent of their spinning motions, and we are interested here only in the latter.

The argument leading to the Bell inequality uses the ordinary theory of QM and the ordinary rules of logic and probability. We first give an account of the argument, then show why it is invalid under the proposed Bayesian way of reasoning.

In the four-dimensional Hilbert space of the spinning motion of the particles there is a basis whose four eigenvectors are labelled by spin components of particles 1 and 2 in the directions $\boldsymbol{a}$ and $\boldsymbol{b}$ respectively. There is another basis whose four eigenvectors are labelled by spin components of particles 1 and 2 in alternative directions $\boldsymbol{a'}$ and $\boldsymbol{b'}$. These bases refer to different, and incompatible properties of the two-particle system.

With chosen directions, $\boldsymbol{a}$ for particle 1 and $\boldsymbol{b}$ for particle 2, let the spin components in these directions be measured, and suppose the experiment is repeated a large number times, say M times, under the same singlet state. Let a_i denote the result of the measurement on particle 1 in the i^{th} trial. The measurement result a_i will be, let us say, $+1$ (spin up) in the $\boldsymbol{a}$ direction or -1 (spin down) in the direction $-\boldsymbol{a}$'.

Let b_i denote, in the same way, the result of the measurement on particle 2. The values of a_i and b_i thus obtained label an eigenket of the first basis mentioned above.

Quantum mechanics predicts the correlation coefficient between a_i and b_i to be $-\cos\theta_{ab}$, where θ_{ab} is the angle between the directions $\boldsymbol{a}$ and $\boldsymbol{b}$. Since the mean value and variance of a_i (and of b_i) are 0 and 1 respectively, the correlation coefficient is equal to[41]

$$\frac{1}{M}\sum_{i=1}^{M} a_i b_i = -\cos\theta_{ab}$$

[41] For a proof of this result using quantum mechanics see, for example, p.41 of Readhead, M., 'Incompleteness, nonlocality and realism' OUP, (1987) loc. cit.

in the limit the as $M \to \infty$. The LHS is a property of an ensemble, i.e. of an infinite set of repetitions of the process, all carried out with a fixed chosen angle θ_{ab}.

Now, if it is supposed that components of spin are properties possessed by the particles, then a_i and b_i have definite values each time the process is run, and the above equation will hold (at least for a large enough number M of trials) whether or not the spin components are measured.

Taking another pair of directions $\boldsymbol{a}'$ and $\boldsymbol{b}'$ in the planes P, each of the following relations hold also.

$$\frac{1}{M}\sum_{i=1}^{M} a_i b_i = -\cos\theta_{ab}$$

$$\frac{1}{M}\sum_{i=1}^{M} a_i b_i^{'} = -\cos\theta_{ab'}$$

$$\frac{1}{M}\sum_{i=1}^{M} a_i^{'} b_i = -\cos\theta_{a'b}$$

$$\frac{1}{M}\sum_{i=1}^{M} a'_i b'_i = -\cos\theta_{a'b'}$$

Combining these,

$$\frac{1}{M}\sum_{i=1}^{M} S_i = -\cos\theta_{ab} - \cos\theta_{ab'} - \cos\theta_{a'b} + \cos\theta_{a'b'}$$

where the summand

$$S_i = a_i(b_i + b'_i) + a'_i(b_i - b'_i)$$

refers to spin components in the i^{th} trial.

As only one of the bracketed terms in the expression for S_i can be non-zero and equal to ± 2, it follows that $S_i = \pm 2$ for each i. As a result,

$$|-\cos\theta_{ab} - \cos\theta_{ab'} - \cos\theta_{a'b} + \cos\theta_{a'b'}| \leq 2$$

This is a Bell inequality. It is not satisfied for all values of the angles, so there is a contradiction.

Therefore, the spin components a_i, b_i, a'_i or b'_i in any one trial, cannot be possessed properties, so, the quantum mechanical formalism can't be interpreted in a realist manner.

However, if in reasoning about the physical world, we take the Bayesian/realist approach including the new rule in Section 4, the inequality is not demonstrable. For, taken together, the equations giving the correlation coefficients for any four angles θ_{ab}, $\theta_{ab'}$, $\theta_{a'b}$ and $\theta_{a'b'}$, involve

incompatible properties (for example, a_i and a'_i). So, while we can (and do) believe in the proposition claiming the truth of the equation for any one of the four correlation coefficients, we are not obliged to believe in their conjunction. The probability of the conjunction of all four propositions is non-existent. So, for us, Bell's inequality is not provable.

Other derived Bell-like contradictions[42], that seem to show the impossibility of a realist interpretation of QM, are unprovable for the same reason.

We have said that the Bayesian/realist approach to QM, allows the modelling of measurement processes. In the next three Sections, we give examples of how such modelling.

7. Measurement in general

Within ordinary QM it is not possible to model measurement processes because they involve wave function collapse, which operate outside the laws of QM.

Using the Bayesian/realist approach, however, it is possible to model measurement processes, to show

[42] See for example, 'Bell's theorem without inequalities.' by Greenberger, D. M., Horne, M. A., Shimony, A. and Zeilinger, A.; Am. J. Phys. 58 (12), 1131-1143 (Dec 1990), and 'Is the moon there when nobody looks?' Physics Today, vol. 38, p. 38 (April 1985).

how, at least in principle, we can get to know precise properties of systems at the quantum level.

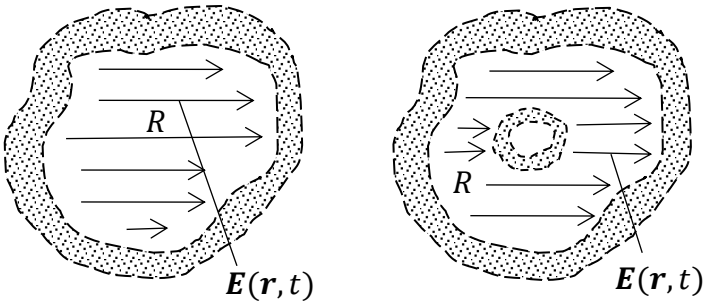
In the modelling of measurements, we may employ any apparatus operating so well into the classical limit that knowledge of its properties is not subject to limitation. We are free to suppose such apparatus produces an external electric or magnetic field acting on particles of any quantum mechanical system, and can be classically manipulated for the purpose of making the measurement.

Of course, in non-relativistic atomic and molecular physics, the particles (electrons and nuclei) have charges that are multiples of the electronic charge and masses that have particular values, and the inter-particle potentials are of Coulomb type. This is not, however, a requirement of quantum mechanics. The only fundamental constant of the theory is Planck's constant, and there is, in principle, no theoretical restriction on the values of charges or masses nor on the magnitude and distribution of external or inter-particle potential fields.

7. Measurement of the position a charged particle

By employing particles of very high mass, carrying charges and sensitive to appropriate inter-particle potentials of non-electromagnetic kind, we can 'construct' means for generating any specified time dependent electric field $\boldsymbol{E}(\boldsymbol{r}, t)$ in a region R of space. We

just have to imagine an assembly of charged particles of high enough mass to be treated classically and to be moved about as we please. Using a large number of densely packed particles of positive and negative charge, we can form a layer, of distributed surface charge and dipole density, over the boundary of the region R.



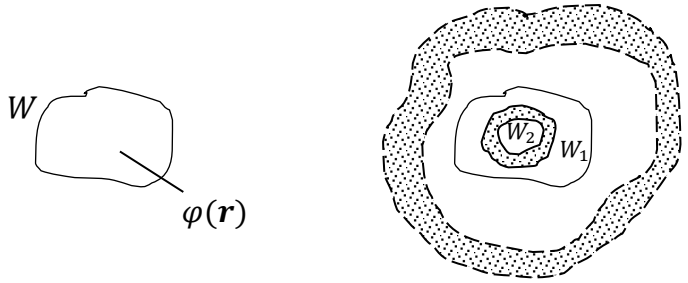
Crowd of charged particles in a thin layer bounding region R and producing a uniform electric field $\boldsymbol{E}(\boldsymbol{r}, t)$ in the region R. Left: Simply bounded region. Right: Region between two boundaries. In the figure the layers of particles are shown with a greatly exaggerated thickness.

At any point on the boundary of R, the jump in electric potential and the jump in the normal component of electric field across the boundary of R, can be controlled by moving the charges about in the layer. With the sum of all particle charges being zero, we can thus produce a large pulsed uniform electric field in the x direction within region R and no electric field outside it.

Ideally, we should let the masses of the particles tend to infinity, their number tend to infinity and their charge tend to zero. The electric field they produce is then a collective effect and the field of any one particle is vanishingly small and has no effect on a charged particle whose position is to be measured.

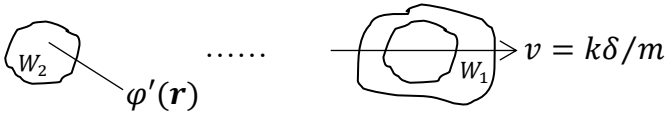
We control our 'apparatus' so it sits dormant producing no electric field till, at a certain time, it generates a pulsed uniform electric field $\boldsymbol{E}(\boldsymbol{r}, t)$ all across region R, and then sits dormant again.

Now, envisage a free quantum mechanical particle of mass m, for example an electron. Suppose we hold a pure state of knowledge of its motion represented by a wave function $\varphi(\boldsymbol{r})$ just before the measurement. Let region R enclose the classically tiny region W in which $\varphi(\boldsymbol{r})$ is essentially confined, and let p be the expected absolute value of the particle's momentum going with $\varphi(\boldsymbol{r})$. Let our apparatus be placed so that the electric field pulse appears over only a part W_1 of W.



Left: Wave function confined to region W. Right: Apparatus placed over W so that the electric field will affect only a part W_1 of W.

Before and after the pulse, our apparatus does not interact with the particle (the charge and dipole densities in the boundaries of region R being zero). The effect of the pulse can be calculated using the Schrödinger equation in which, owing to its very large magnitude, the electric potential dominates in the Hamiltonian during the short time of the pulse. The effect is to leave unchanged the part W_2 of the wave function that was not in the electric field, and to multiply the part W_1 of the wave function, that was in the field, by a phase factor $e^{ikx\delta/\hbar}$, where k is the electric field strength times the particle's charge and δ is the duration of the pulse. With $k\delta \gg p$, this causes the part W_1 of the wave function to start moving away in the x direction at the high speed $k\delta/m$.



The part W_2 of the wave function staying behind, and the part W_1 of the wave function set in rapid motion by the electric field and, by now, far away from its initial position.

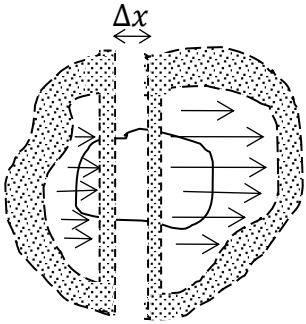
The part W_2 of the wave function we denote by $\varphi'(\boldsymbol{r})$. It occupies a smaller region of space than $\varphi(\boldsymbol{r})$ did. It starts to evolve differently as a result. This evolution is relatively slow and unimportant, while the motion of the wave packet W_1 is fast, and it is soon far away. The wave packet contains many de Broglie wavelengths and therefore, on its own, and in the largeness of space, represents a particle known to move classically, and uniformly in a straight path.

Just after the applied pulse, we believe the particle must have either remained behind in the region occupied by the 'new' wave function or have set off with the wave-packet. Supposing we fail to detect a particle in the path of the wave-packet. We then know the particle has stayed behind, and we may collapse our wave function from $\varphi(\boldsymbol{r})$ to the renormalised function

$$\frac{\varphi'(\boldsymbol{r})}{\sqrt{\int_{W_1}|\varphi'(\boldsymbol{r})|^2 dV}}$$

Our probability distribution over particle position is then sharper than before, and by diminishing the part W_2 of the region in which the electric field doesn't act, we can, by chance, get to know the particle's position to any required accuracy.

Similarly, we might get to know just the x component of the particle's position as accurately as we please, by applying our electric field pulse in two regions, one to the left and one to the right of two closely spaced planes perpendicular to the x axis and crossing W.



Apparatus used to gain knowledge only of the x coordinate of the particle.

Before modelling the measurement of a component of spin of a particle, we need to explain how 'mixed states' should be interpreted in a Bayesian/realist approach to QM.

8. Mixed states and Density matrices

In ordinary QM, mixed states are states of a system more general than those represented by ket vectors.

Any one mixed state is represented by a set of n ket vectors and associated positive weights; it is represented by an array

$$\{|\psi_1\rangle, |\psi_2\rangle, \ldots |\psi_n\rangle; w_1, w_2, \ldots w_n\}$$

where $|\psi_1\rangle, |\psi_2\rangle$,…etc. are (not necessarily orthogonal) kets representing possible pure states of the system, and weight w_1 is associated with ket $|\psi_1\rangle$, weight w_2 with ket $|\psi_2\rangle$,…etc.

The sum of the weights is always equal to 1, and associated with an array is the density operator, or density matrix,

$$\rho = \sum_{i=1}^{n} |\psi_i\rangle w_i \langle\psi_i|$$

Two arrays $\{|\psi_1\rangle, |\psi_2\rangle, \ldots |\psi_n\rangle; w_1, w_2, \ldots w_n\}$ and $\{|\phi_1\rangle, |\phi_2\rangle, \ldots |\phi_m\rangle; v_1, v_2, \ldots v_m\}$ represent the same state when their density matrices are equal:

$$\sum_{i=1}^{n} |\psi_i\rangle w_i \langle\psi_i| = \sum_{i=1}^{m} |\phi_i\rangle v_i \langle\varphi_i|$$

Ordinary QM, claims that if one or other of pure states $|\psi_1\rangle, |\psi_2\rangle, \ldots |\psi_n\rangle$ apply with respective probabilities $p_1, \ldots p_n$, then the system is in a mixed state with array $\{|\psi_1\rangle, |\psi_2\rangle, \ldots |\psi_n\rangle; p_1, p_2, \ldots p_n\}$. The weights are then the probabilities for one or other of the physical states $|\psi_1\rangle, |\psi_2\rangle, \ldots |\psi_n\rangle$ to actually apply.

But ordinary QM does not claim the converse. It does not claim that a mixed state is necessarily one in which one or other of the physical states $|\psi_1\rangle, |\psi_2\rangle, \ldots |\psi_n\rangle$ applies with a probability equal to its weight. This is because the same mixed state can be represented by one or other of two, or more arrays of different ket vectors and weights. So, if the weights were always probabilities, we would have a contradiction; for one or other of the physical states in one array would have to be present, while one or other of the (different) physical states in the other array would also have to be present.

In the proposed Bayesian/realist interpretation, however, we can (and do) claim that any mixed state is one in which

one or other of the pure states of knowledge $|\psi_1\rangle, |\psi_2\rangle, \ldots |\psi_n\rangle$ applies with a probability (i.e. degree of belief) equal to its weight. Then, it is quite imaginable that two or more equivalent arrays might represent the same state of *knowledge*. Indeed, it is necessarily the case, because equivalent arrays demonstratively generate the same probability distributions over the variables quantifying any property of the system.

Unlike a pure state, a mixed state of knowledge can be improved upon. That is, it can be brought closer to a pure state of knowledge by sharpening the probability distribution over the possible pure states in the mixed state array.

Associated with a mixed state of knowledge is an information entropy defined in the usual way as

$$\mathcal{H} = -\sum_{i=1}^{n} p_i \ln p_i$$

the p_i being the probabilities (the weights) in the mixed state array. This entropy measures the degree of our ignorance *over and above our unavoidable ignorance resulting from Heisenberg's uncertainty principle*.

The principle of indifference, the method of transformation groups, and the method of maximum entropy may all be employed to deduce the probabilities

p_i that rationally apply in any particular case. As more information comes to light, the probability distribution p_i $(i = 1,2, \dots n)$ may be sharpened.

A general ket $|\psi\rangle$, or the pure state it stands for, may be represented by a wave function; that is, by the complex values $\langle\psi|x_i\rangle$ of the projections of $|\psi\rangle$ onto each of a complete set of orthonormal kets $|x_i\rangle$ $(i = 1,2, \dots)$.

A mixed state of knowledge can thus be represented by an array

$$\{\langle\psi_1|x_i\rangle, \langle\psi_2|x_i\rangle, \dots \langle\psi_n|x_i\rangle; p_1, p_2, \dots p_n\}$$

of wave functions and their probabilities.

In the case of a single particle in motion, the property might be the particle's position $\boldsymbol{r}$, in which case the $\langle\psi_j|x_i\rangle$ are wave functions $\psi_j(\boldsymbol{r})$ $(j = 1,2, \dots n)$. A mixed state is then represented by the array

$$\{\psi_1(\boldsymbol{r}), \psi_2(\boldsymbol{r}), \dots \psi_n(\boldsymbol{r}); p_1, p_2, \dots p_n\}$$

Under conditions of logical and physical independence, our knowledge of two systems together is represented by the product of their density matrices.

For example, in some circumstances the translational and spinning motion of a particle may be

regarded as separate systems. This requires that there be no spin/orbit interaction, and that our knowledge of the spin and of the orbit be independent. Suppose the mixed state of translational motion is, at one time

$$\{\psi_1(\boldsymbol{r}), \psi_2(\boldsymbol{r}), \ldots \psi_m(\boldsymbol{r}); p_1, p_2, \ldots p_m\}$$

and the mixed state of spin component σ in a specified direction is

$$\{\phi_1(\sigma), \phi_2(\sigma), \ldots \phi_n(\sigma); q_1, q_2, \ldots q_n\}$$

where σ takes the values $\sigma = s, s-1, \ldots -s$, where s is the particle's spin. Then the mixed state representing knowledge of both position and spin at the time in question is

$$\{\psi_1\phi_1, \psi_1\phi_2, \ldots \psi_m\phi_n; p_1q_1, p_1q_2, \ldots p_mq_n\}$$

where variables $\boldsymbol{r}$ and σ are omitted for brevity, and, for example, p_1q_2 is the probability that wave function $\psi_1\phi_2$ applies. There are here, nm possible wave functions and nm associated probabilities, and it is easy to show that the density matrix is the product of the density matrices of the separate systems.

9. Measurement of a particle's spin component

Suppose a beam of silver atoms is created by vaporizing silver in an oven and allowing some of the atoms to pass through collimating slits to create a narrow beam. By subjecting the beam to a pair of opening and closing of gates, we may end up with a single silver atom which, to classical accuracy, is in rectilinear motion at a known speed.

The silver atom has a single electron in its outer shell and can be modelled quantum mechanically as a neutral particle with spin one-half.

Suppose we allow this atom to pass through the strong inhomogeneous magnetic field in a Stern–Gerlach apparatus. Then it will exit the apparatus in, roughly speaking, one of two directions according as the spin is up or down in the direction of the field gradient.

Figure of Stern-Gerlach apparatus.

Before the particle enters the apparatus, our state of knowledge of the spin component σ in any particular direction, is one of indifference. We have, initially, no idea which of the two values $\sigma = \frac{1}{2}$, or $\sigma = -\frac{1}{2}$ the spin component takes. We are thus in a mixed state of

knowledge represented by an array whose wave functions are the eigenfunctions $\delta_{\sigma,\frac{1}{2}}$ and $\delta_{\sigma,-\frac{1}{2}}$ and whose weights are, by the principle of indifference, equal to $\frac{1}{2}$. This mixed state has a density matrix equal (apart from a constant factor) to the unit matrix and is thus independent of which direction of spin is considered. In particular it applies when the direction is that of the magnetic field gradient in the Stern-Gerlach apparatus (i.e. in the $z$direction in the Figure).

Our knowledge of the translational motion of the particle before it enters the apparatus, is more precise. We have (what we view as) a classical particle moving along a known path. Working in the Schrödinger picture, our state of knowledge of the approaching particle's position and velocity might be represented by one or more compact Schrödinger wave packets $\psi(\boldsymbol{r}, t)$ in uniform motion. A pure state of our knowledge of the particle's translational motion is not present, but we can represent our knowledge by a density matrix whose array contains all forms of wave packet $\psi_i(\boldsymbol{r}, t)$ $(i = 1,2, \ldots m)$, that might reasonably represent our knowledge of the translational motion, and probabilities p_i attached to each. As there is no rational to do it, we have to allow a certain degree of subjectivity in choosing both the forms of the wave packets and their associated probabilities. This will not matter if (as is the case) conclusions drawn later are the same no matter which subjective choices were made.

Anyhow, our combined knowledge of the (independent) translational motion and spinning motion of the particle, is then represented by the product of the arrays representing the translational motion and the spinning motion separately. That is, by

$$\left\{\psi_1\delta_{\sigma,\frac{1}{2}},\ldots\psi_m\delta_{\sigma,\frac{1}{2}},\psi_1\delta_{\sigma,-\frac{1}{2}}\ldots\psi_m\delta_{\sigma,-\frac{1}{2}};\frac{1}{2}p_1,..\frac{1}{2}p_m,\frac{1}{2}p_1,..\frac{1}{2}p_m\right\}$$

where variables $\boldsymbol{r}$ and t are omitted for brevity.

During interaction with the magnet in the apparatus, any one of the wave functions, say $\psi_1(\boldsymbol{r},t)\delta_{\sigma,\frac{1}{2}}$, evolves into a wave function $\Phi_\sigma(\boldsymbol{r},t)$ which, strictly speaking, no longer factors. The time dependent Schrödinger equation applying here takes the form

$$-\frac{\hbar}{i}\frac{\partial\Phi_\sigma}{\partial t}=\sum_{\sigma'}H_{\sigma\sigma'}\Phi_{\sigma'}$$

where the Hamiltonian $H_{\sigma\sigma'}$ is given by

$$H_{\sigma\sigma'}=-\delta_{\sigma\sigma'}\frac{\hbar^2}{2m}\nabla^2-\frac{\mu}{s}\boldsymbol{\sigma}.\boldsymbol{H}$$

where m is the mass of the particle, μ its magnetic moment, $\boldsymbol{H}$ the magnetic field strength (a function of

position $\boldsymbol{r}$) and $\boldsymbol{\sigma}$ a vector with the spin matrices as its x, y and z components. As the particle is of spin one-half, the spin matrices are

$$\sigma^x_{\sigma\sigma'} = \tfrac{1}{2}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma^y_{\sigma\sigma'} = \tfrac{1}{2}\begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma^z_{\sigma\sigma'} = \tfrac{1}{2}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

They operate on Φ_σ in vector form:

$$\Phi_\sigma = \begin{pmatrix} \Phi_{1/2} \\ \Phi_{-1/2} \end{pmatrix}$$

The evolution equations for the components of the vector are therefore

$$-\frac{\hbar}{i}\frac{\partial \Phi_{1/2}}{\partial t} = -\frac{\hbar^2}{2m}\nabla^2\Phi_{1/2} - \frac{\mu}{2s}H_x\,\Phi_{-1/2} - \frac{\mu}{2s}H_z\,\Phi_{1/2}$$

$$-\frac{\hbar}{i}\frac{\partial \Phi_{-1/2}}{\partial t} = -\frac{\hbar^2}{2m}\nabla^2\Phi_{-1/2} - \frac{\mu}{2s}H_x\,\Phi_{1/2} + \frac{\mu}{2s}H_z\,\Phi_{-1/2}$$

Here, the term involving the component H_y is omitted because it is zero on account of the design of the apparatus and the midway position of the path of the wave packet along the length of magnetic field in the $y$direction.

The solution of these equations is difficult to carry through, but a qualitative treatment will suffice. In view of the purpose of the apparatus, it is the gradient of H_z in

the z direction that's important. So, we may proceed roughly by setting H_x equal to zero. The equations then separate into equations for each component. They are similar to the Schrodinger equations for a particle in potential fields $V = -\mu H_z/2s$, and $\mu H_z/2s$. The gradient of either of these potentials in the $z$direction amounts to a force on the particle that will deflect the particle's path. The deflection is upward in the case of $\Phi_{1/2}$ and downward in the case of $\Phi_{-1/2}$.

After the motion of the particle through the Stern-Gerlach apparatus, when the two possible paths of the particle are well separated from each other, we can determine the component of spin in the direction of the magnetic field gradient by detecting which path the particle takes after leaving the magnetic field. Until we make an effort to do this, we are still indifferent as to which value of the two spin components is present. We are also indifferent as to which corresponding path the particle is taking.

If we place a particle detector in one of the paths and it fails to detect the particle, then we know the particle is on the other path, and we know the value of its spin component. We have then performed a measurement of the spin component, and we have 'prepared' the particle in a way which puts us into a pure state of knowledge of its spinning motion; a state of knowledge in which we believe the spin component in the $z$direction takes a

certain value. Our state of knowledge is still, however, a mixed state. If we find the spin component is $\frac{1}{2}$, for example, the new mixed state is represented by the array

$$\left\{\psi_1\delta_{\sigma,\frac{1}{2}}, \ldots \psi_m\delta_{\sigma,\frac{1}{2}};\ p_1, .. p_m\right\}$$

The measure of our unavoidable ignorance of the particle's dynamics is thus reduced from

$$-2\sum_{i=1}^{m}\frac{1}{2}p_i \ln\frac{1}{2}p_i \quad \text{to} \quad -\sum_{i=1}^{m} p_i \ln p_i$$

That is, the information entropy falls by $\ln 2$.

10. Experiments confirming Bayesian reasoning

In classical physics, an experiment carried out to test a theory, tests both the physical laws of the theory and the rules of reasoning used in applying them.

Similarly, an experiment carried out to test quantum theory, tests both the physical laws of that theory, and the rules of reasoning used in applying them; rules we take to be based on Bayesian probability.

We now give two examples of experiments that can be said to test the validity of our Bayesian/realist reasoning.

Beam splitter experiment

Suppose a Schrödinger wave packet representing our pure knowledge of a particle's translational motion, breaks into two wave packets as a result of passing through a beam splitter.

Figure

Beam splitter Figure. Packets numbered 1,2,3. No need for caption.

At the start, wave packet 1. represents our knowledge regarding the particle's position. The wave packet evolves in time according to the Schrödinger equation. After passage of the packet through the splitter, the wave function is made up of two wave packets, packets 2. and 3. This leads us to believe that the particle leaves the beam splitter in one or other of two paths.

Now suppose we try to detect the particle at a place along one path where we would expect it to be, had it chosen that path. Suppose we fail to detect it. Then we come to believe the particle went the other way.

The detector would not have affected the motion of the particle any more than would other objects

positioned off the path the particle takes.[43] We can take it that our measurement has produced no effect on the particle motion. But we are logically led to change our wave function to the renormalised wave packet in the other path.

The validity of our reasoning is confirmed when we find the subsequent stochastic behavior of the particle after the null detection is accurately predicted by applying the Schrödinger equation to the single renormalised wave packet. We find this is so, whatever process the particle might subsequently undergo. In every case we get agreement with the predictions of QM following from the premise that the single renormalised wave packet represents our knowledge of the particle's motion after the null detection.

Particle interferometer

The particle interferometer provides a clearer way to think about the paradox associated with the double slit experiment. This is because the wave packets into which an incident wave packet splits, are better separated.

[43] Of course, any Schrödinger wave packet, though small in size, is, as it moves, actually spread out over the whole of space (unless confined by a closed boundary). So, any body, even one very far away, must influence the wave packet's wavefunction to some extent. As the wavefunction falls off very quickly as we move away from the packet, the effect of bodies near to, but not actually in the path of the packet, can be ignored for present purposes.

Figure

Interferometer showing [5] wave packets at various stages of the particle motion. The dotted packet is non-existent, being a supposition of waves of equal and opposite phase. A particle detector D shown dotted, may or may not be present.

We suppose the wave packets emerging from the first beam splitter are of equal amplitude, and there is a high degree of coherence between them. Let the positions of the mirrors be adjusted accurately enough to bring about phase cancelation in one of the paths leaving the final beam spitter. Then, packet 2 (shown dotted in the Figure) is absent, and we believe the particle will always leave the final beam splitter in one particular way.

Now suppose we place a particle detector D in one arm of the interferometer. If the detector fails to detect the particle, it must have gone along the other arm of the interferometer; and we come to believe it is now equally likely to leave the final beam splitter in *either one* of the two possible ways.

But how can it be that the null-detection, which does not affect the particle motion, results in the particle leaving the final beam splitter *either* way?

Ordinary quantum theory answers this question in a rather unsatisfactory way. It takes wave functions to represent physical states, and supposes the wavefunction collapse occurring at the time of the null detection, is due to some physical process operating outside the theory. This outside process, triggered by our null-detection, *does*, it is argued, affect the particle motion, so that its motion is now represented by the renormalised wave packet in the other arm of the interferometer. In this way of thinking, a mysterious outside influence allows the particle to pass either way through the final beam splitter.

Bayesian/realist quantum theory answers the question in another and arguably better way. It claims that after the null detection, we are led *logically* to believe that the particle may take either path out of the final beam splitter. The absence of a physical cause of the unexpected particle behavior is of no matter when there is a logical reason for it. After null-detection it is logical reasoning (which includes the calculation of probabilities using the Schrödinger equation), that leads us to believe that the particle may exit the final beam splitter either way; and this belief is confirmed by experiment. That's all there is

to it. There is no need to find a physical cause of the particle's behavior because it already follows logically.

11. Should probabilities have complex values?

It was Feynman who first treated wave functions as probability distributions. He called them 'probability amplitudes', but noted how they followed rules similar to those followed by ordinary (real valued) probability distributions.

Feynman pointed out that a transformation equation in QM, like the first of those considered in Section 5:

$$\phi(y_j|Y) = \sum_{i=1}^{n} f(y_j|x_i)\psi(x_i|Y)$$

resembles a relation holding in ordinary probability theory between probability distributions $P(x_i|Y)$ and $P(y_j|Y)$ $(i, j = 1, \ldots n)$ under the same conditions Y. For, by the product rule we would have

$$P(y_j x_i|Y) = P(x_i|Y)P(y_j|x_i Y)$$

and using the sum rule, we would have

$$P(y_j|Y) = \sum_{i=1}^{n} P(x_i|Y)P(y_j|x_iY)$$

which, when Y is redundant in the conjunction x_iY, becomes

$$P(y_j|Y) = \sum_{i=1}^{n} P(x_i|Y)P(y_j|x_i)$$

Although the resemblance is clear, Feynman didn't directly take wave functions to be complex valued probability distributions. He chose, instead, to regard them as abstract 'probability amplitudes', from which the actual probability distributions could be arrived at by forming their squared moduli.

The reason Feynman did not think wave functions themselves where probability distributions, was no doubt because, like most physicists, he took probabilities to be relative frequencies. So, for him, probabilities were necessarily real.

But in our Bayesian/realist interpretation of QM, wave functions (like all probability distributions) represent states of knowledge. So we might reasonably claim that a possible state of knowledge of a quantum mechanical system is represented by a complex-valued probability distribution; and that our corresponding

degree of belief distribution is the squared modulus of the probability distribution.

Taking up this idea, a wave function in the $x$representation is a complex-valued probability distribution which we write as $\Phi(x_i|Y)$, and the corresponding degree of belief distribution is written $P(x_i|Y)$ and is given by $P(x_i|Y) = |\Phi(x_i|Y)|^2$.

12. Could complex-valued probabilities help establish a satisfactory realist interpretation of QM?

That is, could we propose certain *purely physical* laws governing the properties of atoms and molecules, and then apply *complex-valued probability theory* to reproduce all the predictions of QM?

At the heart of QM are the transformation functions from one representation to another. So, to achieve a realist interpretation of QM, it may be enough to show how each transformation function can be derived by proposing (i) physical laws relating property values in one representation to property values in the another, and (ii) laws of reasoning employing complex-valued probabilities.

To take a particular case, let's consider the z components of spin, σ and σ', of a spin one-half particle, relative to Cartesian coordinate frames O and O' of different orientation.

From ordinary quantum theory, the transformation equation for the spin components from O and O' is

$$\Phi(\sigma'|Y) = \sum_{\sigma=-1/2}^{1/2} \Phi(\sigma'|\sigma)\Phi(\sigma|Y)$$

The transformation function being given in matrix form by

$$\Phi(\sigma'|\sigma) = \begin{pmatrix} \Phi(\frac{1}{2}'|\frac{1}{2}) & \Phi(\frac{1}{2}'|-\frac{1}{2}) \\ \Phi(-\frac{1}{2}'|\frac{1}{2}) & \Phi(-\frac{1}{2}'|-\frac{1}{2}) \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

where

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} \cos\frac{\alpha}{2}\, e^{i(\beta+\gamma)/2} & i\sin\frac{\alpha}{2}\, e^{-i(\beta-\gamma)/2} \\ i\sin\frac{\alpha}{2}\, e^{i(\beta-\gamma)/2} & \cos\frac{\alpha}{2}\, e^{-i(\beta+\gamma)/2} \end{pmatrix}$$

Here α, β, γ are the Euler angles taking us from Cartesian frame O to Cartesian frame O′, by first rotating frame O through angle β about the z axis, then through angle α about the new x axis, and finally, through angle γ about the new z axis.

We will derive the transformation function $\Phi(\sigma'|\sigma)$ in a way similar to that used by Feynman in Vol. III of his Lectures on Physics.

Rotations about the z axis.

Let us start by consider the case of a simple rotation through angle β about the z axis. We have then that

$$\begin{pmatrix} \Phi(\frac{1}{2}'|\frac{1}{2}) & \Phi(\frac{1}{2}'|-\frac{1}{2}) \\ \Phi(-\frac{1}{2}'|\frac{1}{2}) & \Phi(-\frac{1}{2}'|-\frac{1}{2}) \end{pmatrix} = \begin{pmatrix} e^{i\beta/2} & 0 \\ 0 & e^{-i\beta/2} \end{pmatrix}$$

And therefore

$$\begin{pmatrix} \mathrm{P}(\frac{1}{2}'|\frac{1}{2}) & \mathrm{P}(\frac{1}{2}'|-\frac{1}{2}) \\ \mathrm{P}(-\frac{1}{2}'|\frac{1}{2}) & \mathrm{P}(-\frac{1}{2}'|-\frac{1}{2}) \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

This suggests a physical relation between the z components of spin in frame 0 and frame $0'$. Physically it seems that $\sigma' = \sigma$. So perhaps we should propose a physical law taking the propositional form:

$$"\sigma = \pm 1/2" \implies "\sigma' = \pm 1/2"$$

This would account for why our degrees of belief $\mathrm{P}(\frac{1}{2}'|\frac{1}{2})$ and $\mathrm{P}(-\frac{1}{2}'|-\frac{1}{2})$ are both 1. But it would not account for the phase factors in the probabilities $\Phi(\frac{1}{2}'|\frac{1}{2})$ and $\Phi(-\frac{1}{2}'|-\frac{1}{2})$.

But let us be bold and claim that logical implications should actually carry phases. Then the above physical law can be stated thus

$$"\sigma = \pm 1/2" \quad \Longrightarrow^{\pm\frac{\beta}{2}} \quad "\sigma' = \pm 1/2"$$

where $\frac{\beta}{2}$ and $-\frac{\beta}{2}$ are 'phases of implication'.

If we also propose the following law of Bayesian reasoning with complex-valued probabilities:

$$\text{If } A \Longrightarrow^{\alpha} B, \text{ then } \Phi(B|A) = e^{i\alpha}$$

Then we have succeeded in logically deriving the transformation function from physical law and Bayesian reasoning.

In performing two successive rotations through angles β_1 and β_2 about the $z$axis, the angles add. So the phases of implication must also add. Hence, we need another law of Bayesian reasoning with complex-valued probabilities, namely

$$\text{If } A \Longrightarrow^{\alpha_1} B \text{ and } B \Longrightarrow^{\alpha_2} C, \text{ then } A \Longrightarrow^{\alpha_1+\alpha_2} C$$

This means that any proposition implies itself with a phase equal to zero. For if $A \Longrightarrow^{\alpha} A$ and $\alpha \neq 2\pi n$, then applying this twice we get $A \Longrightarrow^{\alpha} A \Longrightarrow^{\alpha} A$ so $A \Longrightarrow^{2\alpha} A$ which is a contradiction. So we always have that

$$A \Longrightarrow^{0} A$$

At this point it is necessary to resolve a paradox. This arises from the form of the transformation function

$$\begin{pmatrix} e^{i\beta/2} & 0 \\ 0 & e^{-i\beta/2} \end{pmatrix}$$

Under a full rotation, i.e. when $\beta = 2\pi$, this does not reduce to the unit matrix.

It should, because when $\beta = 2\pi$, the propositions "$\sigma = \pm 1/2$ " and "$\sigma' = \pm 1/2$ " are the same, and imply each other with zero phases of implication.

The paradox is removed by proposing another physical law, namely that

> *The group of rotations of any coordinate system or rigid body is represented by the group* SU(2) (*rather than the group* SO(3)). *So, <u>two</u> full rotations about any axis are needed to return a coordinate system, or a rigid body, to its original orientation in space.*

Then, of course, we need get the unit matrix only when $\beta = 4\pi$ or a multiple of 4π, as we do.

Rotations about the y axis.

We can similarly derive the transformation functions for rotation about the $y$axis through angles π and $\pi/2$.

We do this by claiming the physical laws

$$"\sigma = -\frac{1}{2}" \implies^{0} "\sigma' = \frac{1}{2}"$$
$$"\sigma = \frac{1}{2}" \implies^{\pi} "\sigma' = -\frac{1}{2}"$$

From which logically follows the correct transformation function

$$\begin{pmatrix} \Phi(\frac{1}{2}'|\frac{1}{2}) & \Phi(\frac{1}{2}'|-\frac{1}{2}) \\ \Phi(-\frac{1}{2}'|\frac{1}{2}) & \Phi(-\frac{1}{2}'|-\frac{1}{2}) \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}_{\circlearrowleft \pi y} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

Feynman finds the transformation function for a rotation $\pi/2$ about the $y$axis, by equating the transformation functions for two successive rotations through $\pi/2$, to the transformation function for rotation through π. That is by solving

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}_{\circlearrowleft \frac{\pi}{2} y} \begin{pmatrix} a & b \\ c & d \end{pmatrix}_{\circlearrowleft \frac{\pi}{2} y} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

for a, b, c and d. The result is

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}_{\circlearrowleft \frac{\pi}{2} y} = \pm \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$$

In a similar manner we can derive, from the identity

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}_{\circlearrowleft \frac{\pi}{2} y} \begin{pmatrix} a & b \\ c & d \end{pmatrix}_{\circlearrowleft -\frac{\pi}{2} y} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

the result

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}_{\circlearrowleft -\frac{\pi}{2} y} = \pm \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}$$

where the sign in the first result, while not determined, is necessarily the same as the sign in the second.

Rotations about the x axis.

As Feynman explains, the above results, to the transformation equation for rotation through any angle α about the $x$axis. For, such a rotation is equivalent to a rotation $\pi/2$ about the $y$axis, followed by a rotation α about the new $z$axis, followed by a rotation $-\pi/2$ about the new $y$axis.

The result is

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}_{\circlearrowleft \alpha x} = \begin{pmatrix} \cos(\frac{\alpha}{2}) & i\sin(\frac{\alpha}{2}) \\ -i\sin(\frac{\alpha}{2}) & \cos(\frac{\alpha}{2}) \end{pmatrix}$$

Any rotation

A general transformation function for rotation, that is, the transformation function of ordinary QM mentioned at the beginning of this Section, is now reproduced by multiplying in order, the matrices representing the three rotations Euler stipulates.

We are thus led to think that Bayesian reasoning using complex-valued probabilities might well help in establishing a satisfactory realist interpretation of the quantum mechanical formalism.

13. The Schrödinger equation for a spin one-half particle in a magnetic field.

To give further support for the claim that Bayesian reasoning using complex-valued probabilities helps in establishing a realist interpretation of the quantum mechanics, we show how it can lead to the Schrödinger equation for a spin one-half particle in a uniform magnetic field $\boldsymbol{H}$.

We do this by utilizing a transfer function connecting the z components of spin at different times when $\boldsymbol{H}$ points in the z direction, and the transfer function connecting the z components of spin in coordinate systems of different orientation.

When $\boldsymbol{H}$ points in the z direction and is independent of time, $H_x = 0$, $H_y = 0$ and $H_z = H \neq 0$. Under these conditions we obtain the equation in quantum mechanics for the transformation function, i.e.

$$\Phi(\sigma t|\sigma_1 t_1) = \begin{pmatrix} e^{i\frac{\mu H}{\hbar}(t-t_1)} & 0 \\ 0 & e^{-i\frac{\mu H}{\hbar}(t-t_1)} \end{pmatrix}$$

simply by proposing the following physical law:

"$\sigma_1 = \pm\frac{1}{2}$ at time t_1" $\Longrightarrow^{\pm\frac{\mu H}{\hbar}(t-t_1)}$ "$\sigma = \pm\frac{1}{2}$ at time t"

where μ is the magnetic moment of the particle whose dimension is . The choice of phases in this law might be justified as follows.

Firstly, phase normalisation requires that the phases be equal and opposite. Secondly, they must be linear functions of the time, as may be seen by multiply together the transformation matrices for two successive time intervals and equating it the transformation matrix for the sum of the intervals. Thirdly, the only combination of the quantities H, μ, $\hbar$ and $t - t_1$ (which alone are relevant) is $k\frac{\mu H}{\hbar}(t - t_1)$, where k is a real numerical constant. The factor k which can be absorbed into the

magnetic moment and given a value from results of modelled experiments.

To find the transformation function when the magnetic field points away from the origin in a general direction specified by the polar angles θ and φ, we make use of the transformation function relating the z components of spin in Cartesian coordinates of different orientation.

….to be continued.

APPENDIX A

Examples of the calculation of measure densities using the method of transformation groups

---

The problem of calculating the measure density, if we do not already know it, may be difficult. There is as yet, no general way to do it.

However, as noted in Section 11a, it can sometimes be done using the method of transformation groups to find our prior distribution under no knowledge other than the requirement for normalisation. As we noted in Section 11b, this prior distribution coincides with the measure density.

We give below, two examples of the derivation of measure density using this method.

Example 1

Consider the physical variable (the angle θ) specifying the angular position of the minute hand of a mechanical clock relative to a particular angular position taken as the reference. The take any probability density $p(\theta)$ and the

measure density $m(\theta)$, to be continuous periodic functions with period 2π.

In reasoning about probability densities over θ we need to know the measure density $m(\theta)$.

Suppose we are uncertain what this measure density is. The we can find it by finding the prior probability density $\mathrm{p}(\theta)$ when we are quite indifferent as regards the value of the angle θ defining the hand's position. This probability density will coincide with measure density required.

We thus seek the prior probability density $p(\theta)$ in the polar angle θ.

Figure

Let us rotate the reference direction through angle α, and use the angle θ' $(= \theta - \alpha)$ to represent the position of the minute hand. Then, the new probability density $p'(\theta')$ will, by similarity, be the same *function* of the new coordinate θ' with the same periodicity. So

$$p'(x) = p(x)$$

Secondly, the probability the hand is in a given element of angle is, of course the same, no matter which coordinate system is used. This means

$$p'(\theta')d\theta' = p(\theta)d\theta$$

where $d\theta' = d\theta$. Hence

$$p'(\theta - \alpha)d\theta = p(\theta)d\theta$$

And because p and p' are the same function,

$$p(\theta - \alpha) = p(\theta)$$

and, since α is arbitrary $p(\theta)$ can only be a constant. By normalisation its value is $(2\pi)^{-1}$.

Thus

$$m(\theta) = (2\pi)^{-1}$$

That is, the measure density is just a constant.

Example 2

Suppose we hold, initially, only the knowledge that a bead can slide along a straight wire of length d, and is at rest at some position along the wire. What, we may ask, is the measure density relating to its position along the length of the wire?

We would be inclined to think that any segment of the wire has a natural measure, namely it's length. So, the measure density should be constant, i.e. the same at each point along the wire. However, we might worry that

limited length of the wire might affect our prior probability density for bead position along the wire, and hence affect the measure density. That is, that there might be 'end-effects'. We can, however, demonstrate that this is in fact not the case.

In the first place it is clear, by the homogeneity and isotropy of space, that it is of no matter where the wire is situated in space, or how it is orientated. We may, therefore, without lack of generality suppose the wire is on the positive $x$axis of a Cartesian coordinate system.

We consider two situations, the first in which one end of the straight wire is at $x = c$, and the other at $x = c + d$; and our knowledge Y_1 is only that the bead lies at a point in the interval from c to $c + d$. In the second, the wire extends from the origin of the $x$axis to the point $x = b$ where $b > c + d$; and our knowledge Y_2 is only that the bead lies at a point in the interval from 0 and b.

In the first case we denote by $p_1(x - c)$ $c < x < c + d$, the probability density under knowledge Y_1. In the second case we denote by $p_2(x)$ $0 < x < b$, the probability density under knowledge Y_2.

Figure

We have probabilities $P(dx\lfloor Y_1)$ and $P(dx\lfloor Y_2)$ for the bead being in dx at x, over the limited range $c < x < c + d$.

Using the product rule we have,

$$P(Y_1 dx\lfloor Y_2) = P(Y_1\lfloor Y_2)P(dx\lfloor Y_1 Y_2), \quad c < x < c + d$$

Here, Y_1 is redundant on the LHS, and Y_2 is redundant in the last term on the RHS. The probability densities $p_1(x)$ and $p_2(x)$ associated with Y_1 and Y_2 therefore satisfy

$$p_2(x)dx = \alpha(c)p_1(x - c)dx, \quad c < x < c + d$$

where $\alpha(c)$ $(= P(Y_1\lfloor Y_2))$ is independent of x but is possibly a function of c.

Now, the problem of finding $p_1(x - c)$ for $c < x < c + d$ under knowledge Y_1 is similar to that of finding $p_2(x)$ for $0 < x < b$ under knowledge Y_2. The only difference is one of scale. We therein invoke the transformation group of scaling.

From a point in the xyplane, we project points x on the wire $c < x < c + d$ onto points x' on the wire of length b:

Figure

By similar triangles we have

$$\frac{dx'}{dx} = \frac{x'}{x-c} = \frac{b}{d}$$

Similarity on scaling leads to the requirement that

$$p_1(x-c)dx = p_2(x')dx'$$

or

$$p_1(x-c) = p_2\left(\frac{b}{d}(x-c)\right)\frac{b}{d}, \qquad c < x < c+d$$

Combing this result with the relation

$$p_2(x) = \alpha(c)p_1(x-c), \;\; c < x < c+d$$

established above, we have

$$p_2(x) = \alpha(c)p_2\left((x-c)\frac{b}{d}\right)\frac{b}{d}, \qquad c < x < c+d$$

Dividing through by $\alpha(c)$ and differentiating both sides with respect to x, we get

$${p_2}'\left((x-c)\frac{b}{d}\right)\left(\frac{b}{d}\right)^2 = \frac{{p_2}'(x)}{\alpha(c)}$$

By differentiating both sides with respect to c instead we get

$$p_2{}'\left((x-c)\frac{b}{d}\right)\left(\frac{b}{d}\right)^2 = \frac{p_2(x)}{\left(\alpha(c)\right)^2}\alpha'(c)$$

Since the LHSs are the same it follows, by equating the RHSs, that

$$\frac{p_2{}'(x)}{p_2(x)} = \frac{\alpha'(c)}{\alpha(c)}$$

Because the LHS is a function only of x and the RHS only of c, both sides are equal to a constant λ making

$$p_2(x) = \eta e^{-\lambda x}, \qquad \alpha(c) = \mu e^{-\lambda c}$$

Inserting these solutions into

$$p_2(x) = \alpha(c)p_2\left((x-c)\frac{b}{d}\right)\frac{b}{d}\,, \qquad c < x < c+d$$

derived above, we have the requirement

$$e^{\lambda(x-c)} = \mu\frac{b}{d}e^{\lambda(x-c)\frac{b}{d}}$$

This means

$$\frac{b}{d} = 1 \text{ and } \mu = 1, \quad or \quad \lambda = 0 \text{ and } \mu = \frac{d}{b}$$

As $b/d > 1$ only the second option is possible. This means $\alpha(c)$ is just the constant μ and $p_2(x)$ is just the constant η; and because

$$p_2(x) = \alpha(c)p_1(x - c) \qquad c < x < c + d$$

the prior probability density sought. i.e. $p_1(x - c)$, is a constant (equal to $1/d$ for normalisation).

The measure density in the first case is accordingly $m(x) = 1/d$. It follows too, of course, that the measure density in the second case is $m(x) = 1/b$. In both cases, and generally, for a straight wire of any length, the measure density is constant along the wire, as we suspected.

.